

Compiled Public Comments on Proposed
Provisions for a Draft Data Management
and Sharing Policy for NIH Funded or
Supported Research

Notice Number: NOT-OD-19-014

October 10, 2018 – December 10, 2018

Name, Name of Organization

1. ANONYMOUS
2. SHANNON GOURLEY, EMORY UNIVERSITY
3. THOMAS GILL, YALE SCHOOL OF MEDICINE
4. REBECCA KRUKOWSKI, UNIVERSITY OF TENNESSEE HEALTH SCIENCE CENTER
5. HANS IJZERMAN, UNIVERSITE GRENOBLE ALPES
6. STEPHANIE BOHON, UNIVERSITY OF TENNESSEE
7. ELINE APPELMANS, FRED HUTCHINSON CANCER RESEARCH CENTER
8. VARON TOMER, ALBERT EINSTEIN COLLEGE OF MEDICINE
9. CATHERINE BUNCE, UNIVERSITY OF ROCHESTER
10. PETER PREUSCH, **NIGMS, NIH**
11. LAUREN DI MONTE, UNIVERSITY OF ROCHESTER
12. RICHARD KRAVITZ, UC DAVIS
13. MEGAN GUNNAR, UNIVERSITY OF MINNESOTA
14. ANONYMOUS
15. BRIAN SHOICHET, UCSF
16. JOHN GUCKENHEIMER, CORNELL
17. MARA MATHER, USC
18. KAREL SVOBODA, HHMI
19. STEVEN KAWUT, INIVERSITY OF PENNSYLVANIA
20. JOAQUIN ESTRADA, MEDICAL ORGANIZATION FOR LATINO ADVANCEMENT
21. DANIEL GOLDENHOLZ, HARVARD MEDICAL SCHOOL, **BIDMC**
22. MICHAEL BERNAUER, UNIVERSITY OF NEW MEXICO HEALTH SCIENCE LIBRARY AND INFORMATICS CENTER
23. BORRIES DEMELER, UNIVERSITY OF MONTANA
24. CLARICE WEINBERG, NIEHS, NIH
25. ANONYMOUS
26. ANONYMOUS
27. NICHOLAS L CHIA, MAYO CLINIC
28. JEFFREY PENNINGTON, THE CHILDREN'S HOSPITAL OF PHILADELPHIA
29. KIM LITTLEFIELD, UNIVERSITY OF NORTH CAROLINA GREENSBORO
30. ANONYMOUS, BOISE STATE UNIVERSITY
31. DR. RAY UZWYSHYN, TEXAS STATE UNIVERSITY LIBRARIES
32. HUNTER **N.B.** MOSELEY, UNIVERSITY OF KENTUCKY
33. LYLE G. BEST, MISSOURI BREAKS INDUSTRIES RESEARCH INC
34. BRYANT THOMAS KARRAS MD, STATE OF WASHINGTON
35. SHELLEY COLE, TEXAS BIOMED
36. ANONYMOUS
37. QINGLING SUN, SUN TECHNOLOGIES & SERVICES, LLC
38. CLARK C. EVANS, PROMETHEUS RESEARCH, LLC
39. ALEXANDER TSAI, HARVARD MEDICAL SCHOOL
40. ANONYMOUS, INTERNATIONAL COMMITTEE OF MEDICAL JOURNAL EDITORS (ICMJE)
41. CHRIS PAPAIOANNOU, RUTGERS, THE STATE UNIVERSITY OF NEW JERSEY
42. HOWARD Fox, UNIVERSITY OF NEBRASKA MEDICAL CENTER

43. ELINOR SCHOENFELD, STONY BROOK UNIVERSITY
44. BRIANNA R LINDSAY, UNIVERSITY OF PENNSYLVANIA
45. PITTR ADAMCZVK, UNIVERSITY OF WISCONSIN
46. KATHY HELZLSOUER, NCI
47. MARTIN GRUEBELE, U. OF ILLINOIS
- 48.** NORBERT PERRIMON, HARVARD MEDICAL SCHOOL
49. GAIL ADLER, MD, PHD, BRIGHAM AND WOMEN'S HOSPITAL/HARVARD MEDICAL SCHOOL
50. JENNIFER DARRAGH, DUKE UNIVERSITY
51. DAVID R. BOBBITT, CDISC
52. YONGJIAN LIU, WASHINGTON UNIVERSITY SCHOOL OF MEDICINE
53. JERRY POWER, USC
- 54.** DENISE STURDY, DUKE CLINICAL RESEARCH INSTITUTE
55. HELEN, BERMAN
56. ELISA A. HURLEY, PUBLIC RESPONSIBILITY IN MEDICINE AND RESEARCH **(PRIM&R)**
57. REBECCA H. LI, VIVLI, INC.
58. SARAH WRIGHT, CORNELL UNIVERSITY
59. C. TITUS BROWN, UNIVERSITY OF CALIFORNIA DAVIS
60. JONATHAN PETERS, DATA SERVICES, VIRGINIA TECH UNIVERSITY LIBRARIES
61. MARY ELLEN K. DAVIS, EXECUTIVE DIRECTOR ACRL, ASSOCIATION OF COLLEGE AND RESEARCH LIBRARIES
62. PAUL ANDERSON, BRIGHAM AND WOMEN'S HOSPITAL
63. DAVID MELLOR, CENTER FOR OPEN SCIENCE
- 64.** JAMES KENT, UNIVERSITY OF IOWA
- 65.** TODD CONSTABLE, YALE UNIVERSITY
66. BROOKE N. MACNAMARA, CASE WESTERN RESERVE UNIVERSITY
67. KERRY RESSLER, MCLEAN HOSPITAL
68. WADE HARPER, HARVARD MEDICAL SCHOOL
69. ANDREW REIMER, CASE WESTERN RESERVE UNIVERSITY
70. KEVIN MCGHEE, NEW YORK GENOME CENTER
71. MARYROSE FRANKO, HEALTH RESEARCH ALLIANCE
72. ANONYMOUS
73. SIRIMON O'CHAROEN, CROHN'S & COLITIS FOUNDATION
74. FINLAY MACRAE, UNIVERSITY OF MELBOURNE
75. BRUCE R. THOMADSEN, PHD, PRESIDENT, AMERICAN ASSOCIATION OF PHYSICISTS IN MEDICINE
76. ALLEN A. DIPALMA, UNIVERSITY OF PITTSBURGH
77. ANDRE NOEL PORTER, AMERICAN SOCIETY OF BIOCHEMISTRY AND MOLECULAR BIOLOGY
- 78.** MEIR STAMPFER, BRIGHAM AND WOMEN'S HOSPITAL
79. HOLLY J FALK-KRZESINSKI, PHD, ELSEVIER
80. HARRY W. ORF, MASSACHUSETTS GENERAL HOSPITAL
- 81.** XIA JING, OHIO UNIVERSITY
82. JAMES P SLUKA, INDIANA UNIVERSITY
- 83.** RAJA MAZUMDER, THE GEORGE WASHINGTON UNIVERSITY
- 84.** REBECCA OSTHUS, AMERICAN PHYSIOLOGICAL SOCIETY
85. GREG RASCHKE, SENIOR VICE PROVOST AND DIRECTOR OF LIBRARIES, NORTH CAROLINA STATE UNIVERSITY LIBRARIES

86. ANONYMOUS
87. JULIET P. LEE, PREVENTION RESEARCH CENTER OF PIRE
88. UC DAVIS LIBRARY, UNIVERSITY OF CALIFORNIA, DAVIS
89. WENDY D. STREITZ, UNIVERSITY OF CALIFORNIA
90. MARK MUSEN, STANFORD UNIVERSITY
91. MARA BLAKE ON BEHALF OF JHU DATA SERVICES, JOHNS HOPKINS UNIVERSITY
- 92. ANA SANCHEZ, DUKE UNIVERSITY**
- 93. SALVATORE LA ROSA, CHILDREN'S TUMOR FOUNDATION**
- 94. LOIC LE MARCHAND, UNIVERSITY OF HAWAII CANCER CENTER**
95. SANORA ORCHARD (ISB CHAIR), INTERNATIONAL SOCIETY FOR BIOCURATION
96. ALEX BATEMAN, THE UNIPROT CONSORTIUM
97. NICOLE HENWOOD, NF2 BIOSOLUTIONS
98. CLAIRE ZHU, NCI
99. MARY JO HOEKSEMA, POPULATION ASSOCIATION OF AMERICA/ASSOCIATION OF POPULATION CENTERS
100. MARCIN CIESLIK, UNIVERSITY OF MICHIGAN
101. MICHAEL LITZINGER, PROJECT DATA SPHERE, LLC
102. ARMAN VASHAR KHOJANOI, NATIONAL INSTITUTE OF MENTAL HEALTH
103. MARGARET LEVENSTEIN, INTER-UNIVERSITY CONSORTIUM FOR POLITICAL AND SOCIAL RESEARCH
104. IAIN HRYNASZKIEWICZ, SPRINGER NATURE
105. ANNE KLIBANSKI, **M.D.**, PARTNERS HEALTHCARE
106. CHUCK COOK, EMBL-EUROPEAN BIOINFORMATICS INSTITUTE
107. EMILY HAOZOUS, PACIFIC INSTITUTE OF RESEARCH AND EVALUATION
108. CAROLE MITNICK ON BEHALF OF, HARVARD MEDICAL SCHOOL, DEPARTMENT OF GLOBAL HEALTH & SOCIAL MEDICINE
109. CHRIS BOURG, **MIT** LIBRARIES
110. MARY **M.** LANGMAN, MEDICAL LIBRARY ASSOCIATION
111. JASON WILLIAMS, COLD SPRING HARBOR LABORATORY
- 112. TANEISHA WILSON, SOCIETY OF ACADEMIC EMERGENCY MEDICINE AND ACADEMY OF EMERGENCY PHYSICIANS**
- 113. AUDIE ATIENZA, PHD (REPRESENTING ICF GENERALLY), ICF**
114. CARL MCKINLEY, REGENSTRIEF INSTITUTE
115. MATTHEW TRUNNELL, FRED HUTCHINSON CANCER RESEARCH CENTER
116. JAMES M. MUSSER, MD, PHD, FEDERATION OF AMERICAN SOCIETIES FOR EXPERIMENTAL BIOLOGY
{FASEB}
117. AMERICAN COLLEGE OF RADIOLOGY, AMERICAN COLLEGE OF RADIOLOGY
118. ANONYMOUS
119. CATHERINE LURIA, LABORATORY OF SYSTEMS PHARMACOLOGY, HARVARD MEDICAL SCHOOL
120. CHRISTINE ZAROECKI, RCSB PROTEIN DATA BANK
121. MICHAELA SEIBER, MPH, COLLABORATIVE RESEARCH CENTER FOR AMERICAN INDIAN HEALTH
122. LISA SIMPSON, ACADEMYHEALTH
123. ANONYMOUS
124. MICHAEL MABE, INTERNATIONAL ASSOCIATION OF SCIENTIFIC TECHNICAL AND MEDICAL PUBLISHERS
(STM)
125. BRETT HARNET, UNIVERSITY OF CINCINNATI

126. ALESSIA DANIELE, WEILL CORNELL MEDICINE
127. AMY KOSHOFFER, UNIVERSITY OF CINCINNATI LIBRARIES
128. DANIEL SHRINER, NATIONAL HUMAN GENOME RESEARCH INSTITUTE
129. FELICE J. LEVINE, AMERICAN EDUCATIONAL RESEARCH ASSOCIATION
130. VINCE MOR AND JULIE LIMA, BROWN UNIVERSITY
131. ERIN GARRISON, TRIBAL NATIONS RESEARCH GROUP
132. DUSHANKA KLEINMAN AND MARY SHELLEY, UNIVERSITY OF MARYLAND SCHOOL OF PUBLIC HEALTH
133. DENIS WIRTZ, JOHNS HOPKINS UNIVERSITY
- 134.** KAREN ESTLUND, ROBYN REED, CYNTHIA HUDSON VITALE, , PENNSYLVANIA STATE UNIVERSITY LIBRARIES
135. LAURE HAAK, ORCID, INC
136. JAMES GLAZIER, INDIANA UNIVERSITY
137. ANONYMOUS
- 138.** JULIE PALMER, BOSTON UNIVERSITY
139. LOIS BRAKO, UNIVERSITY OF MICHIGAN HRPP
140. MARGARET MCCARTHY, YALE COLLABORATION FOR RESEARCH INTEGRITY AND TRANSPARENCY
141. SHARAD VERMA, NTAP AT JOHNS HOPKINS
142. LAURA THORNHILL, ALZHEIMER'S ASSOCIATION
143. FERNANDO RIOS, CHRIS KOLLEN, UNIVERSITY OF ARIZONA - OFFICE OF DIGITAL INNOVATION AND STEWARDSHIP
144. LUBA SMOLENSKY, THE MICHAEL J. FOX FOUNDATION
145. ANURUPA DEV, ASSOCIATION OF AMERICAN MEDICAL COLLEGES
- 146.** PETER SCHIFFER, YALE UNIVERSITY
- 147.** ARA TAHMASSIAN, HARVARD UNIVERSITY
- 148.** TOM SELLERS, PHD, CENTER DIRECTOR, H. LEE MOFFITT CANCER CENTER & RESEARCH INSTITUTE, INC.
- 149.** ADAM THOMAS, **NIMH IRP, NIH**
150. SANTIAGO SCHNELL, STRENDA COMMISSION
151. KITCKI CARROLL, UNITED SOUTH AND EASTERN TRIBES
152. YVETTE ROUBIDEAUX MD MPH, NCAI POLICY RESEARCH CENTER
153. JEFFERY SMITH, **AMIA**
154. JUERGEN KLENK, DELOITTE CONSULTING LLP
155. PAMELA A. WEBB | LISA JOHNSTON, UNIVERSITY OF MINNESOTA
156. HEIDI REHM, PHD, FACMG, CLINICAL GENOME RESOURCE
157. JAMES REECY, IOWA STATE UNIVERSITY
158. EVERETT R. RHOADES MD, FACP (RET.), PRIVATE CITIZEN MEMBER KIOWA TRIBE OF OKLAHOMA
159. JULIE STONER, UNIVERSITY OF OKLAHOMA HEALTH SCIENCES CENTER
160. HEATHER STEVENS, ACCENTURE FEDERAL SERVICES LLC
161. HEIDI IMKER, UNIVERSITY OF ILLINOIS AT URBANA CHAMPAIGN
162. FRANCIS P. CRAWLEY, GOOD CLINICAL PRACTICE ALLIANCE - EUROPE (GCPA) & STRATEGIC INITIATIVE FOR DEVELOPING CAPACITY IN ETHICAL REVIEW (SIDCER)
163. YOORI KIM, GILBERT FAMILY FOUNDATION
164. JENNIFER HALL, AMERICAN HEART ASSOCIATION
165. MARK CULLEN, STANFORD UNIVERSITY SCHOOL OF MEDICINE
166. LAURA PLATERO, NORTHWEST PORTLAND AREA INDIAN HEALTH BOARD

167. MEGAN POTTERBUSCH, HIROMI SANDERS, NINA HAMBURG, ANNE LINTON, GEORGE WASHINGTON UNIVERSITY
168. CHRIS SHAFFER & JOHN CHODACKI, UNIVERSITY OF CALIFORNIA SAN FRANCISCO & CALIFORNIA DIGITAL LIBRARY
169. RAJNI SAMAVEDAM, Booz ALLEN HAMILTON INC.
170. JAMES LUTHER, DUKE UNIVERSITY
171. MARY PIORUN, PH.O., DIRECTOR, LAMAR SOUTTER LIBRARY, LAMAR SOUTTER LIBRARY, UNIVERSITY OF MASSACHUSETTS MEDICAL SCHOOL
172. MELISSA HAENDEL, OREGON STATE UNIVERSITY
173. LAURA QUILTER, UNIVERSITY OF MASSACHUSETTS AMHERST, LIBRARIES
174. ELAINE MARTIN, JULIE GOLDMAN, COUNTWAY LIBRARY OF MEDICINE, HARVARD MEDICAL SCHOOL
175. DANIEL HANDWERKER, **NIMH** (WRITING IN MY PERSONAL CAPACITY, NOT AS A REPRESENTATIVE OF **NIMH**), **NIH**
176. JASON BRET HARRIS, COLLABORATIVE DRUG DISCOVERY, INC. (COD)
177. KATIE STEEN, ASSOCIATION OF AMERICAN UNIVERSITIES (AAU)
178. RICARDO DE MIRANDA AZEVEDO, MAASTRICHT UNIVERSITY
179. ANDREW TEIN, WILEY
180. ALEXANDER (SASHA) WAIT ZARANEK, CUROVERSE RESEARCH
181. DIANE LEHMAN WILSON, UNIVERSITY OF MICHIGAN MEDICAL SCHOOL
182. SARAH GREENE, HEALTH CARE SYSTEMS RESEARCH NETWORK
183. WILLIAM HERSH, OREGON HEALTH & SCIENCE UNIVERSITY

Submission #1**Date:** 10/11/2018**Name:** Anonymous**Name of Organization:****Type of Organization:** Other**Other Type of Organization:** Consulting Group**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

NA - all research involving human subjects

II. The requirements for Data Management and Sharing Plans

NIH should define deidentification and describe standards for de-identification when data sharing includes individual-level data. There is much confusion in the scientific and human protections community on this topic given the various concepts of identifiability and the growing literature (and availability of algorithms) about reidentification of supposedly anonymized or de-identified data. Does de-identified mean that "investigators cannot reasonably ascertain the identity of subjects" per the Common Rule? Does it mean that the data has been de-identified in accordance with the Safe Harbor or Expert Determination methods per HIPAA? Is data identifiable when there is "at least a very small risk, as determined by current scientific practices or statistical methods, that some combination of the information, the request, and other available data sources could be used to deduce the identity of an individual" per the standards for Certificates of Confidentiality at 42 U.S.C. 241(d)? What about when information is coded? And when the code is provided to the repository so that a GUID can be issued to link individual data across studies?

Likewise, NIH should address acceptable standards for broad consent and, when embedded within the consent for the original research, whether an opt-in method of consent is required and when (e.g., when the research includes genomic data, when the research under which the data will be generated confers the possibility of direct benefit to individual participants, when the data was obtained pursuant to HIPAA authorization, when data sharing is not intrinsic to the aims of the original research (i.e., the purpose of the research is not to establish a repository), when the research involves tribal or indigenous populations, etc.).

Submission #2

Date: 10/11/2018

Name: Shannon Gourley

Name of Organization: Emory University

Type of Organization: University

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

neuroscience

I. The definition of Scientific Data

reasonable

II. The requirements for Data Management and Sharing Plans

Generating these sorts of additional documents for the grant review process feels to me like an activity meant to eliminate a few bad apples in the field, but in reality, will burden all of us in a climate in which we are already over-burdened with administrative duties. I urge the NIH to remember that every one of these kinds of activities take us away from actually leading/conducting scientific research. Further, given that more and more journals are requiring data sharing, this activity seems unnecessary because we will already be required to share data upon publication.

Uzwyshyn: There are great reasons to share research data, related to discovery and the promotion and progress of the scientific enterprise - especially at universities where most primary research on every topic imaginable is carried out.

Online data repositories allow easy global availability, sharing, and download. *Online research data repositories are now pragmatic realities. Most discovery in the future will be predicated on the sharing and synthesis of data.*

From data regarding the search for the cure of diseases to double checking conclusions for new scientific and social scientific discoveries, possibilities are manifold. It's the vast majority of research data being produced, for which online data repositories are best. Given that the research data isn't classified and will be personally de-identified, the situation currently is that there are specific federal mandates to make a researcher's data available and publicly accessible if they are receiving federal funding from any of the large U.S. granting agencies.

Grush: Beyond making datasets accessible, are researchers who use federal funds required to outline their data management plans?



Any large-grant-seeking researcher must now produce a Data Management Plan (DMP), especially when applying for large grants, (say NSF or NIH) and they must describe how they will make their data accessible. Applicants may wish to note that there is a good documentation and policy planning tool, the DMPTool (see DMPTool, <https://dmptool.org> and <https://dmptool.org/video>), available online through the California Digital Library, that helps researchers create their DMP. An online data repository to house a researcher's data is a central piece of that puzzle.

Submission #4**Date:** 10/11/2018**Name:** Rebecca Krukowski**Name of Organization:** University of Tennessee Health Science Center**Type of Organization:** University**Role:** Scientific Researcher**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Behavioral research

I. The definition of Scientific Data

Scientific data should include intervention materials for behavioral interventions, so that researchers can replicate previous research. Without these materials, it is not possible to replicate the intervention. If these intervention are created with taxpayers' dollars, they should be freely accessible.

II. The requirements for Data Management and Sharing Plans

NIH should provide a central repository for data, so individual investigators do not have to develop these themselves.

Submission #5

Date: 10/12/2018

Name: Hans IJzerman

Name of Organization: Université Grenoble Alpes

Type of Organization: University

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

psychology/neuroscience

I. The definition of Scientific Data

I don't have specific feedback to your questions, because the questions require more than three text boxes to discuss. I did want to point you to a recent initiative I was involved with for data sharing in psychological science, as I think it can address many of your ideas/concerns: <https://www.collabra.org/articles/10.1525/collabra.158/>. Here is one way we implement having a solid sharing workflow in our lab: <https://osf.io/q29nf/>

Submission #6**Date:** 10/14/2018**Name:** Stephanie Bohon**Name of Organization:** University of Tennessee**Type of Organization:** University**Role:** Scientific Researcher**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Demographic research on health

II. The requirements for Data Management and Sharing Plans

I agree that sharing of data is essential to good science; however, better infrastructure is needed for sharing both data and the methods by which data are analyzed (both need to be shared). Platforms for sharing quantitative data are better (and better supported) than those for sharing qualitative data (which arguably needs more oversight due to heightened concerns with confidentiality). ICPSR is well-funded and provides a good platform for sharing quantitative data. QDR is less well-funded. Currently, QDR requires a deposit fee, which can be written into an application, but it not clear for how long data can be deposited and how often a researcher will have to pay. Our own libraries are probably not sufficiently well-versed in data protection to be trusted with the distribution of confidential data.

Another issue that will need careful consideration are issues related to data (and methods) shared from research generated within FRDCs. NIH will need to coordinate carefully with the Census Bureau and other agencies on this issue.

Finally, NIH should pay careful attention to the possibility of re-identification that is now possible with supercomputing and the creation of large network models. You may want to consult with John Abowd (Census/Cornell), as he is at the forefront of people considering the hazards of re-identification in network modeling.

Submission #7

Date: 10/14/2018

Name: Eline Appelmans

Name of Organization: Fred Hutchinson Cancer Research Center

Type of Organization: Nonprofit Research Organization

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

Biostatistics, Bioinformatics, and Epidemiology Program

II. The requirements for Data Management and Sharing Plans

There is a need for separate data management and data sharing plan requirements for biostatistics, bioinformatics, and epidemiology methodology grants as well as secondary data analysis grants. The current plans, as proposed, place too much burdensome on PI's whose grants do not collect data.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and

Currently, data sharing agreements are often formalized after grant funding. The section 6 component should be incorporated in either the JIT or first year progress report.

Submission #8**Date:** 10/15/2018**Name:** Yaron Tomer**Name of Organization:** Albert Einstein College of Medicine**Type of Organization:** University**Role:** Scientific Researcher**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Cancer, Immune therapies, Neuroscience, Metabolic Disorders, Healthcare delivery, RNA science

II. The requirements for Data Management and Sharing Plans

I believe that Data Sharing should happen upon publication and replace the current practice of having enormous amount of Supplementary Information attached to many manuscripts. Instead the raw data and detailed methods associate with the publication should be available publicly in a shared lab notebook which could be provided either by the journal, or by the institution where the research was performed, or by the NIH.

I don't think that sharing all data produced every day in a lab will be useful for the following reasons:

- (1) Too much irrelevant data will be stored publicly, and it will be very difficult to reach relevant data.
- (2) It will create significant additional burden on researchers already dealing with constantly increasing regulations.
- (3) It will create issues in situations when a lab is making a breakthrough and they don't want to expose the results before publication.

However, requiring that all methods and raw data associated with a publication be deposited in a publicly available shared lab notebook will be easy to access (there will be no need for complex query mechanisms, the data will be available through PubMed searches since it will be linked to the manuscript). In addition it will not create significantly additional burden, and it will ensure confidentiality of new findings until publication.

Submission #9

Date: 10/16/2018

Name: Catherine Bunce

Name of Organization: University of Rochester

Type of Organization: University

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

HIV/AIDS, Infectious Disease

II. The requirements for Data Management and Sharing Plans

Ability to access data in a systematic and smooth process

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and

Over the next six months with phasing in over 18 months to 2 years

Submission #10

Date: 10/16/2018

Name: Peter Preusch

Name of Organization: NIGMS

Type of Organization: Government Agency

Role: Government Official

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

Basic research, multiple disciplines, translational and clinical research in a few areas affecting all organ systems.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and

- 1, The policy should be applicable to all grant mechanisms and should be rolled out in a single launch so that one does not have to guess whether the policy is applicable or not.
2. If NIH is going to require data sharing, then NIH also has to adequately support the archives and other resources that will host the data.
3. Further specifics are needed for the implementation and enforcement. Specification of flags in the eRA system that will enable IC staff to monitor compliance, particularly the post-award compliance.

Submission #11**Date:** 10/16/2018**Name:** Lauren Di Monte**Name of Organization:** University of Rochester**Type of Organization:** University**Role:** Institutional Official**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

My portfolio includes developing infrastructures and services to support data management and sharing, and open science more broadly.

I. The definition of Scientific Data

The proposed definition of scientific data is very good and covers a lot of important ground. However, in the first sentence, I think that the word "replicate" should be replaced with "reproduce." Replication means that a researcher should be able to repeat the study with out the use of the original data, where as reproducibility means that other researchers can reproduce reported findings using the shared code and data. I'm very happy to see a comprehensive list of what is not scientific data. I think it might also be helpful to define minimum requirements for what should be shared, for example, data and code to reproduce figures and stats described in a publication.

II. The requirements for Data Management and Sharing Plans

I think its important for DMPs to be part of the application process and that evaluation of the plan should be embedded into as many processes as possible--its the only way to make publicly-funded scientific research available to the broader community. If DMPs are part of technical evaluations for contracts it is important for NIH to provide clear and consistent minimum technical requirements for repositories so that institutions can ensure compliance. I'm very happy to see this new set of standardized plan elements. There needs to be as much consistency as possible across all NIH ICs so that Universities and Hospitals can better support their researchers in these planning efforts. I think the specific elements make a lot of sense. It's good to see software and code included, and to emphasize free and open software. Also good to see the emphasis on CDEs--standardizing metadata will help a lot. Defining minimum standards for digital preservation would be helpful as we evaluate and build technical infrastructures. I also think that NIH should prioritize funding shared infrastructures for sharing

and preservation, versus funding specific projects to do their own sharing and preservation. Collective action in these areas is much more effective than, smaller, individualized efforts. In terms of licensing, any "upon request" language should be specifically rejected and insufficient. It may be helpful to point to particular data licensing frameworks (e.g., Creative Commons).

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and

A better understanding of minimum standards for sharing and preservation, and fo what compliance and enforcement rules will be instituted, are required to establish a reasonable timeline for change. Institutions need to audit and evaluate their existing infrastructures to determine what kinds of changes must occur. To this end, it might be valuable for NIH to consider supporting or coordinating alternative repository forms, for example decentralized or distributed data sharing and archiving across multiple institutional nodes, rather than a few centralized repositories. This is a much more modern and sustainable approach to data storage, discovery, and preservation, that also leverages the resources, skills, and social capital of major research institutions.

Submission #12**Date:** 10/16/2018**Name:** Richard Kravitz**Name of Organization:** UC Davis**Type of Organization:** University**Role:** Scientific Researcher**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Health services research

II. The requirements for Data Management and Sharing Plans

The most important considerations to ensure successful implementation of these worthy goals are:

- 1) Making sure the rules are flexible enough to accommodate different kinds of data (clinicaltrials.gov is an example of a government initiative that generates researcher frustration when innovative designs don't fit NIH parameters);
- 2) Making every effort to make definitions clear, with lots of examples of how the standards can be met (these examples should be posted online as text but also as part of training videos for investigators and their staff); and
- 3) Assuring that there is adequate funding provided within grants so that investigators can hire appropriate personnel to implement the rules. The gap between datasets suitable for analysis and datasets suitable for sharing is large. Considerable time and expertise are needed to transform one into the other.

Submission #13**Date:** 10/18/2018**Name:** Megan Gunnar**Name of Organization:** University of Minnesota**Type of Organization:** University**Role:** Scientific Researcher**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Human research using all of the above, however, am responding as a researcher who conducts development psychology/neuroscience research with children and adolescents.

II. The requirements for Data Management and Sharing Plans

I would strongly encourage different levels of requirements for R21 vs other mechanisms. What you have is an addition 2 pages just on data management on top of a 6 page research proposal. Seems like overkill.

Annual reports and check on data sharing make little sense as in many human developmental studies, you are collecting the data up until nearly the end of the award period, processing it and getting ready to analyze and write it up in the last year and into the post-funding period. You aren't ready to put the data out for others to see until several years AFTER the award period is over. So, annual reporting DURING the award period will be a bit silly. This also points to one of the major problems with the requirement. It is an unfunded initiative. The major work on it will be done by the researcher AFTER they no longer have funding to pay the people who will be doing the work, leaving it on the researcher's shoulders and/or leaving the researcher to hunt for funds or "illegally" use funds from another grant to support posting of data from a previous grant. Asking researchers to make their data available to other while they are collecting it and before they can analyze it doesn't make sense either. Perhaps mechanisms where you can post while to collect but only release later would work. I know that some archives do work like this.

Submission #14**Date:** 10/18/2018**Name:** Anonymous**Name of Organization:****Type of Organization:** Government Agency**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

clinical

I. The definition of Scientific Data

A single research parameter can often be measured by different assays, one study may use assay X while another uses assay Y. If a secondary data user were to query the data from both studies, they need to know that different assays were used so any comparison's they make allow for assay variances. Scientific data should include specific details on methodology, assay, and algorithms used to generate and process the experimental data.

II. The requirements for Data Management and Sharing Plans

The FAIR principles should enable greater utility from the NIH funded research data for purposes beyond the initial research question. However, NIH has not made FAIR a required condition that users follow FAIR principles in either the grants or contracts they award. As such, projects often proceed to generate data in structures that are not based on a standard. Without NIH requiring funded projects to deposit data in a form that adheres to established standards, the FAIR principles are not being followed.

This issue has been recognized by NHI and the NHI has taken fragmented uncoordinated steps in trying to address them. Specifically, for large high profile projects, the NIH has funded Data Coordinating Centers (DCC) that clean-up, converting and harmonize the data from the project participant sites to establish a single standardized set of data. But that is a fragmented solution, not every funded project has a DCC and the DCC's are not required to adhere to a standard- they function independently. If a project does not feed into a DCC then this level of converting/harmonization does not occur and the data while in a database is not FAIR. Given the costs and time it takes for downstream conversion/harmonization of data, it is more economical to avoid these costs and delays by gathering the data into a data structure that is already FAIR compliant. The NIH could enable FAIR compliance by providing training resources, provide tools and technical help. Moreover, NIH could establish FAIR recognition as a primary

feature when evaluating grants and establishing contracts- something that is not done now, as FAIR issues are discussed after grants and contracts are awarded. Researcher citations could include a FAIR score that would also benefit the researcher's standing for grants and contracts. Additionally, funds could be held back until data was delivered and it accepted into FAIR compliant databases. As a major source of funding for research, the NIH has an implicit and explicit duty to look after the interests of the patients and the researchers. It is in the patients interest to ensure data is FAIR as quickly as possible and this is balanced against the researchers more focused interests, but those focused interests have out-weighed the needs of the patients. It is time the balance was redressed and the individual interests of researchers are balanced against the needs of the patients.

Submission #15

Date: 10/18/2018

Name: Brian Shoichet

Name of Organization: UCSF

Type of Organization: University

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

Basic biomedical research.

II. The requirements for Data Management and Sharing Plans

I worry that this will add further burdens to an already burdensome process (grant applications), and distract reviewers further from their main remit, reviewing scientific impact and innovation. This is among the proposals that sounds good to outsiders, but could fill the days of investigators with yet more administrivia.

Submission #16

Date: 10/19/2018

Name: John Guckenheimer

Name of Organization: Cornell

Type of Organization: University

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

Genomics

I. The definition of Scientific Data

The success of genomic databases has been greatly facilitated by NIH/NLM investment in establishing standards for data formats and building tools that enable broad groups of scientists to utilize these data. Data structures are much more complicated in other areas, but NIH can facilitate data sharing by promulgating standards and creating software for the collection of data.

Make the lives of bench scientists easier!

Submission #17

Date: 10/19/2018

Name: Mara Mather

Name of Organization: USC

Type of Organization: University

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

Cognitive neuroscience

II. The requirements for Data Management and Sharing Plans

Bravo! I am so happy to see that a data sharing plan will be required finally for all proposals. This draft looks well thought out and allows flexibility while still being clear that data sharing is a requirement.

Submission #18

Date: 10/21/2018

Name: karel svoboda

Name of Organization: hhmi

Type of Organization: Nonprofit Research Organization

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

neuroscience

I. The definition of Scientific Data

Data is the output of a study.

This includes raw data (If possible), analyzed data, metadata (formal descriptions of the data), links to protocols.

All needs to be provided in a carefully documented, ideally standardized data format.

The primacy of research papers is an anachronism.

Data, not research papers, is the primary (!) output.

The focus needs to be on data.

II. The requirements for Data Management and Sharing Plans

A plan for distributing raw data (If possible), analyzed data, metadata (formal descriptions of the data), links to protocols.

All needs to be provided in a carefully documented, ideally standardized data format.

The data needs to be made available in publ repositories (figshare; dryad, crcns etc)

Given that raw data can be extremely voluminous (PBs) it may sometimes not be possible to serve the raw data.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and

Data sharing requires effort.

But data sharing is possible and it requires carrots and sticks.

I would argue that data sharing is behind because NIH has not provided carrots or sticks.

Carrots - funding for data sharing (repositories, data formats and related software; supplements for data sharing for large data sets)

Sticks - requirements for data sharing

Submission #19**Date:** 10/21/2018**Name:** Steven Kawut**Name of Organization:** University of Pennsylvania**Type of Organization:** University**Role:** Scientific Researcher**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

All

I. The definition of Scientific Data

There should be some guidance on exactly what data need to be shared. That is currently left a bit nebulous and up to the researcher.

II. The requirements for Data Management and Sharing Plans

The enforcement of this policy will need some teeth. The reality is that data sharing in the vast majority of cases will occur after the end of the funding period of the grant in question. Any enforcement has to be planned for post-funding/support period and the current language appears weak. I would propose that non-compliance with the data sharing plan/requirements would result in withholding funds from other current grants to the investigator until the data sharing proposed and approved has occurred.

The other policy might want to include use of global unique identifiers for patients in human research, using GUID engines. These will be important for data sharing and harmonization of human studies. this would mean that researchers would have to collect identifiers locally (and create GUIDs) or plan for central transmission with generation of GUIDs centrally with destruction of identifiers.

It should also be recognized that data sharing might consume resources after the award period, so that funding during the award creating the data may not be sufficient or appropriately timed for the actual data sharing process, which will most commonly occur after the conclusion of the award period, in patient-oriented research anyway.

Submission #20

Date: 10/21/2018

Name: Joaquin Estrada

Name of Organization: Medical Organization for Latino Advancement

Type of Organization: Professional Org/Association

Role: Medical Provider

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

Clinical

Diversity and Inclusion

Epidemiology

Culturally Competent Care

Health Disparities

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and

MOLA is a non-profit association of Chicagoland Hispanic/Latino physicians working for career advancement, linguistic and cultural competency, personal wellness, and reduced health disparities for the good of the entire Hispanic/Latino community. We support increasing data sharing to improve Latino health and reduce health disparities in the Latino community.

Submission #21

Date: 10/23/2018

Name: Daniel Goldenholz

Name of Organization: Harvard Medical School, BIDMC

Type of Organization: University

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

data science and clinical epilepsy

II. The requirements for Data Management and Sharing Plans

There is no meaning to a data management and sharing plan unless NIH plans to increase budgets commensurate with the burdens associated with this additional activity.

Simply requiring a very time intensive and potentially equipment or service intensive process without any financial hooks results in a decrease in the overall productivity of investigators.

I say this as someone who has built a research career on shared data, so I don't speak lightly here.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and

10 years, phased in starting in 7 years.

Submission #22**Date:** 10/23/2018**Name:** Michael Bernauer**Name of Organization:** University of New Mexico Health Science Library and Informatics Center**Type of Organization:** University**Role:** Scientific Researcher**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Biomedical research informatics and observational/secondary use of electronic health records.

I. The definition of Scientific Data

Data that can be used either prospectively or retrospectively to address questions of scientific interest. A key point is to make sure the definition does not only focus on data collected explicitly for research. In the current digital era, much data are used for purposes other than for what they were initially intended. For example, terms and phrases entered into Google may be considered scientific data. Additionally, data collected routinely for clinical care may be considered scientific data when being used to test hypotheses or to produce generalizable knowledge/advance a field even though these data originally exists as artifact of care.

II. The requirements for Data Management and Sharing Plans

Reproducibility should be a major concern. Specific to this are issues of data provenance. When possible all data manipulations and transformations should be coded/scripted to ensure reproducibility. In addition, version control should be used to track changes to the data longitudinally and to maintain a record of manipulation. Sharing of raw data (or processed data given the steps have been adequately documented) should be required when possible. For some this may present certain challenges...technical/logistic challenges include selecting appropriate data formats and having adequate infrastructure to host the data...challenges associated with privacy/security should be addressed, especially in cases where data contain protected health information (PHI). Possible ways around this include federated databases where original author maintain control/ownership of the data. Methods of anonymization and deidentification can also help in this respect. Finally concerns around attribution should be addressed...data sharing should be recognized as a significant contribution to the field and rewarded by promotion and tenure. Also, original authors should be cited and recognized with their data have been used by others. Researchers may be reluctant to share data as hoarding it

may offer certain competitive advantages. In these cases, embargo periods may be used to allow original authors finite time to complete additional work on the dataset prior to making it available to other researchers. This may also be addressed by requiring only a limited subset of data to be shared (e.g. data sufficient to reproduce a particular result).

Submission #23**Date:** 10/25/2018**Name:** Borries Demeler**Name of Organization:** University of Montana**Type of Organization:** University**Role:** Scientific Researcher**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Biophysics

I. The definition of Scientific Data

Experimental data (raw and processed) obtained with scientific instruments, experimental designs, and ALL software used to acquire, analyze, and process such data.

II. The requirements for Data Management and Sharing Plans

Requirements for me are:

- * Data are shared in a format that is published and conforms with an open or open source standard, preferably licensed by GPL or similar license scheme
- * Data are stored in a device that can be transferred to a newer technology once it becomes obsolete (floppies, CD, DVD, BlueRay, harddrive, SSD progression)
- * Data transfers to new technology are guaranteed as they become available (must be funded!)
- * Process in place to assure data curation.

I am also a VERY STRONG proponent for requiring that all software developed with taxpayer funds (i.e., NIH, NSF, DOD, etc) have to be licensed under GPL or LGPL or other open source license that guarantees free access to such software's source code for future adoption and adaptation by other research groups to assure that funding agencies do not pay for the same work twice. If the taxpayer pays for it through NIH funding, the taxpayer should not be required to purchase this again! GPL and LGPL licenses guarantee perpetual open source licensing for codes and code snippets that can be reused by others, and LGPL even allows for flexible commercialization of such codes.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and

I don't have an opinion on this.

Submission #24

Date: 10/25/2018

Name: Clarice Weinberg

Name of Organization: NIEHS

Type of Organization: Government Agency

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

Environmental epidemiology, statistical genetics, biostatistics, reproductive epidemiology

I. The definition of Scientific Data

Looks fine.

II. The requirements for Data Management and Sharing Plans

Data sharing sounds like a great idea, and I like the idea that crowd sourcing can involve all kinds of talented people in re-analyses that might provide further insights. However, privacy concerns are serious, especially for environmental data and health data. I hope that before there are serious consequences the policy makers will take the time to read the column that recently appeared in the New England Journal of Medicine, by Joel Schwartz at Harvard. Here is the link:

https://www.nejm.org/doi/full/10.1056/NEJMp1807751?query=featured_secondary

While this piece targets the use of epidemiological data on environmental effects, the points made about the need for respecting privacy in the use of human data apply much more generally. There are additional concerns, e.g. for genetic data. For example, suppose genotypes within a family are made public and it turns out the father is not the father or the two sisters are not sisters after all. Information does not get more private than this.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and

No comment.

Submission #25

Date: 10/25/2018

Name: Anonymous

Name of Organization:

Type of Organization: University

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

Molecular biology

I. The definition of Scientific Data

In a wet lab, even the making of reagents (solutions, agar plates, handling of reagents) is data. These can determine the outcome of experiments and the reporting of supposed discoveries that later turn out to be artifacts.

II. The requirements for Data Management and Sharing Plans

This paperwork merely allows some incompetent PIs to write long plans that make them look credible. Meanwhile, the PIs who are actually in the lab trying to teach their students and make sure reagents and experiments are done correctly have less time to game the system by writing long "Data Mgt & Sharing Plans". I realize that this was well-intended. But leave it to the non-performing administrators who do not do experiments to come up with more and more items like this to detract from actual research.

Submission #26

Date: 10/26/2018

Name: Anonymous

Name of Organization :

Type of Organization: University

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

Rare diseases

I. The definition of Scientific Data

I'm not sure how you can define this--we start with observations, then develop hypotheses and systematically collect data to test the hypotheses. I would define scientific data that should be considered for sharing as any systematically collected data designed to test a hypothesis.

II. The requirements for Data Management and Sharing Plans

Data collected using tax dollars should be made available. However, there should be guidelines for managing the data, and financial support for this process.

These guidelines should be comprehensible. As in, I can understand them. I've tried understanding common data elements so I can create surveys and manage rare disease data in a way that could be easily shared, but the guidelines for NIH are incomprehensible to me.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and

Funding considerations, including funding post project, ,need to be considered.

Submission #27**Date:** 10/26/2018**Name:** Nicholas L Chia**Name of Organization:** Mayo Clinic**Type of Organization:** Nonprofit Research Organization**Role:** Scientific Researcher**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Microbiome

I. The definition of Scientific Data

All data and materials that can lead to potential benefit to future treatment or understanding of disease that does not compromise patient rights. This includes sequence data, clinical data, protocols, microbial strains, bioinformatics pipelines, and animal models.

II. The requirements for Data Management and Sharing Plans

Large data such as sequencing data should be managed in a central repository in order to prevent loss and to unburden the individual labs with the need to keep backups across many years. New microbial strains should be deposited into ATCC. All other materials and protocols should be available upon request and each program should be allowed to withhold funding for any investigator that fails to share materials that are relevant to the central mission of healthcare or public health.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and

In theory, NIH's data management should be a service for the depositing investigator and the scientific community. That means additional clarity (and personal help when needed) for each investigator, especially as they deposit data. This will ensure that these processes do not become overly burdensome or that because of difficulties, results in lost data, sample labels, or lost links between the sample data and clinical data.

Submission #28**Date:** 10/29/2018**Name:** Jeffrey Pennington**Name of Organization:** The Children's Hospital Of Philadelphia**Type of Organization:** Healthcare Delivery Organization**Role:** Scientific Researcher**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Informatics, clinical, genomic, computational biology, machine learning

I. The definition of Scientific Data

Primary (raw) data and metadata collected by direct observation linked to analytic products comprised of data, metadata, and analysis code. Linking potentially extends across many related analytic products derived at multiple levels of analysis. Scientific Data is complete when data, metadata, and analysis code are sufficient to reproduce the entire analytic chain of events used to produce a result.

II. The requirements for Data Management and Sharing Plans

Data Management plans should include a commitment to use formal archival methods and systems that implement an archival standard such as the Open Archival Information System reference model. The plan should describe how metadata will be encoded in a computable format, and how metadata will be used to describe both structure (e.g. columns, data types) and content (e.g. categorical variable values, numerical ranges). The plan should detail how data held in an archival repository will be linked to analysis code held in a version control code repository that documents the change history final code versions used to produce analysis results. IDEALLY the plan should document how the researcher will engage the support of archival experts to assist in these complex activities. We are experiencing at CHOP exponential growth in the volume of fragmented, poorly documented, and irretrievable scientific data; a chronic problem today with decades-long impact on productive use of biomedical data.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and

NIH should phase adoption starting with an educational program to train researchers in the basics of data archival and code management methods, followed by a second infrastructure phase to seed data archival and library science expertise and tools in data-generating organizations. The education phase is a timely and cost-effective way to both help researchers where they are right now (aware of the problem, confused about how to address) and prepare the community for broader and systematic data and code archival.

Submission #29**Date:** 10/29/2018**Name:** Kim Littlefield**Name of Organization:** University of North Carolina Greensboro**Type of Organization:** University**Role:** Institutional Official**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Research Training, Clinical, Musculo-skeletal, behavioral psychology

II. The requirements for Data Management and Sharing Plans

Even with these excellent updates, the Data Management and Sharing Plan will still be considered a "3rd Class" document. It will be developed at the last minute before proposal submission and will likely mostly contain minimal information that will serve to satisfy the submission requirement but will contain little meaningful information. How could it not? The majority of PIs consider the data management plan "boilerplate." Most PIs don't know how and where and when to share data. Making adherence to a data management plan a "term and condition" is a natural next step but few PIs read their NIH award terms and conditions and scramble to address data management activities in the annual RPPR. Institutions can put up policy, training and resources to support open data practices but without an incentive that drives the use of the training and infrastructure resources, adoption of open data and open scholarship simply will not be practiced robustly. I submit that the most compelling incentive to change the culture and make data management and sharing a central tenant and integral practice, is to implement that the data management plan be considered as an "additional review criteria." I make this comment as an institutional official, as a Co-Investigator, and as a NIH-reviewer. Implementing this change, on both sides, for PIs and reviewers, will provide a natural mechanism to indoctrinate responsible, open, meaningful data management and sharing plans into proposals/applications. A great first mechanism to beta test this is the F pre- and post-doctoral fellowship programs. It is this group of developing researchers that are the most receptive, and frankly, more comfortable, with the premise of open data and data sharing. The NIH has many data storage resources; have each institute name their repository of importance. Provide indicators for reviewers to be able to meaningfully evaluate the data management and sharing plan. I would be more than willing, as a reviewer, to assist a PI, through the review process, develop a meaningful data management and sharing plan. I want

to stress that this implementation could go further than the RCR presentation which is largely formulaic and largely "2nd class." Through diligence the compliance documents - "use of human subject" and "animal models" have been elevated to "1st class" documents-the information must be complete, meaningful and align with the research plan. We can, and I submit, I think we must, bring the data management and sharing plan into this same stature/importance. I would be more than happy to assist in these efforts. I would like to drive the use of the resources (training and infrastructure and expertise) to enhance the competitiveness of proposals and research projects and importantly the discovery and use of research data supported by tax payer dollars.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and

Incentivizing meaningful data management and sharing plan development and practices: encourage presentation of PIs data, code, figure citations (top 5) in Biosketch, encourage presentation of data sets discovered and used for project projection by PI (top 5) in research strategy - approach and listed in bibliography; encourage PIs to cite pre-registration activities in Biosketch;

Submission #30

Date: 10/30/2018

Name: Anonymous

Name of Organization: Boise State University

Type of Organization: University

Role: Institutional Official

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

Biomedical

Attachment:

Comments for “Proposed Provisions for a Draft NIH Data Management and Sharing Policy”
https://osp.od.nih.gov/wp-content/uploads/Data_Sharing_Policy_Proposed_Provisions.pdf

- The proposal often uses the term “could” when describing requirements for “Data Management and Sharing Plans” (ex. “Plans could be evaluated as an additional Review Consideration... - page 3). Although the document includes “proposed provisions”, we recommend funders provide clear and definitive requirements and guidelines to grantees. For many researchers, data management as described in most funder policies, is a new activity and results in poorly developed plans. Clear-cut language would help researchers in creating and implementing useful data management plans resulting in better stewardship of NIH funded data sets.
- NIH Intramural Research Projects (third bullet point on page 3) - The proposal states, “Plans could be reviewed by the Scientific Director (or designee) or Clinical Director (or designee) of the researcher’s funding IC and integrated into approval conditions as appropriate.” In addition to these individuals, we strongly recommend that plans also be reviewed by data management librarians and others from the data management community. Management of scientific data, particularly description and publishing activities, requires specialized knowledge and skill sets. Including individuals trained in these kinds of responsibilities would allow for a more thorough review and the ability for NIH to identify plans that are not feasible or simply won’t result in good data stewardship.
- Extramural Grants (first bullet point on page 3) and Plan Elements (page 3) - We recommend making data management plans a scorable part of the application and not limiting their length to two pages. The implementation of data management plan requirements has produced tremendous awareness of and movements towards the goal of rigorous and reproducible research. However, these proposed guidelines will undermine the intent of NIH’s policy and the benefits and purposes of public sharing articulated in the original 2013 OSTP “Increasing Access to the Results of Federally Funded Scientific Research” memorandum and communicate to researchers that data stewardship is not a priority. In most cases, placing a limit of 2 pages makes it impossible for reviewers to effectively evaluate the feasibility of the proposed plan. To realistically provide the level of detail needed to ensure proper management of research data, plans for most projects would need to be expanded. Similar to a Commercialization Plan, a well written data management plan should demonstrate in detail how these responsibilities will be met throughout and after the research project lifecycle. Once again, we recommend that the full data management plan be reviewed by individuals with expertise in planning, managing, publishing, and preserving research data.
- Data Type (pages 3 -4) - The descriptions provided for identified elements are clear and should help researchers understand the required content. We recommend adding a conditional requirement to this section, requesting a description for the amount of data, either for bytes or numbers of files/objects when expected to be over a certain threshold

(ex. >1TB or 1000 files). Researchers, particularly those new to the field, often underestimate the impact of size on their data management practices. Knowing at the beginning of a project when a grantee should arrange for additional storage or use automated metadata and organizing practices can assist reviewers in determining the feasibility of the plan, as well as help the researcher make the necessary arrangement before the project begins.

- In reviewing the provided Plan Elements, we noticed the absence of requirements regarding oversight. Implementing a data management plan will require some kind of human intervention and coordination. Identifying who will carry out these tasks helps ensure that they will be done, as well as allocating the needed resources or training.
- Scientific Data Archiving (page 6) - Assuming that data can be made accessible for extended periods of time without additional resources may not be realistic for most large datasets. Ongoing storage and management comes at a cost. Although it may not be possible to address this issue through these updated data sharing guidelines, we would encourage NIH to allocate resources to ensure that grantees have access to an adequate data preservation infrastructure.
- Compliance and Enforcement (page 6) - the proposal states, “NIH encourages the sharing of data for as long as it is useful to the scientific community.” This statement is vague and we recommend striking and replacing it with either a minimum retention period or a requirement to state specifically how long the grantee will share the data.
- NIH should consider explicitly stating that general websites (ex. personal or university department websites) do not have the necessary repository infrastructure to ensure long term access to research data and other grant outputs.

Submission #31**Date:** 11/02/2018**Name:** Dr. Ray Uzwyszyn**Name of Organization:** Texas State University Libraries**Type of Organization:** University**Role:** Institutional Official**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Scientific, Social Sciences and Humanities Data

I. The definition of Scientific Data

Scientific Data needs to be more broadly defined. Firstly, the data itself, but also the more important need of including robust metadata, paratextual material, field notes, methodologies, scientific procedures surrounding and for obtaining the data for reproducibility. Data needs to be broadly defined and with attention to best practices disciplinary metadata schemas.

II. The requirements for Data Management and Sharing Plans

For Data Management and Sharing Plans, researchers should be encouraged to work with their associated academic library for data sharing and this provision should be included in any data management plan. Sharing plans should also include lockdown and quality assurance procedures and recommendations as many researchers wish to change the data once it is in a repository and this should be discouraged to keep the integrity of the data and experiment for the future. Also, reproducibility and replicability should be encouraged with any data submission. It goes without saying that the hallmark of any scientific experiment should be the ability to reproduce the data by others and same results from the same experiment - the line between science and pseudoscience. Many researchers give excuses as to why their data can never be replicated and this should be discouraged and written into grant requirements to discourage fraudulent research. Also, the larger 'metadata' necessity of sharing involves furthering the course of knowledge through later accessibility (findability of the data) and synthesis by others - the scientific enterprise. Many researchers actually make their data though unusable to prevent competition. They are satisfying 'federal requirements' but they do not wish competition or others actually using their data to build on there work (publish other papers etc. on it) so they construct the data so that other scientists cannot use the data. This should be discouraged and usability, aggregation and sharing encouraged to advance the

scientific process, encourage collaboration and drive discoveries forward. This is making the case for consortial and shared repository to aggregate similar knowledge sets for comparison and synthesis

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and

NIH policies should be standardized and coordinated with other federal agencies so researchers are not left wondering with a huge selection of varying federal data management policies. NIH should work with the California Digital Library Digital Management Plan (DMP), the standard tool used by universities to simplify and make these timelines more standardized across federal agencies. Best disciplinary practices should be implemented with Metadata schemas for scientific disciplines and a smaller list of preferred schemas generated. Standards bodies should be maintained with all scientific data metadata standards. Finally, the larger problem of very large data sets and public sharing should be addressed perhaps on federal levels and funding made available or incorporated into grant standards/awards for long term data storage. Academic Libraries and universities are equipped for smaller to medium sized data sets but as research projects increase data sizes, storage becomes unmanageable for most academic libraries and university research computing centers. These types of questions should be addressed.

Attachment:

Coming of Age: The Online Research Data Repository

A Q&A with Ray Uzwyshyn

- By Mary Grush
- 12/13/16



"Online research data repositories are now pragmatic realities. Most discovery in the future will be predicated on the sharing and synthesis of data." - Ray Uzwyshyn

Technology advancements surround us, and sometimes the sheer volume of new tools and services is overwhelming. Can we identify which technologies are poised to make significant changes in the way we work? One technology that's been relatively under the radar may be about to make a huge difference in scholarly research practice and has the potential to help move scientific and social scientific discovery ahead as never before.

Here, CT asks Ray Uzwyshyn, the director of digital and collection services for the Texas State University library, about research data repositories - a technology that is just coming into its own. Uzwyshyn served on the implementation, planning, and policy committees for the Texas Data Repository, which launched December 2016. He offers both a current view of the technology and insight into its impact.

Mary Grush: What's the main objective of an online academic research data repository?

Ray Uzwyshyn: Research data repositories enable academic researchers to access, cite, and share data for a particular project- not just the final paper or project summary, but the actual data and paratextual material associated with it. They house both datasets and the material surrounding the data: field notes, documents, multimedia, and even specialized software used to process this data.

Most scholars today are working with an online community of colleagues who may be geographically dispersed around the globe. For them, the online data repository is becoming the storage application of choice - access speeds and software no longer present significant barriers to entry.

Grush: Why would an academic researcher choose to share their data via an online research data repository?

Uzwysbyn: There are great reasons to share research data, related to discovery and the promotion and progress of the scientific enterprise - especially at universities where most primary research on every topic imaginable is carried out.

Online data repositories allow easy global availability, sharing, and download. *Online research data repositories are now pragmatic realities. Most discovery in the future will be predicated on the sharing and synthesis of data.*

From data regarding the search for the cure of diseases to double checking conclusions for new scientific and social scientific discoveries, possibilities are manifold. It's the vast majority of research data being produced, for which online data repositories are best. Given that the research data isn't classified and will be personally de-identified, the situation currently is that there are specific federal mandates to make a researcher's data available and publicly accessible if they are receiving federal funding from any of the large U.S. granting agencies.

Grush: Beyond making datasets accessible, are researchers who use federal funds required to outline their data management plans?



Any large-grant-seeking researcher must now produce a Data Management Plan (DMP), especially when applying for large grants, (say NSF or NIH) and they must describe how they will make their data accessible. Applicants may wish to note that there is a good documentation and policy planning tool, the DMPTool (see DMPTool, <https://dmptool.org> and <https://dmptool.org/video>), available online through the California Digital Library, that helps researchers create their DMP. An online data repository to house a researcher's data is a central piece of that puzzle.



Grush: What are the important steps for research faculty - or for that matter, graduate students - as they create datasets and utilize a research data repository?

Uzwyshyn: All research data has an associated data repository life cycle path. Typically, researchers capture project data from experiments, instruments, surveys, and field work. They assign a disciplinary taxonomy, classification, or what we call a metadata schema to the data - essentially more data or description describing the primary data. This schema is key for the repository's search capabilities and later database search.

Ultimately, this classificatory work done properly allows effective searching across repositories so that datasets can be aggregated and harvested for later insight. Most researchers are looking for datasets similar to theirs: Do the datasets I found confirm my data? Can I use other researchers' results to build on my own experiments and data?

By searching research data repositories, researchers can also find examples of "negative data" or experiments that have failed - so they do not have to recreate the wheel and go down paths that previously have been dead ends or found to be unproductive. They can avoid duplicating such work that has already been done.

The final, and very important stage in the research data repository life cycle, is the long-term archiving and storage of datasets. This is both for the historical record and so that experiments won't be needlessly repeated. Basic research done today may not find its use value until twenty years hence. Because of this, it's important that the data be more transparently archived, stored, and kept accessible through file normalization and updates to software formats. This is especially true for time series data that tracks changes over time.

Grush: What types of data and data formats will you find in research data repositories?

Uzwyshyn: Most repositories are format-neutral and accept wide ranges of data formats. Of course, everyone knows Excel, but different disciplines also have their own specific data formats and preferences. For example, biochemistry may have specific data formats and software for data that the repository needs to accept.

There are also many specific types of data repositories: *project-specific*, *discipline-specific*, *institutional*, or *consortia/* data repositories. Project-specific repositories are project-oriented and typically contain research data created by a single faculty or a small team. Discipline-specific data repositories are usually subject-focused and

aggregate data from a certain discipline, say experiments surrounding nanotechnology- Purdue's [Nanohub](#) is an example. Institutions may also possess their own research data repository that goes across disciplines - this is an increasing trend. And finally, there are consortia! research data repositories. The [Texas Data Repository](#) (TDR), which launched in December 2016, is the first statewide academic consortia! repository. My institution, Texas State University, is one of the institutional repositories that make up the larger network within this consortia! repository.

Grush: How big are these various types of data repositories, and what size datasets do they accept?

Uzwysbyn: There is wide variation among repositories, depending on storage requirements and the sizes of datasets being gathered. The majority of online research data repositories for academic institutions accept what we might think of as regular or medium-sized datasets. These are typically of a size small enough to allow that the data may be housed right in the repository itself. A researcher or research group can upload their data from their desktop computer or research group server. To help you get your bearings on this question, for the Texas Data Repository each researcher may currently add as many files as they like up to 2GB in total, and research data groups within repositories may possess up to 10GB. These are very loose and flexible size limitations though, and it's safe to say they are constantly expanding and being re-evaluated in light of researchers' needs.

If there are larger datasets, the data repository might be considered a specialized project-specific or discipline-specific variant rather than institutional or consortia!. Very large datasets - with voluminous amounts of data, like the Seti project generates - might be more effectively treated with pointers from the repository to the actual storage places where the project data is housed. In such cases, the research data repository becomes a metadata repository - a place for data describing the data. Again, the value here is that the data repository enables researcher discovery, searchability, interoperability, and aggregation of datasets for further research.



Grush: What are the search capabilities in general? And what are a few of the benefits of searchability?

Uzwysghyn: Searchability is a primary value of research data repositories - the scholar is able to search across institutions, a consortium, or an entire discipline's experiments in specialized areas. Researchers can identify someone else's data that may help validate their findings; they can share data in partnerships with other researchers for new levels of discovery in their field; and they may aggregate or mash up data from various fields to create new knowledge and insight.

The concept of the research data mashup is similar to most software mashups, where, for example, one database provides GIS geospatial data, another provides real estate data, and a common field provides a link to combine disparate knowledge sets. By mashing these together, greater relational insight is achieved.

The analogy holds for scientific and social scientific experimentation through this linking field. This becomes especially interesting in linking datasets for disciplines that wouldn't normally "talk" to each other, academically speaking, but have commonality of one or more data fields. Research can then be synthesized, validated, or invalidated through the examination of a global scholarly community. Researchers can also gain insight from access to previously unavailable relevant datasets. It's probably also important to mention that data visualization technology becomes an important tool and infrastructure within the data repository ecology.

Grush: How can institutions approach all this? What kind of infrastructure do you need to provide, and what factors should you consider in choosing to build a repository? Do you have to build this from scratch?

Uzwysghyn: Today you'll find a variety of new solutions for housing and sharing your data, both open source and proprietary. A good data repository should have a permanent linking strategy - citation and access capabilities, typically with a Digital Object Identifier (DOI) or a Universal Numerical Fingerprint (UNF) that give the data a permanent location on the Internet. The repository could either be installed on a university server or hosted somewhere else, and a good solution will include administrative and collaborative options. Capacity for ingesting a wide range of data types, from Excel, to SPSS, to various discipline-specific data formats is also an important factor.

The number of good examples and models to investigate is increasing over time. Here are a couple links to what we did in Texas as we created the Texas Data Repository: <http://tinyurl.com/h36w93v> and <http://tinyurl.com/j5pcccz>.

Grush: What is the landscape now, for research data repositories? Is this a good time for institutions to think of getting "in" on this?

Uzwysghyn: The top research institutions in the U.S. have adopted, so the early adopters are all in. The early majority adoption is presently occurring and we're somewhere in the middle of this cycle. It's an excellent time to start thinking about adoption, especially if your institution has research faculty or aspires to be a research institution.

Even before selecting or implementing anything though, the best place to begin is with an environmental scan to examine your institution's needs. This means both polling your researchers and reviewing the state of current focused solutions. Harvard is very much behind a product called [Dataverse](#) - the product is flexible across disciplines and has its roots in the social sciences. Purdue University came out with a different orientation and originally advocated a more discipline-centered approach - their product, [HubZero](#), is used for more specific research interests, often in highly specialized technical scientific research areas.

Grush: Given that researchers in any field may be far-flung geographically, are there any global organizations that are pointing the way to help join data repositories together, promote better access to them, or help them interoperate via universal standards?

Uzwysghyn: In general, this idea of the research data repository is still fairly new and hasn't yet adopted official, universally accepted bodies of standards. But there are plenty of organizations beginning to think very seriously about data repositories, both present realities and emergent possibilities.

There are also emerging standards strongly in place for various specialized disciplinary metadata schema. This standardization of metadata schema ranges from the more general Dublin Core standard, to other, more specialized schema - for example, geophysical, life sciences, or astrophysics data. All of these enable interoperability.

Another organization advocating open data, SPARC, has just published an important [tool that helps researchers navigate federal data requirements](#) for their grants in the U.S.

Grush: When institutions take the plunge, if you will, and select a research data repository solution, is there any way they can plan for agility in the future, or even just have a reasonable exit strategy?

Uzwyshyn: Well, the evolving academic record of how research is being carried out today is rapidly changing, so the hard answer is that you must be thinking with both this moving target and a new generation of researchers in mind. You shouldn't be focusing on exit strategies these days, but rather thinking about evolutionary and developmental scenarios and those variables that will allow migration down the road.

To generalize, academics are not going to cease research efforts, and the possibilities for organizing, sharing, and housing research data have exponentially expanded through technology. You just need to keep your eyes open and more importantly, keep an open mind to the technological possibilities as the research data repository continues to evolve.



Grush: Are there other factors behind understanding and planning for a research data repository? What about planning for staff and the multiple roles that will be needed to build and support a repository?

Uzwyshyn: We've already mentioned several of the characteristics of research data repositories and what unique services and discovery advances they can bring to an institution. Beyond things we've highlighted specific to

this technology, other factors campus leaders should consider are more typical of any major technology initiative.

Behind the research data repository lie technical factors like emerging data and metadata standards, QC standards, and a range of technology issues; administrative, policy, and legal issues such as copyright and intellectual property; and outreach information, user education, and operational and service expectations. Challenges in implementing a research data repository will be similar to those you find with any important technology initiative. An institution should plan to leverage expertise from its previous technology implementation successes and be prepared for the research data repository to dialogue with various levels of the university campus.

Human resource expertise in research data repositories does also especially need to be developed over time. All research institutions will have to do something at some point, to plan for and create their repository infrastructure, and the need for staffing and staff expertise is a reality. Now is a very good time for leadership to begin considering staff roles, along with the related discussion of whether they want their institution at the back of the pack tomorrow or in the middle today.

Grush: What is the outlook for research data repository consortia? Can institutions gain advantage through consortia - maybe in their own region or even globally?

Uzwyshyn: Historically, most research collaborations among academic researchers have been more localized, with nearby universities, states in their region, or collegial institutional networks. Consortia! efforts increasingly allow researchers to enhance the possibilities opened by aggregated datasets, leveraging visibility and the expertise of colleagues both locally and globally. Sharing data globally leads to recognition, grant and project collaboration, and traction for new areas of investigation. These new paradigms have the power to move research ahead more quickly in the disciplines.

On technological levels, consortia also often build a community of technological human resource expertise regarding implementing and building technology offerings like research data repositories. Often these are state or interstate technology groups that can be very helpful in navigating the myriad of issues that will arise. Leveraging the cooperation of numerous institutions as a repository is created has many benefits.

Grush: Are researchers ready, in general, to share their research data more openly and work towards shared discovery? Are we looking ahead at good changes in research practice?

Uzwyshyn: There are great possibilities and I believe most researchers see or will see the value here. Currently our largest granting agencies (including NSF, NIH, or USDA) mandate and encourage data sharing processes, so the future is bright on pragmatic levels. Historically, the advancement of scientific discovery has been predicated on the sharing of knowledge and data. This is true from Newton to Einstein. As Newton put this, "If I have been able to see a little further, it is because I have been allowed to stand on the shoulders of Giants". The larger idea is that no researcher is working in a vacuum, but rather within a scholarly community with a past trajectory and a forward telos. Because of this, I'm encouraged by these new technological possibilities for organization and sharing from this great ocean of data opening before us. Our software infrastructures are now able to allow this next renaissance of discovery, enabling new insights and synthesis from current results. Hopefully this will also allow a few of those next intrepid explorers to stand on the shoulders of giants.

About the Author

Mary Grush is Editor and Conference Program Director, Campus Technology.



Submission #32**Date:** 11/04/2018**Name:** Hunter N.B. Moseley**Name of Organization:** University of Kentucky**Type of Organization:** University**Role:** Scientific Researcher**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

bioinformatics and systems biology

I. The definition of Scientific Data

Software as “scientific data” should be mentioned. There are special considerations with the sharing of software, especially open access and methods for broad dissemination through software repositories.

II. The requirements for Data Management and Sharing Plans

The FAIR principles have a catchy acronym, but the four principles are incomplete, lacking an emphasis on scientific rigor and reproducibility. Thus, making the FAIR principles the core of NIH data management and sharing policy without listing additional core principles will likely lead to an incomplete policy. The definition of “scientific data” appears to capture some of the concepts of rigor and reproducibility, but stands separate from the FAIR principles, which emphasize re-usability. Therefore, the principles of re-usability and reproducibility both need to be clearly defined and co-supported by the new data management and sharing policy. These two principles are distinct and require different approaches to their implementation, especially in the context of growing data resources.

Do not put a two-page limit on the data management and sharing plan. This may be inadequate for certain proposals that focus on data management and analysis, especially involving the integration of large heterogeneous datasets.

The description of “standards” is limited. Neither the use of ontologies nor structured data repository formats are directly mentioned or encouraged. Both should be mentioned and encouraged if common data elements are going to be directly mentioned and encouraged.

Submission #33**Date:** 11/05/2018**Name:** Lyle G. Best**Name of Organization:** Missouri Breaks Industries Research Inc**Type of Organization:** Other**Other Type of Organization:** American Indian owned, SBA-certified HUBZone small business**Role:** Scientific Researcher**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Epidemiology and population genetics translated into improved clinical care and public health interventions.

I. The definition of Scientific Data

No comment in this area.

II. The requirements for Data Management and Sharing Plans

Dear Sirs:

I appreciate the opportunity to respond to:

Request for Information (RFI) on Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research

Notice Number: NOT-OD-19-014

As a physician that has provided care for American Indian patients for over 30 years, both in the Indian Health Service and the private sector, I am very much aware of the tremendous health disparities present in these communities. As a biomedical investigator having served as PI for 3 different NIH funded studies over the past two decades, I am also very much aware of the potential contribution of biomedical research to improved health of these populations. Biomedical research requires cooperation between the investigator, the participant and the community at large. This cooperation is fragile and easily disrupted. For this reason it is critical that NIH come to grips with the legitimate concerns of American Indian tribes regarding the sharing of data. They have been very generous in the past; but what appears to some as imperious "requirements" do not come off well and can jeopardize this tradition.

I am in agreement with the data-sharing rationale and have worked hard to point out the benefits of this to my colleagues and friends in Indian communities. I accept the broad outline of this policy; but have some suggestions and concerns:

1) The policy seems all-encompassing in application and would appear to apply to even very small studies with data of a very subjective nature (eg interview transcripts). This would necessitate a very large expansion in database storage to rather little purpose. I feel the previous genetic data-sharing policy provides an example of a more limited, rational constraint, wherein the policy was limited to data from one gene on over 1,000 participants, etc etc.

2) As a consequence of above, the cost of storing this greatly expanded collection of datasets will inevitably increase substantially and at some point the currently free, or low cost, repositories will begin to expect compensation or even a profit for their efforts.

3) the suggestion on page 3 " Extramural Grants: Plans could be evaluated as an Additional Review Consideration, i.e., evaluated as acceptable or unacceptable by reviewers, but not be factored into the overall impact score through the peer review process." In my opinion, allowing the scientific reviewers to determine the "acceptability" of a data-sharing plan is not wise. These reviewers are a constantly shifting group of non-NIH employees that frequently have very incomplete and often flawed understanding of tribal sovereignty and governance, not to mention the local infrastructure. I think it is much more likely that NIH will provide consistent and accurate interpretation of this policy by giving this authority solely to the NIH program officers or others familiar with this area.

Sincerely,

Lyle G. Best, MD

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and

No comment.

Submission #34

Date: 11/07/2018

Name: Bryant thomas Karras MD

Name of Organization: State of Washington

Type of Organization: Government Agency

Role: Government Official

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

Epidemiology, Public Health

Submission #35

Date: 11/07/2018

Name: Shelley Cole

Name of Organization: Texas Biomed

Type of Organization: Nonprofit Research Organization

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

Genetics and genetic risk factors for metabolic disease in understudied population groups.

II. The requirements for Data Management and Sharing Plans

The policy suggestion regarding plan review for extramural grants that proposes that the plans be evaluated as acceptable or unacceptable by peer reviewers, even when this would not be factored into the score, is very concerning to me. First, it is highly unlikely that any given set of reviewers would have the background and knowledge to judge the adequacy of a data sharing plan from diverse sources and for diverse data sets especially if the plan includes specific sharing restrictions or sharing is not possible because of licensing, ownership, or other barriers (e.g. tribal sovereignty, data from sensitive populations, etc.). Secondly, peer reviewers are already overburdened with a long list of administrative review items during peer review. Adding to this burden would not benefit the peer review process nor provide adequate overview of data sharing plans.

Submission #36

Date: 11/08/2018

Name: Anonymous

Name of Organization:

Type of Organization: Other

Other Type of Organization: Hospital

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

my research team focuses on surgical oncology

II. The requirements for Data Management and Sharing Plans

I would perceive difficulties in communicating this sharing plan to clinical trial participants, if this is a requirement, some individuals may choose to decline a trial due to this increased data sharing. Variability in choice of software/code may reduce ability of multiples groups to use the shared data. Care will be required to determine similarities and differences in data abstraction processes for each study in order to determine if results should be combined.

Submission #37

Date: 11/09/2018

Name: Qingling Sun

Name of Organization: Sun Technologies & Services, LLC

Type of Organization: Other

Other Type of Organization: LLC

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

Drugs/Medicine for Metastatic Breast Cancer to Different Organs of the Body

I. The definition of Scientific Data

The Quantitative Data on Drugs/Medicine on Metastatic Breast Cancer to Different Organs of the Body, including those FDA have approved and is currently in trial

II. The requirements for Data Management and Sharing Plans

Make these data available to public research to improve the treatment of metastatic breast cancer

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and

Make the data on metastatic breast cancer drugs available to public research as soon as possible. If phase is adopted, at phase I, publish the data on drugs name, function, preliminary data, efficiency to the metastasis to what organ. This may facilitate the research to find the most effective drug for metastatic breast cancer.

Submission #38**Date:** 11/11/2018**Name:** Clark C. Evans**Name of Organization:** Prometheus Research, LLC**Type of Organization:** Other**Other Type of Organization:** Software and Informatics Provider**Role:** Member of the Public**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

medical informatics, the application of technology to address medical knowledge challenges

I. The definition of Scientific Data

Currently, it's not super clear where there is a boundary between "data" and "code". If the NIH makes rules for data, but not code, then I'm concerned those acquiring grants would be ungenerous with what falls outside the definition. Hence, I'd change the viewport to discuss "Scientific Information", and I'd expressly include software source code, neural network training data, terminologies, instruments, and system configuration as well as results. For reproducible science, it's important that we have every bit of information necessary to reproduce a study.

II. The requirements for Data Management and Sharing Plans

Sharing plans should permit (a) not only access to all materials, but the (b) ability to make and publish derivative works. If the materials are to be licensed, it should be non-discriminatory and have broad exceptions for fair-use, such as refuting a study or producing derivative that has seemingly minor changes but comes to a contradictory, if surprising result. It should be impossible to share results, or even compiled objects which generate results, without also sharing the source code (which includes training data for a neural network, or a terminology, etc)

Submission #39**Date:** 11/13/2018**Name:** Alexander Tsai**Name of Organization:** Harvard Medical School**Type of Organization:** University**Role:** Scientific Researcher**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Epidemiology, Anthropology

I. The definition of Scientific Data

Any NIH policy should provide clarification about how the policy accommodates sharing of qualitative data. The data collected in qualitative studies are typically obtained through in-depth interviews, focus groups, direct observation, document review, and audio recording review. These data, while typically not aimed at establishing generalizability, lend themselves to generating new theoretical insights about certain phenomena in greater depth and detail than is possible through quantitative designs (Patton, 2002). Given the inherently intersubjective nature of qualitative data collection, the iterative nature of qualitative data analysis, and the unique importance of interpretation as part of the core contribution of qualitative work, sharing data for the purposes of verification is likely to be impossible in the setting of qualitative research. First, some scholars have argued that interview transcripts, even when accompanied by detailed field notes, cannot represent with sufficient fidelity the actual interview that took place. According to this understanding, we should question the extent to which interview transcripts may be considered “raw data” for external investigators to use in the same manner as a dataset taken from a randomized controlled trial of the latest unoriginal antidepressant medication. Second, the interview transcripts disseminated to external investigators are unlikely to be the data they would have collected had they conducted the study themselves. A qualitative study guided by the method of grounded theory, for example, follows an inductive process with concurrent review of the data being collected, filtering of the data for relevance and meaningfulness, and grouping and naming of patterns observed in the data (Glaser and Strauss, 1967). Even if the authors of a particular study uploaded the entire set of field notes or interview transcripts to a secure data repository, what do these data mean to an external investigator who might not have the same kinds of embedded cultural experiences

(that would help contextualize the interview and field observation data) and who would have collected the data differently?

II. The requirements for Data Management and Sharing Plans

There are a number of challenges that could hamper the implementation of data sharing policies for qualitative data. Most qualitative researchers use respondent validation (e.g., reviewing emerging themes and analyses with study participants or key informants) to ensure rigor, and the practice is highlighted as a key process component of qualitative research in most reporting checklists (Clark, 2003; O'Brien et al., 2014; Tong et al., 2007). This method of data sharing through member-checking of interim findings is carefully supervised. In contrast, data sharing policies that make interview transcripts available to study participants in a completely unstructured fashion may have negative effects. Because qualitative study designs often lend themselves to the in-depth study of highly sensitive subject material (Kelly et al., 2011; King et al., 2013; Parkinson, 2013; Wade et al., 2005), field notes and interview transcripts would need to be anonymized prior to dissemination in order to conform with prevailing legal and ethical guidelines. Because interview transcripts contain verbatim quotations, it is likely that some transcripts cannot be sufficiently anonymized to prevent deductive disclosure, or what Tolich (2004) has called violations of “internal confidentiality.” To minimize the risk of deductive disclosure, a data sharing policy might, in lieu of obliging the release of interview transcripts, require investigators to implement procedures to enhance transparency. Many aspects of the qualitative analysis (e.g., transcription rules, data segmentation, coding units, process for code development, finalized codes) could be shared with minimal risk to study participants. Taking transparency a step further, investigators could export coding queries and make these available to external investigators. Because coding queries consist of excerpted and possibly disembodied interview text, they may offer greater anonymity compared with full transcripts. Depending on the interview content, investigators may still need to redact some of the text to preserve anonymity – which would entail added burden – but the risk of deductive disclosures would be reduced. The release of coding queries has not been suggested in the ongoing conversation on data sharing in qualitative research but should be regarded as a viable and potentially more ethical way to promote transparency than the release of full transcripts. In addition to the risk of deductive disclosures, a number of other unintended consequences could result from data sharing policies if they are not properly tailored to the unique aspects of qualitative and mixed methods research. For example, the burden of organizing qualitative data for inspection or use by external investigators could easily exceed the work of writing the manuscript itself. These and other concerns are spelled out in Tsai et al., *Social Science & Medicine* 169 (2016) 191-198, included as an attachment.

Attachment:



Contents lists available at ScienceDirect

Social Science & Medicine

journal homepage: www.elsevier.com/locate/socscimed

Promises and pitfalls of data sharing in qualitative research



Alexander C. Tsai ^{a, b, c, *}, Brandon A. Kohrt ^d, Lynn T. Matthews ^a, Theresa S. Betancourt ^{b, e},
Jooyoung K. Lee ^f, Andrew V. Papachristos ^g, Sheri D. Weiser ^h, Shari L. Dworkin ⁱ

^a Chester M. Pierce, MD Division of Global Psychiatry, Massachusetts General Hospital, Boston, USA

^b Harvard Center for Population and Development Studies, Cambridge, USA

^c Mbarara University of Science and Technology, Mbarara, Uganda

^d Duke Global Health Institute, Duke University, Durham, USA

^e Department of Global Health and Population, Harvard T.H. Chan School of Public Health, Boston, USA

^f Department of Sociology, University of Toronto, Toronto, Canada

^g Department of Sociology, Yale University, New Haven, USA

^h Department of Medicine, University of California at San Francisco, USA

ⁱ Department of Social and Behavioral Sciences, School of Nursing, University of California at San Francisco, San Francisco, USA

a r t i c l e i n f o

Article history:

Received 20 May 2016

Received in revised form

30 July 2016

Accepted 2 August 2016

Available online 9 August 2016

Keywords:

Confidentiality

Data sharing

Ethnography

Mixed methods

Qualitative research

Reproducibility

Transparency

a b s t r a c t

The movement for research transparency has gained irresistible momentum over the past decade. Although qualitative research is rarely published in the high-impact journals that have adopted, or are most likely to adopt, data sharing policies, qualitative researchers who publish work in these and similar venues will likely encounter questions about data sharing within the next few years. The fundamental ways in which qualitative and quantitative data differ should be considered when assessing the extent to which qualitative and mixed methods researchers should be expected to adhere to data sharing policies developed with quantitative studies in mind. We outline several of the most critical concerns below, while also suggesting possible modifications that may help to reduce the probability of unintended adverse consequences and to ensure that the sharing of qualitative data is consistent with ethical standards in research.

© 2016 Elsevier Ltd. All rights reserved.

1. Introduction

In 2014, the Public Library of Science (PLOS) journals unveiled a policy stipulating that authors must make available all data underlying the findings described in their published manuscript (Bloom et al., 2014). The implementation of this new policy was something of a watershed moment; although *PLOS Medicine* was not the first high-impact medical journal to require data sharing as a matter of policy, it is the only one that routinely publishes findings from qualitative studies and qualitative meta-syntheses. While the new guidance permits authors some latitude in circumventing data sharing, in some ways it does resemble the obligatory and much more rigorous conditions of publication already in place at leading journals in biostatistics (Peng, 2009), economics

(Ashenfelter et al., 1986; Bernanke, 2004), and political science (Meier, 1995). At the *American Economic Review*, for example, authors make publicly available the raw data and statistical programming code needed to reproduce all of the findings in the published manuscript, and these materials are uploaded to the journal web site prior to publication (Bernanke, 2004). The experiences in these fields suggest that leading journals can implement unilateral changes that eventually contribute to building a culture in which data sharing becomes the norm.

The movement to promote reproducible research in the medical and public health literature has lagged, perhaps for myriad reasons. First, concerns are frequently voiced about intellectual property protections and/or the potential hazard of disclosing protected health information (Hrynaskiewicz et al., 2010; Mello et al., 2013; Tudur Smith et al., 2015). Second, because medical and public health research can often carry enormous financial implications for specific products (Rennie, 1997; Shuchman, 2005) or entire industries (Kaiser, 1997; Michaels and Monforton, 2005; Muggli et al.,

* Corresponding author. Massachusetts General Hospital, MGH Global Health, 125 Nashua Street, Ste. 722, Boston, MA 02114, USA.

E-mail address: actsai@partners.org (A.C. Tsai).

2001) that are implicated in the findings, requests for data may be driven by financial motivations that extend well beyond any disinterested concerns about science for science's sake. A researcher might be appropriately wary, for example, of responding to an industry representative's seemingly benign request for data. Finally, there are also structural barriers to data sharing, because faculty members at schools of medicine and public health are incentivized to publish secondary findings from a given data collection effort. For example, it is not uncommon for investigators to publish secondary analyses of data from randomized trials (Rotheram-Borus et al., 2015; Tsai et al., 2016) or multiple analyses of data from the same cohort (Colditz and Hankinson, 2005; Colditz et al., 1997). These concerns apply less strongly in the social sciences. Yet because this type of research often has direct relevance for patient care, data sharing should (in general) be regarded as an imperative for ensuring transparent analysis of data and reproducibility of research findings (Doshi et al. 2012; Le Noury et al., 2015).

The movement for research transparency has gained irresistible momentum over the past decade (Groves, 2010; Hanson et al., 2011; Laine et al., 2007; Miguel et al., 2014; Nosek et al., 2015; Peng et al., 2006; PLOS Medicine Editors, 2014; Stodden et al., 2013; Tsai, 2011). Although qualitative research is rarely published in the more high-impact journals (Greenhalgh et al., 2016; Shuval et al., 2011) that have adopted, or are most likely to adopt, data sharing policies, qualitative and mixed methods researchers who publish work in these and similar venues will likely encounter questions about data sharing within the years ahead, especially as mixed methods studies integrating qualitative and quantitative data become increasingly prominent (Creswell et al., 2011). The substantive ways in which qualitative and quantitative data differ should be considered when assessing the extent to which qualitative and mixed methods researchers should be expected to adhere to data sharing policies developed with purely quantitative studies in mind. We outline several of the most critical concerns below, while also suggesting possible modifications that may help to reduce the probability of unintended adverse consequences and to ensure that the sharing of qualitative data is consistent with ethical standards in research.

2. Reliability, validity, and reproducibility in qualitative research

2.1. Unique features of qualitative data production and analysis

Qualitative studies are based on data that are fundamentally different from the data collected in other observational study designs. The standardized measures employed in quantitative studies constrict the diverse perspectives of study participants along pre-determined continua (e.g. categorical or continuous) so that they can be statistically aggregated. Quantitative data analysis plans (Olken, 2015) and study protocols (Horton, 1997) can be pre-specified and disseminated. The data can be anonymized and uploaded to secure data repositories. The statistical code used to process the data, and the process through which the output is translated into the manuscript text and tables, can just as easily be shared and replicated (Gandrud, 2013; Peng, 2009; Stodden et al., 2014; Vickers, 2006). External investigators can then use the electronic paper trail to verify the published findings (Dewald et al., 1986; Jefferson and Doshi, 2014; Le Noury et al., 2015; McCullough and Vinod, 2003). Data sharing, in effect, is "a threat that might keep potential cheaters honest" (p.722) (Hamermesh, 2007).

In contrast, the data collected in qualitative studies are typically obtained through in-depth interviews, focus groups, direct

observation, document review, and audio recording review. These data, while typically not aimed at establishing generalizability, lend themselves to generating new theoretical insights about certain phenomena in greater depth and detail than is possible through quantitative designs (Patton, 2002). While complementary to other forms of social measurement, these data are also neither collected nor analyzed in as linear a manner, and it has been argued that the concept of reliability does not directly translate from the quantitative (rationalistic) to the qualitative (naturalistic) paradigm (Guba and Lincoln, 1981). In her influential essay, Stenbacka (2001) goes so far as to argue, "It is obvious that reliability has no relevance in qualitative research ... If a qualitative study is discussed with reliability as a criterion, the consequence is rather that the study is no good" (p.552). The extremity of her viewpoint notwithstanding, more recent work in the field has sought to address questions about the validity and reliability of qualitative research findings, through the use of descriptive approaches (e.g., verification strategies (Morse et al., 2002)), quantitative approaches (e.g., calculating inter-rater reliability for comparing the assessments of multiple coders (Cohen, 1960) or proportional reduction in loss (Rust and Cooil, 1994)), and reporting checklists (Clark, 2003; O'Brien et al., 2014; Tong et al., 2007).

2.2. Reproducible research and qualitative data

For most readers of high-impact medical and public health journals, the term "reproducibility" will evoke the idea that external investigators ought to be able to arrive at the same published findings when given the data and analysis code (Claerbout and Karrenbach, 1992; King, 1995). In Clemens' (in press) recently published typology of replication and robustness, this particular type of check is described as but one form of "replication" and given the label "verification": "ensuring that the exact statistical analysis reported in the original paper gives materially the same results reported in the paper, either using the original dataset or remeasuring with identical methods the same traits of the same sample of subjects." This definition corresponds closely to the concept of "methods reproducibility" suggested by Goodman et al. (2016). Notably, other researchers have ignored the distinction between "replication" and "reproducibility." For example, the Open Science Collaboration (2012) have written: "Some distinguish between 'reproducibility' and 'replicability' by treating the former as a narrower case of the latter (e.g., computational sciences) or vice versa (e.g., biological sciences). We ignore the distinction" (p.659).

Verification does not translate well to a data sharing policy for qualitative studies. Given the inherently intersubjective nature of qualitative data collection, the iterative nature of qualitative data analysis, and the unique importance of interpretation as part of the core contribution of qualitative work, verification is likely to be impossible in the setting of qualitative research. We discuss two principal reasons below.

First, some scholars have argued that interview transcripts, even when accompanied by detailed field notes, cannot represent with sufficient fidelity the actual interview that took place. Even audio and video recordings, which are generally considered the most complete observational data that can be captured, cannot convey valuable tactile and/or olfactory data obtained in the field (Bernard and Ryan, 2009). Drawing on focus groups conducted with qualitative researchers, Broom et al. (2009) showed that many of them were of the immoderate opinion that their transcript data were "an encoded account only decipherable to the individual who collected it" (p.1170). According to this understanding, we should question the extent to which interview transcripts may be considered "raw data" for external investigators to use in the same manner as a dataset taken from a

randomized controlled trial of the latest unoriginal antidepressant medication.

Second, the interview transcripts disseminated to external investigators are unlikely to be the data they would have collected had they conducted the study themselves. A qualitative study guided by the method of grounded theory, for example, follows an inductive process with concurrent review of the data being collected, filtering of the data for relevance and meaningfulness, and grouping and naming of patterns observed in the data (Glaser and Strauss, 1967). Investigators may also choose to collect additional data, if necessary, to deepen understanding into emerging phenomena via “theoretical sampling” (Glaser, 1978). Even if the authors of a particular study uploaded the entire set of field notes or interview transcripts to a secure data repository, what do these data mean to an external investigator who might not have the same kinds of embedded cultural experiences (that would help contextualize the interview and field observation data) and who would have collected the data differently? An external investigator conducting a secondary analysis of a grounded theory dataset must be aware that, even if the same research questions are considered at the outset, s/he likely would have made very different decisions during the course of the study that would have led to an entirely different dataset being constructed. If external investigators perceive there to be gaps in the dataset they are provided by the study authors, the potential explanations for the missing data are legion: are data missing because the concepts of interest occurred too infrequently to be meaningful to the initial guiding propositions, because the phenomena were simply not present in the sample, or because the study authors’ interview probes were driven by a different conceptual lens?

Some researchers might view these unique features of qualitative modes of inquiry as befitting their position in the conventional “hierarchy” of evidence (Atkins et al., 2004; Guyatt et al., 1995). The economist Amitabh Chandra has quipped, for example, “If ethnography is a legitimate way to learn things ... why aren't [pharmaceutical] manufacturers allowed to do it?” (Chandra, 2015) Yet even quantitative data are subject to what Goodman et al. (2016) have labeled as “inferential reproducibility”: “... scientists might draw the same conclusions from different sets of studies and data or could draw different conclusions from the same original data, sometimes even if they agree on the analytical results” (p.4). Furthermore, it is important to note that secondary analyses of qualitative data *would* likely be able to reproduce at least some, if not all, of the major themes identified in the primary published article. However, that is not the aim of a verification test – which is, rather, to reproduce “materially the same results reported in the paper” (Clemens, in press). Given these difficulties, it is likely that external qualitative investigators would not seek “verification” but rather “reproduction,” defined by Clemens as being another form of replication similar to verification except that reproduction studies are conducted with a different sample of study participants from the same population. This definition corresponds to the concept of “results reproducibility” suggested by Goodman et al. (2016). For example, Lewis' (1951) re-study of the Mexican village Tezapotlan 20 years after Redfield (1930) might be considered, had it been conducted somewhat earlier, a reproduction test of a qualitative study. In theory, reproduction of a qualitative study does not require a data sharing policy. The authors' description of the study's methods, especially if guided by a reporting checklist (Clark, 2003; O'Brien et al., 2014; Tong et al., 2007), should be sufficient to enable another team of investigators to conduct a reproduction test. But if reproduction, rather than verification, is the goal, then of what relevance is a data sharing policy?

3. Data sharing in qualitative research

Beyond attempts to increase transparency in the production of qualitative data, it is likely that qualitative and mixed methods researchers will need to address qualitative data sharing in some fashion. Applying these standards uncritically, one might presume that data sharing involves providing the following in an [online supplementary appendix](#): interview guides and interview transcripts, in the original language and in the translated language of the investigators (if different from the original); field notes; data used, if any, to establish inter-coder reliability; full code books; and documents, if any, describing the process of open coding, selection of codes for inclusion in the final codebook, and category construction. The “audit trail” supports reliability and validity, so even if it is recognized that no two groups would conduct identical qualitative studies, the information available to external investigators would enable them to understand how the study authors arrived at the published conclusions. Most computer-assisted qualitative data analysis software packages offer export functions that enable users to save an entire “project” (e.g., raw data, codebook, coding links, and memos), which could facilitate dissemination. While these types of maneuvers might be consistent with a data sharing policy, there are a number of challenges that could hamper their implementation in practice. Below we highlight the most significant challenges facing data sharing in qualitative research.

3.1. Preserving the anonymity or pseudonymity of study participants

Data sharing policies should carefully consider the potential effects of data sharing on study participants. Most qualitative researchers use respondent validation (e.g., reviewing emerging themes and analyses with study participants or key informants) to ensure rigor, and the practice is highlighted as a key process component of qualitative research in most reporting checklists (Clark, 2003; O'Brien et al., 2014; Tong et al., 2007). This method of data sharing through member-checking of interim findings is carefully supervised. In contrast, data sharing policies that make interview transcripts available to study participants in a completely unstructured fashion may have negative effects. Chief among these are the potential psychosocial consequences of compromising study participant anonymity.

Because qualitative study designs often lend themselves to the in-depth study of highly sensitive subject material (Kelly et al., 2011; King et al., 2013; Parkinson, 2013; Wade et al., 2005), field notes and interview transcripts would need to be anonymized prior to dissemination in order to conform with prevailing legal and ethical guidelines. Institutional Review Board concerns about participant anonymity, discussed in the PLOS policy (Bloom et al., 2014), have been identified as a leading barrier to data sharing. Consequently, investigators lacking proper guidance on how to comply with data sharing guidelines in a way that provides adequate anonymity protections may simply default to data withholding. For example, in the Data Availability Statement for their qualitative study recently published in *PLOS Medicine*, Christopoulos et al. (2015) stated, “Public availability of data could potentially compromise participant privacy. Participants did not consent to have their full transcripts or excerpts of transcripts made publicly [*sic*] available.” Qualitative studies published in *PLOS One* subsequent to the PLOS policy adoption have made similar claims (Natoli et al., 2015; Tang et al., 2015) (although there have also been notable, and welcome, exceptions (Lo et al., 2016)).

While Institutional Review Board restrictions are commonly cited to justify withholding of quantitative data (Campbell et al.,

2002), in fact it may be possible to release de-identified versions of transcripts that preserve the anonymity of qualitative study participants. The nature of any anonymization procedures would depend on the nature of the data collected and the extent to which the data can be linked with publicly available information to reveal specific identities. At a minimum, the anonymization procedures would entail redaction or alteration of protected health information and any specific encounter details that reveal, however indirectly, the identity of any of the parties to the encounter, with obfuscated information shown in brackets. The investigator might keep a detailed record of these procedures in a secure location should it become necessary to revisit the data after publication (Table 1), similar to the recommendations made in the Privacy Certificate Guidance of the U.S. National Institute of Justice (2007). As a cautionary note, depending on the size of the dataset, the redaction or anonymization process could require tremendous time and effort of the investigators and could also potentially introduce errors and inconsistencies (Goffman, 2014; Lewis-Kraus, 2016). Additionally, for some studies, the nature of the research (Parkinson, 2013) may be such that any suitably redacted or anonymized transcripts might be so unserviceably thin that they would be devoid of meaningful content. Wolcott (1973) discusses this possibility in the introduction of his classic ethnography: "To present the material in such a way that even the people central to the study are 'fooled' by it is to risk removing those very aspects that make it vital, unique, believable, and at times painfully personal" (p. 4).

Because interview transcripts contain verbatim quotations, it is likely that some transcripts cannot be sufficiently anonymized to prevent deductive disclosure, or what Tolich (2004) has called violations of "internal confidentiality." That is, study participants could recognize themselves, their communities, or other study participants (if they belong to the same community) (Larossa et al., 1981). van den Hoonaard (2003) holds that anonymity is "a virtual impossibility in ethnographic research" (p.141). Depending on the sensitivity of the subject matter, deductive disclosure could result in harm to study participants and their relationships with others in the community. Ellis (1995), Schepher-Hughes (2000), and Stein (2010) have famously written about being angrily received by study participants over deductive disclosures following the publication of their celebrated books (Ellis, 1986; Schepher-Hughes, 1977; Stein, 2001). If such aggravated harm could result from the publication of books and journal articles in which verbatim quotations are carefully curated, one can imagine the harm resulting from a data sharing policy requiring entire interview transcripts to be shared.

Certain types of studies may carry even greater risks of deductive disclosure. These include studies of small-scale societies;

studies that rely on respondent-driven sampling and other variations of snowball sampling to identify hard-to-reach populations; and studies in which permission to access a small community must be first secured from highly networked research gatekeepers, such as village leaders or community advisory boards. In these settings, a minor, idiosyncratic detail – such as a manner of speaking or a specific phrase – that is of unknown significance to the investigator (and therefore likely to go unredacted) could result in deductive disclosure and potential harm. In addition to the risk of harm to study participants, deductive disclosure also raises important questions about potential risks to third-party non-participants when study participants disclose sensitive information about social network ties that arises from their shared history with others (Larossa et al., 1981; Lounsbury et al., 2007; McLellan et al., 2003).

Related to the above, data sharing potentially further limits qualitative research done through "studying-up" (Nader, 1969) or "studying over" (Markowitz, 2001) – approaches in which persons in positions of power (e.g., hospital administrators, pharmaceutical company executives, heads of governmental or multilateral organizations) become the subject of ethnographic study (Abramowitz and Panter-Brick, 2015; Closser, 2010). Because elites are more empowered to articulate concerns about confidentiality and disclosure, data sharing could unintentionally perpetuate power differentials in which health program beneficiaries endure as research subjects while health program funders and implementers remain understudied (Schneider and Aguiar, 2012).

Given the greater risks of deductive disclosure through unregulated data sharing (as contrasted with the carefully curated release of specific quotations through publication of study findings), consent documents for qualitative studies would need to properly inform prospective study participants that the interview transcripts could potentially be uploaded to a shared data repository for public consumption. Even researchers who have no intentions to share the data might be advised to seek informed consent from study participants at the outset simply to preserve the option in the future (Groves, 2010). Although study participants' exposure to such risk would ultimately be contingent on the researchers' decision to publish their findings in a journal where a data sharing policy is enforced, it is likely that such a caveat – however conditional – would result in selection on unobserved heterogeneity. These selective pressures could shape the types of persons who agree to participate in qualitative and mixed methods studies; alternatively, these selective pressures could have no impact on the types of persons who agree to participate but could shape the nature of the data they are willing to share with investigators. Either of these selective pressures would likely compromise the quality of the research, thereby upending one of the distinctive advantages of qualitative research, which is the ability to conduct in-depth

Table 1
Supplementary information table of procedures taken to anonymize or redact hypothetical interview transcripts prior to dissemination in a data repository (N = 53).

Study participant	Line number	Original	Anonymized
Clinic patient 2	79	"My husband has been beating me regularly since I was married to him at age 18"	"My husband has been beating me regularly since I was married to him at [a young age]"
Clinic patient 2	85	"Just the other day he got angry with me because there was no water and our eldest went to school in a soiled uniform. He threw the empty jerricans at me and you now see the bruise on my left eye"	"[] He got angry with me because there was no water []. He [attacked me] and you now see [my face]"
Community member 8	243	"I am the headmaster of the Buhingo Boarding School. What would the parents say if they knew I was HIV positive?"	"I am the headmaster of [a school]. What would the parents say if they knew I was HIV positive?"
...			
Clinic patient 53	164	"I was in the hospital for a week after injuring my left leg in a <i>boda boda</i> accident. The nurse at the Mbarara Hospital chastised me when she found out my HIV status."	"I was in the hospital [after a transportation accident]. The nurse [] chastised me when she found out my HIV status."

examinations of sensitive subject material (Kelly et al., 2011; King et al., 2013; Parkinson, 2013; Wade et al., 2005).

To minimize the risk of deductive disclosure, a data sharing policy might, in lieu of obliging the release of interview transcripts, require investigators to implement procedures to enhance transparency. Many aspects of the qualitative analysis (e.g., transcription rules, data segmentation, coding units, process for code development, finalized codes) could be shared with minimal risk to study participants. Taking transparency a step further, investigators could export coding queries and make these available to external investigators. Because coding queries consist of excerpted and possibly disembodied interview text, they may offer greater anonymity compared with full transcripts. Depending on the interview content, investigators may still need to redact some of the text to preserve anonymity – which would entail added burden – but the risk of deductive disclosures would be reduced. The release of coding queries has not been suggested in the ongoing conversation on data sharing in qualitative research but should be regarded as a viable and potentially more ethical way to promote transparency than the release of full transcripts.

An example of a coding query, applied to data from Kohrt et al. (2010) and Morley and Kohrt (2013), is provided in the [Electronic Supplementary Appendix](#). Coding queries would provide external investigators with comprehensive information that could be used to qualitatively assess the internal coherence of the coding scheme (Box 1). In qualitative research, study participants often present conflicting or contradictory views on the same topic based on varying influences such as the nature of the question and the time elapsed during the interview (LeCompte and Schensul, 1999). Discrepant data may be especially important in longitudinal studies where study participants provide serial interviews during the course of an illness or throughout their lifetimes, thereby gaining increasing familiarity with a particular interviewer. These processes are rarely, but with some exceptions (Groleau et al., 2006), captured in academic publications that tend to present views as static and internally coherent. Ultimately, much like the sharing of data from quantitative studies can provide opportunities to conduct detailed interrogations of the scientific record (Le Noury et al., 2015), coding queries can help reviewers and external investigators assess whether the quotes provided in manuscripts and journal articles capture the overall content of the data or whether they represent selective reporting of study participants' perspectives in a way that suits the authors' theses.

3.2. Other unintended consequences of qualitative data sharing

In addition to the risk of deductive disclosures, a number of other unintended consequences could result from data sharing policies if they are not properly tailored to the unique aspects of qualitative and mixed methods research. First, the burden of organizing qualitative data for inspection or use by external investigators could easily exceed the work of writing the manuscript itself. How should the interests of research transparency be weighed against the potential costs of documentation burden? Redacting the hundreds of pages of transcripts collected during the course of a small qualitative study would require months of work. Moreover, there are no standards in the field for systematically documenting the hours of conversations, conference calls, and e-mail exchanges required for code selection and category construction. Guidelines would need to be developed so that documentation of these procedures is uniform across studies. Larger qualitative and mixed methods studies would entail an even greater documentation burden. For example, the longitudinal qualitative study by Maman et al. (2014) involved 657 study participants and 1059 in-depth interviews, with each interview

Box 1

Using coding queries to evaluate the internal coherence of the coding scheme

1. Do the quotes represent similar concepts to a sufficient degree to justify a coherent theme?
2. Is the concept shared among study participants throughout the sample, or is it limited to specific subset? If limited to a specific subset, is the circumscribed nature of the concept adequately described in the manuscript?
3. Does the description or valence of the concept change during the course of the interview or during the course of multiple interviews with the same study participant? If so, are these changes adequately described in the manuscript?
4. Does the choice of quotes, and their accompanying descriptions, presented in the manuscript adequately capture the content and diversity of the coding query?

averaging 30–60 minutes in duration. Even redacting just the 175-page *summary reports* for each of the 48 sites – much less the primary interview transcripts – would have required the review of more than 8000 pages of data. In what format should such data be made available to meet the conditions of a reasonable data sharing policy?

Second, and related to the above, journals should consider the possibility that, in response to data sharing policies, study participants and qualitative researchers may alter their behavior in undesirable ways. Will qualitative researchers, whose work is already *de facto* excluded from most high-impact journals (Greenhalgh et al., 2016; Shuval et al., 2011), shy away from submitting their work to these journals, where data sharing policies are increasingly enforced? Will they be discouraged from conducting large-sample qualitative studies, knowing the documentation burden that will be involved? Furthermore, it is one thing to make available several hundred pages of interview transcripts from a two- to three-year qualitative study conducted by paid research assistants. It is another thing to make available thousands of pages of field notes and journal entries – some of which may be intensely personal in content – accumulated during the course of a five-year ethnography. Ethnographic note-taking guidelines that separate field notes according to observation, interpretation, and personal reflection (Bernard, 2006) could potentially facilitate data sharing by restricting dissemination to material related to observation. Unless qualitative researchers have a secure understanding that certain types of material can be shielded from dissemination, they may be motivated to alter the underlying data, i.e., by withholding this material from the written or transcribed record (Baez, 2002; Goodwin et al., 2003; McLellan et al., 2003; Scheper-Hughes, 2000) or by maintaining a set of private “shadow files” separate from the official research record (similar to the detailed “psychotherapy notes” that therapists store apart from the medical record). Box 2 summarizes our recommendations for journal policies that would promote transparency and, in some cases facilitate sharing of qualitative data, while remaining sensitive to their unique attributes that require their distribution to be handled somewhat differently than quantitative data.

4. Conclusion

Data sharing in medical and public health research is becoming increasingly normative, but medical and public health journals

Box 2

Summary of recommendations for journal editors

1. Require a statement from authors about whether the consent process included a description of any public availability of data. Prior to public dissemination of data, authors should provide a statement to journal editors about whether or not study participants were informed about future plans for public availability of data and the manner in which this issue was addressed, if at all, during the informed consent process.
2. Require adherence to minimum standards for de-identification of publicly shared data. Under the 1996 U.S. Health Insurance Portability and Accountability Act, protected health information includes 18 identifiers (e.g., names, addresses, serial numbers) that must be treated with special care in quantitative datasets. These same identifiers should be removed from qualitative data prior to dissemination. The geographic subdivision requirement, which stipulates that geographic units contain 20,000 or fewer people, requires special attention. If qualitative researchers are working in a village or community with fewer than 20,000 people, then site pseudonyms or larger geographic divisions should be used in published reports (e.g., providing the sub-county name rather than the parish or village name).
3. Encourage authors to use, and publish, data from multiple informants and/or institutions per selection category. Whenever possible, authors should be encouraged to recruit more than one informant and more than one institution per category. For example, interviewing only one surgeon at a hospital or only one official at a ministry of health increases the probability that the study participant's comments may be traced back to that study participant (or study participant's institution). If two or more informants are recruited per selection category and a range of institutions are included, the probability of identification may be reduced. Journal policies related to this provision should be cognizant of the lesser amounts of funding granted for qualitative research and the smaller scale of qualitative studies.
4. Permit coding queries to be shared as an alternative to full transcripts. Coding queries may offer greater anonymity compared with full transcripts because statements are grouped by theme rather than by study participant. Furthermore, coding queries allow a form of verification of the findings reported in results and conclusion. For the purposes of promoting transparency in qualitative research, these should be considered acceptable, or possibly even preferable, alternatives to full transcripts.
5. Encourage anonymization of field notes. Ethnographers frequently rely on field notes as a source of data. These could be anonymized in the same fashion as interview transcripts before being made publicly available. Because field notes include a range of objective, subjective, and interpretative documentation, requests for field notes should be limited to objective excerpts. Field notes, as with other forms of qualitative data, could also be submitted in the form of coding queries, with the same advantages as discussed above.
6. Encourage authors to document social audits or other stakeholder dissemination at the time of manuscript submission. A major source of participant-researcher

dispute occurs when participants feel that their responses are selectively represented in the reported results or in recommendations drawn from the data. Public availability of qualitative data may therefore be especially contentious if study participants, or their representatives (e.g., local leaders), have not signed off on the researchers' interpretations. Social audits or other stakeholder dissemination of results and conclusions prior to public availability of data will foster participants' perceptions of inclusiveness and accurate representation.

7. Encourage manuscript reviewers with requisite expertise in qualitative and mixed methods research to comment on the adequacy of anonymization. Study authors are ultimately responsible for anonymization. However, to promote good scientific practice, journal editors should encourage manuscript reviewers with requisite expertise in qualitative and mixed methods research to comment on the adequacy of anonymization and to raise any concerns they may have regarding potential maleficence resulting from data sharing.
8. Establish a petitioning process for non-disclosure of data. Authors should have the option of petitioning for non-disclosure of qualitative data in select instances. These include scenarios in which the study could not have yielded important results if participants were to have been required to consent to public disclosure of data, or in which anonymization could not be adequate given the uniqueness of the study population or the data.

have yet to grapple with how to feasibly and ethically promote data sharing for qualitative and mixed methods research. Recent advances in the field have begun to enhance the reliability and validity of qualitative data. Data sharing may help to increase confidence in qualitative research findings, but the concept of reproducible research does not translate as straightforwardly from quantitative data to qualitative data. Data sharing policies may be feasible for qualitative studies, but leading medical and public health journals should consider modifying their policies to be more relevant to the unique aspects of qualitative and mixed methods study designs; they must also address concerns about potential violations of participant anonymity and other unintended adverse consequences. Such policies, if appropriately implemented, can build a culture of data sharing that also facilitates critical, patient-oriented qualitative and mixed methods research.

Funding

No specific funding was received for the preparation of this manuscript. The authors acknowledge salary support through K23MH096620, K01MH104310, and K23MH095655. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests

ACT is an Editorial Associate for *Social Science and Medicine*, Associate Editor for *SSM - Population Health*, and a Specialty Consulting Editor for *Public Library of Science Medicine*. SLD is Associate Editor of the *Archives of Sexual Behavior*.

Acknowledgments

We thank Norma C. Ware, PhD for her comments on an earlier draft of the manuscript.

Appendix A. Supplementary data

Supplementary data related to this article can be found at <http://dx.doi.org/10.1016/j.socscimed.2016.08.004>.

References

- Abramowitz, S.A., Panter-Brick, C. (Eds.), 2015. *Medical Humanitarianism: Ethnographies of Practice*. University of Pennsylvania Press, Philadelphia.
- Ashenfelter, O., Haveman, R.H., Riley, J.G., Taylor, J.T., 1986. Editorial statement. *Am. Econ. Rev.* 76 (4).
- Atkins, D., Best, D., Briss, P.A., Eccles, M., Falck-Ytter, Y., Flottorp, S., et al., 2004. Grading quality of evidence and strength of recommendations. *BMJ* 328 (7454), 1490.
- Baez, B., 2002. Confidentiality in qualitative research: reflections on secrets, power and agency. *Qual. Res.* 2 (1), 35e58.
- Bernanke, B.S., 2004. Editorial statement. *Am. Econ. Rev.* 94 (1), 404.
- Bernard, H.R., 2006. *Research Methods in Anthropology: Qualitative and Quantitative Approaches*, fourth ed. AltaMira Press, Lanham.
- Bernard, H.R., Ryan, G.W., 2009. *Analyzing Qualitative Data: Systematic Approaches*. Sage Publications, Inc., Los Angeles.
- Bloom, T., Ganley, E., Winker, M., 2014. Data access for the open access literature: PLOS's data policy. *Public Libr. Sci. Med.* 11 (2), e1001607.
- Broom, A., Cheshire, L., Emmison, M., 2009. Qualitative researchers' understandings of their practice and the implications for data archiving and sharing. *Sociology* 43 (6), 1163e1180.
- Campbell, E.G., Clarridge, B.R., Gokhale, M., Birenbaum, L., Hilgartner, S., Holtzman, N.A., et al., 2002. Data withholding in academic genetics: evidence from a national survey. *J. Am. Med. Assoc.* 287 (4), 473e480.
- Chandra, A (amitabhchandra2). (2015). "If ethnography is a legitimate way to learn things... why aren't Rx manufacturers allowed to do it?" June 15, 2015, 5:33 AM. Tweet.
- Christopoulos, K.A., Olender, S., Lopez, A.M., Lekas, H.M., Jaiswal, J., Mellman, W., et al., 2015. Retained in HIV care but not on antiretroviral treatment: a qualitative patient-provider dyadic study. *Public Libr. Sci. Med.* 12 (8), e1001863.
- Claerbout, J.F., Karrenbach, M., 1992. Electronic documents give reproducible research a new meaning. *SEG Technical Program Expanded Abstracts*. Society of Exploration Geophysicists, Tulsa, pp. 601e604.
- Clark, J.P., 2003. How to peer review a qualitative manuscript. In: Godlee, F., Jefferson, T. (Eds.), *Peer Review in Health Sciences*, second ed. BMJ Books, London, pp. 219e235.
- Clemens, M.A., 2016. The meaning of failed replications: a review and proposal. *J. Econ. Surv.* <http://dx.doi.org/10.1111/joes.12139>. Epub ahead of print 26 Dec 2015, (in press).
- Closser, S., 2010. *Chasing Polio in Pakistan: Why the World's Largest Public Health Initiative May Fail*. Vanderbilt University Press, Nashville.
- Cohen, J., 1960. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* 20 (1), 37e46.
- Colditz, G.A., Hankinson, S.E., 2005. The Nurses' Health Study: lifestyle and health among women. *Nat. Rev. Cancer* 5 (5), 388e396.
- Colditz, G.A., Manson, J.E., Hankinson, S.E., 1997. The Nurses' Health Study: 20-year contribution to the understanding of health among women. *J. Women's Health* 6 (1), 49e62.
- Creswell, J.W., Klassen, A.C., Plano Clark, V.L., Smith, K.C., for the Office of Behavioral and Social Sciences Research, 2011. *Best Practices for Mixed Methods Research in the Health Sciences*. U.S. National Institutes of Health, Washington, D.C.
- Dewald, W.G., Thursby, J.G., Anderson, R.G., 1986. Replication in empirical economics: the *Journal of Money, Credit and Banking* project. *Am. Econ. Rev.* 76 (4), 587e603.
- Doshi, P., Jefferson, T., Del Mar, C., 2012. The imperative to share clinical study reports: recommendations from the Tamiflu experience. *Public Libr. Sci. Med.* 9 (4), e1001201.
- Ellis, C., 1986. *Fisher Folk: Two Communities on Chesapeake Bay*. University Press of Kentucky, Lexington.
- Ellis, C., 1995. Emotional and ethical quagmires in returning to the field. *J. Contemp. Ethnogr.* 24 (1), 68e98.
- Gandrud, C., 2013. *Reproducible Research with R and R Studio*. Chapman and Hall/CRC, London.
- Glaser, B.G., 1978. *Theoretical Sensitivity*. Sociology Press, Mill Valley.
- Glaser, B.G., Strauss, A.L., 1967. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Aldine Transaction, Chicago.
- Goffman, A., 2014. *On the Run: Fugitive Life in an American City*. University of Chicago Press, Chicago.
- Goodman, S.N., Fanelli, D., Ioannidis, J.P., 2016. What does research reproducibility mean? *Sci. Transl. Med.* 8 (341), 341ps312.
- Goodwin, D., Pope, C., Mort, M., Smith, A., 2003. Ethics and ethnography: an experiential account. *Qual. Health Res.* 13 (4), 567e577.
- Greenhalgh, T., Annandale, E., Ashcroft, R., Barlow, J., Black, N., Bleakley, A., et al., 2016. An open letter to The BMJ editors on qualitative research. *BMJ* 352, i563.
- Groleau, D., Young, A., Kirmayer, L.J., 2006. The McGill Illness Narrative Interview (MIND): an interview schedule to elicit meanings and modes of reasoning related to illness experience. *Transcult. Psychiatry* 43 (4), 671e691.
- Groves, T., 2010. BMJ policy on data sharing. *BMJ* 340, e564.
- Guba, E.G., Lincoln, Y.S., 1981. *Effective Evaluation: Improving the Usefulness of Evaluation Results through Responsive and Naturalistic Approaches*. Jossey-Bass, San Francisco.
- Guyatt, G.H., Sackett, D.L., Sinclair, J.C., Hayward, R., Cook, D.J., Cook, R.J., 1995. Users' guides to the medical literature. IX. A method for grading health care recommendations. Evidence-Based Medicine Working Group. *J. Am. Med. Assoc.* 274 (22), 1800e1804.
- Hamermesh, D.S., 2007. Viewpoint: replication in economics. *Can. J. Econ.* 40 (3), 715e733.
- Hanson, B., Sugden, A., Alberts, B., 2011. Making data maximally available. *Science* 331 (6018), 649.
- Horton, R., 1997. Pardonable revisions and protocol reviews. *Lancet* 349 (9044), 6.
- Hrynaskiewicz, I., Norton, M.L., Vickers, A.J., Altman, D.G., 2010. Preparing raw clinical data for publication: guidance for journal editors, authors, and peer reviewers. *Trials* 11, 9.
- Jefferson, T., Doshi, P., 2014. Multisystem failure: the story of anti-influenza drugs. *BMJ* 348, g2263.
- Kaiser, J., 1997. Showdown over clean air science. *Science* 277 (5325), 466e469.
- Kelly, J.T., Betancourt, T.S., Mukwege, D., Lipton, R., Vanrooyen, M.J., 2011. Experiences of female survivors of sexual violence in eastern Democratic Republic of the Congo: a mixed-methods study. *Confl. Health* 5, 25.
- King, G., 1995. Replication, replication. *PS Political Sci. Polit.* 28 (3), 444e452.
- King, R., Barker, J., Nakayiwa, S., Katuntu, D., Lubwama, G., Bagenda, D., et al., 2013. Men at risk: a qualitative study on HIV risk, gender identity and violence among men who have sex with men who report high risk behavior in Kampala, Uganda. *Public Libr. Sci. One* 8 (12), e82937.
- Kohrt, B.A., Tol, W.A., Pettigrew, J., Karki, R., 2010. Children and revolution: the mental health and psychosocial wellbeing of child soldiers in Nepal's Maoist Army. In: Singer, M., Hodge, G.D. (Eds.), *The War Machine and Global Health*. AltaMira Press, Lanham, pp. 89e116.
- Laine, C., Goodman, S.N., Griswold, M.E., Sox, H.C., 2007. Reproducible research: moving toward research the public can really trust. *Ann. Intern. Med.* 146 (6), 450e453.
- Larossa, R., Bennett, L.A., Gelles, R.J., 1981. Ethical dilemmas in qualitative family research. *J. Marriage Fam.* 43 (2), 303e313.
- Le Noury, J., Nardo, J.M., Healy, D., Jureidini, J., Raven, M., Tufanaru, C., et al., 2015. Restoring Study 329: efficacy and harms of paroxetine and imipramine in treatment of major depression in adolescence. *BMJ* 351, h4320.
- LeCompte, M.D., Schensul, J.J., 1999. *Designing and Conducting Ethnographic Research*. AltaMira Press, Walnut Creek.
- Lewis, O., 1951. *Life in a Mexican Village: Tepoztlán Restudied*. University of Illinois Press, Urbana.
- Lewis-Kraus, G., 2016. The Changeling. *NYT Sunday Magazine*, 31e37, 56-60.
- Lo, C., Ilic, D., Teede, H., Cass, A., Fulcher, G., Gallagher, M., et al., 2016. The perspectives of patients on health care for co-morbid diabetes and chronic kidney disease: a qualitative study. *Public Libr. Sci. One* 11 (1), e0146615.
- Lounsbury, D.W., Reynolds, T.C., Rapkin, B.D., Robson, M.E., Ostroff, J., 2007. Protecting the privacy of third-party information: recommendations for social and behavioral health researchers. *Soc. Sci. Med.* 64 (1), 213e222.
- Maman, S., van Rooyen, H., Stankard, P., Chingono, A., Muravha, T., Ntogwisangu, J., et al., 2014. NIMH Project Accept (HPTN 043): results from in-depth interviews with a longitudinal cohort of community members. *Public Libr. Sci. One* 9 (1), e87091.
- Markowitz, L., 2001. Finding the field: notes on the ethnography of NGOs. *Hum. Organ.* 60 (1), 40e46.
- McCullough, B.D., Vinod, H.D., 2003. Verifying the solution from a nonlinear solver: a case study. *Am. Econ. Rev.* 93 (3), 873e892.
- McLellan, E., MacQueen, K.M., Neidig, J.L., 2003. Beyond the qualitative interview: data preparation and transcription. *Field Methods* 15 (1), 63e84.
- Meier, K.J., 1995. Replication: a view from the streets. *PS Political Sci. Polit.* 28 (3), 456e459.
- Mello, M.M., Francer, J.K., Wilenzick, M., Teden, P., Bierer, B.E., Barnes, M., 2013. Preparing for responsible sharing of clinical trial data. *N. Engl. J. Med.* 369 (17), 1651e1658.
- Michaels, D., Monforton, C., 2005. Manufacturing uncertainty: contested science and the protection of the public's health and environment. *Am. J. Public Health* 95 (Suppl. 1), S39eS48.
- Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K.M., Gerber, A., et al., 2014. Social science. Promoting transparency in social science research. *Science* 343 (6166), 30e31.
- Morley, C.A., Kohrt, B.A., 2013. Impact of peer support on PTSD, hope, and functional impairment: a mixed-methods study of child soldiers in Nepal. *J. Aggress. Maltreatment Trauma* 22 (7), 714e734.
- Morse, J.M., Barrett, M., Mayan, M., Olson, K., Spiers, J., 2002. Verification strategies for establishing reliability and validity in qualitative research. *Int. J. Qual. Methods* 1 (2), 13e22.
- Muggli, M.E., Forster, J.L., Hurt, R.D., Repace, J.L., 2001. The smoke you don't see: uncovering tobacco industry scientific strategies aimed against environmental

- tobacco smoke policies. *Am. J. Public Health* 91 (9), 1419e1423.
- Nader, L., 1969. Up the anthropologist e perspectives gained from studying up. In: Hymes, D. (Ed.), *Reinventing Anthropology*. Pantheon, New York, pp. 284e311.
- Natoli, L., Guy, R.J., Shephard, M., Causer, L., Badman, S.G., Hengel, B., et al., 2015. "I do feel like a scientist at times": a qualitative study of the acceptability of molecular point-of-care testing for chlamydia and gonorrhoea to primary care professionals in a remote high STI burden setting. *Public Libr. Sci. One* 10 (12), e0145993.
- Nosek, B.A., Alter, G., Banks, G.C., Borsboom, D., Bowman, S.D., Breckler, S.J., et al., 2015. SCIENTIFIC STANDARDS. Promoting an open research culture. *Science* 348 (6242), 1422e1425.
- O'Brien, B.C., Harris, I.B., Beckman, T.J., Reed, D.A., Cook, D.A., 2014. Standards for reporting qualitative research: a synthesis of recommendations. *Acad. Med.* 89 (9), 1245e1251.
- Olken, B.A., 2015. Promises and perils of pre-analysis plans. *J. Econ. Perspect.* 9 (3), 61e80.
- Open Science Collaboration, 2012. An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspect. Psychol. Sci.* 7 (6), 657e660.
- Parkinson, S.E., 2013. Organizing rebellion: rethinking high-risk mobilization and social networks in war. *Am. Political Sci. Rev.* 107 (3), 418e432.
- Patton, M.Q., 2002. *Qualitative Research and Evaluation Methods*, third ed. Sage Publications, Thousand Oaks.
- Peng, R.D., 2009. Reproducible research and biostatistics. *Biostatistics* 10 (3), 405e408.
- Peng, R.D., Dominici, F., Zeger, S.L., 2006. Reproducible epidemiologic research. *Am. J. Epidemiol.* 163 (9), 783e789.
- PLOS Medicine Editors, 2014. Observational studies: getting clear about transparency. *Public Libr. Sci. Med.* 11 (8), e1001711.
- Redfield, R., 1930. *Tepoztlán: a Mexican Village*. University of Chicago Press, Chicago.
- Rennie, D., 1997. Thyroid storm. *J. Am. Med. Assoc.* 277 (15), 1238e1243.
- Rotheram-Borus, M.J., Tomlinson, M., Roux, I.L., Stein, J.A., 2015. Alcohol use, partner violence, and depression: a cluster randomized controlled trial among urban South African mothers over 3 years. *Am. J. Prev. Med.* 49 (5), 715e725.
- Rust, R.T., Cooil, B., 1994. Reliability measures for qualitative data: theory and implications. *J. Mark. Res.* 31 (1), 1e14.
- Scheper-Hughes, N., 1977. *Saints, Scholars and Schizophrenics: Mental Illness in Rural Ireland*. University of California Press, Berkeley.
- Scheper-Hughes, N., 2000. Ire in Ireland. *Ethnography* 1 (1), 117e140.
- Schneider, C.J., Aguiar, L.M. (Eds.), 2012. *Researching Amongst Elites: Challenges and Opportunities in Studying up*. Ashgate Publishing, Ltd, Farnham.
- Shuchman, M., 2005. *The Drug Trial: Nancy Olivieri and the Science Scandal that Rocked the Hospital for Sick Children*. Random House Canada, Toronto.
- Shual, K., Harker, K., Roudsari, B., Groce, N.E., Mills, B., Siddiqi, Z., et al., 2011. Is qualitative research second class science? A quantitative longitudinal examination of qualitative research in medical journals. *Public Libr. Sci. One* 6 (2), e16937.
- Stein, A., 2001. *The Stranger Next Door: the Story of a Small Community's Battle over Sex, Faith, and Civil Rights*. Beacon Press, Boston.
- Stein, A., 2010. Sex, truths, and audiotape: anonymity and the ethics of exposure in public ethnography. *J. Contemp. Ethnogr.* 39 (5), 554e568.
- Stenbacka, C., 2001. Qualitative research requires quality concepts of its own. *Manag. Decis.* 39 (7), 551e556.
- Stodden, V., Guo, P., Ma, Z., 2013. Toward reproducible computational research: an empirical analysis of data and code policy adoption by journals. *Public Libr. Sci. One* 8 (6), e67111.
- Stodden, V., Leisch, F., Peng, R.D., 2014. *Implementing Reproducible Research*. Chapman and Hall/CRC, London.
- Tang, X., Yang, F., Tang, T., Yang, X., Zhang, W., Wang, X., et al., 2015. Advantages and challenges of a village doctor-based cognitive behavioral therapy for late-life depression in rural China: a qualitative study. *Public Libr. Sci. One* 10 (9), e0137555.
- Tolich, M., 2004. Internal confidentiality: when confidentiality assurances fail relational informants. *Qual. Sociol.* 27 (1), 101e106.
- Tong, A., Sainsbury, P., Craig, J., 2007. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. *Int. J. Qual. Health Care* 19 (6), 349e357.
- Tsai, A.C., 2011. Managing nonfinancial conflict of interest: how the "New McCarthyism" could work. *Am. J. Bioeth.* 11 (1), 42e44.
- Tsai, A.C., Tomlinson, M., Comulada, W.S., Rotheram-Borus, M.J., 2016. Food insufficiency, depression, and the modifying role of social support: evidence from a population-based, prospective cohort of pregnant women in peri-urban South Africa. *Soc. Sci. Med.* 151, 69e77.
- Tudur Smith, C., Hopkins, C., Sydes, M.R., Woolfall, K., Clarke, M., Murray, G., et al., 2015. How should individual participant data (IPD) from publicly funded clinical trials be shared? *BMC Med.* 13, 298.
- U.S. National Institute of Justice, 2007. *Privacy Certificate Guidance*. U.S. National Institute of Justice, Washington, D.C.
- van den Hoonaard, W.C., 2003. Is anonymity an artifact in ethnographic research? *J. Acad. Ethics* 1 (2), 141e151.
- Vickers, A.J., 2006. Whose data set is it anyway? Sharing raw data from randomized trials. *Trials* 7, 15.
- Wade, A.S., Kane, C.T., Diallo, P.A., Diop, A.K., Gueye, K., Mboup, S., et al., 2005. HIV infection and sexually transmitted infections among men who have sex with men in Senegal. *AIDS* 19 (18), 2133e2140.
- Wolcott, H.F., 1973. *The Man in the Principal's Office: an Ethnography*. Holt, Rinehart, and Winston, Toronto.

Submission #40

Date: 11/15/2018

Name: Anonymous

Name of Organization: International Committee of Medical Journal Editors (ICMJE)

Type of Organization: Professional Org/Association

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

Clinical

I. The definition of Scientific Data

The definition of scientific data needs to be made more explicit. It needs to state that this does not include the primary acquired data, e.g. x-rays, ECG, pathology specimens, but it does include the interpretation of the primary acquired data, e.g., interpretation of images, ECGs or pathology specimens.

II. The requirements for Data Management and Sharing Plans

- The NIH should explicitly request that grant applications have an identifiable line item allocating a portion of the study budget to the Data Sharing Plan.
- It is incumbent on the NIH to support the Data Sharing Plan and provide or identify a repository for data sharing.
- Reviewers of applications should consider a Data Sharing Plan that widely limits access to data as unacceptable.
- The NIH should clarify if they are mandating data sharing – and provide guidance about which data, when, etc.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and

There are enough clinical trial data sharing platforms now available as to make such adoption of clinical trial data sharing easy for all to accomplish. Given that data sharing can increase the utility and applicability of scientific research findings, we would prefer implementation of this plan by the NIH as soon as possible.

Submission #41

Date: 11/16/2018

Name: Chrissa Papaioannou

Name of Organization: Rutgers, The State University of New Jersey

Type of Organization: University

Role: Institutional Official

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

All

Submission #42**Date:** 11/16/2018**Name:** Howard Fox**Name of Organization:** University of Nebraska Medical Center**Type of Organization:** University**Role:** Scientific Researcher**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Neuroscience, clinical and translational research

I. The definition of Scientific Data

Information collected in a study collected in an explainable fashion. This includes subjective and objective information, with documentation on how the information was collected, the value(s) determined, standards used in determination of the value(s), and if available the validation of the instrument used.

II. The requirements for Data Management and Sharing Plans

Data should preferably in a digital format if possible. Metadata are crucial and should be of sufficient depth to enable others to use and interpret the data (e.g. how subjects were chosen, type of data collected, any quality control on the method/assay, units and standards. Ideally should be as open access as possible, any constraints should be well-justified. Methods to insure de-identification of any human subjects and plans for maintenance of the data and its accessibility should be stated.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and

The main issue arising is how this will be paid for by the investigators, especially once the study is done. If this became the institution's responsibility (and I think it should, speaking as a Dean for Research) this can be included in the F&A calculations. This makes sense as the grants in reality are assigned to institutions. Many already have digital repositories for various sorts of files/information. Thus I could see this as a 3-year requirement for institutions to make a plan, with simultaneous consensus determinations of standards for privacy, intellectual property, security, and other concerns.

Submission #43**Date:** 11/20/2018**Name:** Elinor Schoenfeld**Name of Organization:** Stony Brook University**Type of Organization:** University**Role:** Scientific Researcher**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

epidemiology, aging, aging in place, technology to support aging, opioid addiction

II. The requirements for Data Management and Sharing Plans

I have some concerns about IV.2 tools - understand the goal to use open source but the majority of clinical trials use statistical software that provides robust analysis of study data. Making recommendations to use open source for clinical research data analysis may limit the robustness of the analyses performed and create even more variability in data format and code written for analysis.

It may be better to say that one should use statistical software that has the function to export to a number of other data formats along with their accompanying statistical codes so that there is interoperability and not limit encourage freeware.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and

You can set a date say of January 1, 2020 and have a phased in program. Active studies will be the most difficult to determine when to include. The appropriate NIH agency should work with the study teams to roll out updated sharing plans with each new grant year until all have addressed the sharing meeting new guidelines.

Submission #44

Date: 11/20/2018

Name: Brianna R Lindsay

Name of Organization: University of Pennsylvania

Type of Organization: University

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

Epidemiology

I. The definition of Scientific Data

I do not support the included definition of metadata as I believe categorizing intermediate, descriptive, or phenotypic observational variables is a mistake. Metadata to me are data that help to explain the scientific data. Using metadata to refer to data phenotypic observational variables is not appropriate as many time the information about the phenotype is some of the most important data about a sample. Limiting scientific data to data about samples and metadata as data about the participant/study subject is misguided. Metadata is more process oriented; used to aid in the conduct of the research; not the data that is analyzable.

Submission #45

Date: 11/25/2018

Name: Peter Adamczyk

Name of Organization: University of Wisconsin

Type of Organization: University

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

rehabilitation

I. The definition of Scientific Data

Definition is reasonable.

II. The requirements for Data Management and Sharing Plans

The whole Data Management and Sharing Plan should be negotiated as part of the Just-In-Time information for funding arrangements. NIH should provide a service to help researchers develop these plans, including up-to-date information on NIH-sponsored repositories and data management and sharing best-practices. This approach would allow researchers to spend their proposal time on their core science, and it would also improve the quality of data management plans due to expert input from NIH staff. And, it would provide an opportunity for NIH to promote uniformity and interoperability in the resulting data.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and

Submission #46**Date:** 11/26/2018**Name:** Kathy Helzlsouer**Name of Organization:** NCI**Type of Organization:** Government Agency**Role:** Government Official**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Epidemiology

I. The definition of Scientific Data

The examples in the policy should be border than just genetic/genomic. It is critically important to include all data collected in a study (genomic and non-genomic- questionnaire data as well as other laboratory assessment) in order to reproduce study aims. The data sharing should apply to all types of data - genomic and other.

II. The requirements for Data Management and Sharing Plans

It should be clearly noted that the data management and sharing plans must be approved by the funder prior to the award. The plan should be noted in the terms of the award. Data management and sharing plans should apply to all types of grants as well as intramural research - regardless of the amount of the funding (award).

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and

This should be implemented as soon as possible. The 2003 data sharing policy is out of date and this needs to be brought in line with the genomic data sharing policy. Those receiving awards are aware of the data sharing requirements and have included data sharing plans but these have not been adequately enforced. Therefore time for adoption should be short. One of the main problems has been enforcement of data sharing by intramural researchers. They should be leading by example. For epidemiologic research - the intramural program for most projects does not have a controlled access policy to de-identified data in compliance with the GDS/2003 policies. For example, they will say it is "available" and no one is turned down but researchers

must submit detailed plans of analysis and an approval process for the plans (which places roadblocks and is more stringent than for dbGAP or BioLNC.)

The enforcement of the data sharing for intramural activities should be done by those outside of intramural research. In the epidemiological world the intramural group is in direct competition with the extramural activity. While their charge is to do "high risk" activities that cannot be done by extramural researchers - this is rarely the case. As they are government workers, their work should be fully accessible - and data sharing implemented and enforced.

Submission #47**Date:** 11/26/2018**Name:** Martin Gruebele**Name of Organization:** U. of Illinois**Type of Organization:** University**Role:** Scientific Researcher**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

biophysics, chemistry

I. The definition of Scientific Data

I think the definition of digital data is OK. There is however still some analog data that could be given attention, such as dated and personalized lab notebooks. While these are slowly switching to digital, many people still use paper.

II. The requirements for Data Management and Sharing Plans

The current data management plan requirement is satisfactory. NIH needs to continue to work with databases such as Swissprot to make sure data availability is completely transparent, yet at the same time does not require time-wasting redundancy from investigators (i.e. submitting the same information in multiple places to satisfy requirements).

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

To allow all stakeholders (library centers, institutions, data warehouses, NIH, individual PIs, users of science data) time to adapt, any more rigorous data dissemination requirements should be phased in over a period of at least a year, or twice the period of current requirements. (e.g. if a deposition of article requirement currently has a 1 year embargo, shortening that time period should be phased in over 2 years).

Submission #48**Date:** 11/26/2018**Name:** Norbert Perrimon**Name of Organization:** Harvard Medical School**Type of Organization:** University**Role:** Scientific Researcher**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

No comments

I. The definition of Scientific Data

(1) We see that the idea of standards is raised in Section IV, part 3 of the Proposed Provisions for a Draft NIH Data Management and Sharing Policy document. We think that the idea of requiring that researchers meet specific standards is a critical aspect of data management, including as it relates to metadata. We provide the example of RNAseq studies in our field, including both bulk RNAseq studies. Many RNAseq datasets are available in a standard data format at NCBI GEO. However, re-use of these data sets is limited because the metadata do not use standard terms or provide information sufficient to map to standard terms. Standards exist for naming a specific organ or tissue types, for example, controlled vocabularies (CVs) and ontologies have been established by experts for curation of model organism and human databases. We urge the NIH to be more specific and clear with regards to standards, including but not limited to the use of standard CVs in metadata and data sets, for example in describing the specific organ or tissue from which a sample subjected to genomics analysis was taken, so that large-scale genomic data, including bulk RNAseq and single-cell RNAseq data, adhere to FAIR data principles.

(2) In our field, one of the most difficult data types to share is high-content images generated in large-scale functional genomics screens. The difficulty lies in both the file structure (many files) and in the total data size (the set of image data associated with a high-content image-based screen can add up to several terabytes). The ability to acquire and analyze the data have outpaced the ability to effectively manage and share these data. At our institution, there have been investments in image file management that are beginning to have positive impact. Good solutions for image data management and sharing do exist. However, there is the lack of a centralized or a standard solution with long-term support. We feel that the value of storing this

type of image data should be reviewed and if deemed valuable, new investments in infrastructure be made so that these data sets can be managed and shared in a way that adheres to FAIR data principles.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

For our field, the optimal timing for NIH to start requiring improved data management and sharing plans is now, particularly for single-cell RNAseq (scRNAseq) data. Defining and requiring standards for this data type now could have significant positive impact, whereas waiting will mean that an increasing amount of scRNAseq data will not adhere to FAIR data principles. We urge NIH not to miss an opportunity to intervene early in establishing standards for scRNAseq data management and sharing.

Attachment:

Comments on “Proposed Provisions for a Draft NIH Data Management and Sharing Policy”

Nov. 26, 2018

Provided by:

Norbert Perrimon, PhD

Professor of Genetics at Harvard Medical School
Investigator at Howard Hughes Medical Institute

Stephanie E. Mohr, PhD

Lecturer on Genetics at Harvard Medical School

I. Definition of Scientific Data

No comments

II. The requirements for Data Management and Sharing Plans

(1) We see that the idea of standards is raised in Section IV, part 3 of the Proposed Provisions for a Draft NIH Data Management and Sharing Policy document. We think that the idea of requiring that researchers meet specific standards is a critical aspect of data management, including as it relates to metadata. We provide the example of RNAseq studies in our field, including both bulk RNAseq studies. Many RNAseq datasets are available in a standard data format at NCBI GEO. However, re-use of these data sets is limited because the metadata do not use standard terms or provide information sufficient to map to standard terms. Standards exist for naming a specific organ or tissue types, for example, controlled vocabularies (CVs) and ontologies have been established by experts for curation of model organism and human databases. We urge the NIH to be more specific and clear with regards to standards, including but not limited to the use of standard CVs in metadata and data sets, for example in describing the specific organ or tissue from which a sample subjected to genomics analysis was taken, so that large-scale genomic data, including bulk RNAseq and single-cell RNAseq data, adhere to FAIR data principles.

(2) In our field, one of the most difficult data types to share is high-content images generated in large-scale functional genomics screens. The difficulty lies in both the file structure (many files) and in the total data size (the set of image data associated with a high-content image-based screen can add up to several terabytes). The ability to acquire and analyze the data have outpaced the ability to effectively manage and share these data. At our institution, there have been investments in image file management that are beginning to have positive impact. Good solutions for image data management and sharing do exist. However, there is the lack of a centralized or a standard solution with long-term support. We feel that the value of storing this type of image data should be reviewed and if deemed valuable, new investments in infrastructure be made so that these data sets can be managed and shared in a way that adheres to FAIR data principles.

III. The optimal timing ...

For our field, the optimal timing for NIH to start requiring improved data management and sharing plans is now, particularly for single-cell RNAseq (scRNAseq) data. Defining and requiring standards for this data type now could have significant positive impact, whereas waiting will mean that an increasing amount of scRNAseq data will not adhere to FAIR data principles. We urge NIH not to miss an opportunity to intervene early in establishing standards for scRNAseq data management and sharing.

Submission #49

Date: 11/26/2018

Name: Gail Adler, MD, PhD

Name of Organization: Brigham and Women's Hospital/Harvard Medical School

Type of Organization: University

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

Basic and clinical research related to hormones and cardiovascular disease. Both the clinical and basic studies are physiology type studies.

II. The requirements for Data Management and Sharing Plans

To avoid placing onerous demands on grant writers and grant reviewers, detailed data management and sharing plans should be required at the JIT stage rather than at the grant submission stage. At the grant submission stage a brief statement indicating that the principal investigator will utilize data management and sharing plans consistent with NIH guidelines should be sufficient.

NIH should allow for additional monies beyond the 500,000 annual direct costs to pay for the additional personnel needed to comply with new data management and sharing policies. These additional monies would pay for the training of personnel (turnover for research assistants is every 1-2 years) and for the time needed for compliance with the new policies.

All required data management platforms should easily integrate into multiple data analyses programs.

All data sharing plans should preserve the publication rights of the investigators collecting the data. Also, the investigators who collected the data should have the option of being collaborators in future publications using their data. The investigators who collected the data should be kept aware of all analyses being performed with the collected data.

Submission #50

Date: 11/28/2018

Name: Jennifer Darragh

Name of Organization: Duke University

Type of Organization: University

Role: Institutional Official

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

Duke University is a leader in biomedical research across multiple areas. As the research data management and curation group within the Duke University Libraries, we are engaged with various different research disciplines and often review DMPs for various disciplines and funding agencies.

I. The definition of Scientific Data

See attached

II. The requirements for Data Management and Sharing Plans

See attached

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

See attached

Attachment:

RESEARCH AREA MOST IMPORTANT TO YOU OR YOUR ORGANIZATION

As a research data management and curation group within an academic library, we are engaged with various different research disciplines and often review DMPs for various disciplines and funding agencies.

I. DEFINITION OF SCIENTIFIC DATA

Generally, the definition is broad enough to accommodate various disciplines while also specific enough to avoid an “all things are data” perspective; however, we would suggest not omitting “physical objects” from the definition of scientific data. Specimens, cell lines, samples, etc. serve as a foundation for medical and scientific research (i.e., HeLa Cells). These physical samples should be preserved and shared as appropriate. Digital surrogates should be created in tandem with preserving primary physical objects/items in the event that sharing or preservation is not feasible. This is keeping in line with the idea of broad consent when physical samples are collected (latest revision to Common Rule) to ensure proper bio-banking and data reuse.

II. REQUIREMENTS FOR DATA MANAGEMENT AND SHARING PLANS

Plan Review and Evaluation: Requiring successful applications to modify their plans after submission but before funding is granted and allowing an acceptable/not-acceptable assessment by reviewers for extramural grants is a useful step towards improving plans. However, it might be even more effective to avoid the “check the box” phenomenon and a back flow of revisions, to include plans within the overall impact score. This might also serve as a means to see better quality plans from the start.

Plan Elements:

- Page limits can be very restrictive if you want plans to be thorough. Since medical and health related research often involves privacy and security concerns, expanding the page limit to three pages would provide more flexibility for providing comprehensive and thoughtful responses.
- *Data Type:* Section 1.2 provides a brief reference to “other information necessary to interpret data” (i.e., accompanying documentation). The types of supporting materials should be expanded to include items such as data dictionaries or codebooks that define variables, values, weighting, etc. This type of documentation allows the data to stand alone as a discrete scientific source and is very important for data reuse. NIH may also want to consider expanding the “Standards” section to more explicitly include less structured types of documentation that might be necessary to use if a community does not have an established data standard.
- *Related Tools, Software and/or Code:* Information on software should also include software versions, since many software packages can become obsolete over time. In addition, it is a common best practice to include a recognized open or portable format that allows the data to be read into different software packages (e.g. Comma Separated Values File vs. Excel Spreadsheets). This section of the plan could more directly address file formats and suggest creating open, non-proprietary versions of files when appropriate.

- *Data Preservation, Access and Sharing*: Given the specificity you are asking for with regards to preservation and access as well as licensing, NIH should be more prescriptive on recommending deposit within an established repository. Identifying a repository early in the process will answer most of the questions pertaining to these areas in the plan. If the researcher is unsure about what repository they should use, their program officer should be able to recommend an appropriate option. NIH could also suggest researchers reach out to library or other research support units on campus to get help identifying an appropriate repository. In addition, the proliferation of multiple types of repositories can make decisions more difficult. Serious thought should be put into whether “new” repositories need to be developed

III. TIMELINE

These new requirements do not seem so onerous that they could not be implemented in a timely manner. However, giving a 6-month lead-time will allow both researchers and institutional stakeholders time to acclimate to the new requirements.

Submission #51**Date:** 11/28/2018**Name:** David R. Bobbitt**Name of Organization:** CDISC**Type of Organization:** Other**Other Type of Organization:** Standards development organization**Role:** Institutional Official**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Clinical research and nonclinical research

I. The definition of Scientific Data

Data developed in any activity related to the research enterprise in either academic or industry settings can be considered scientific data. By this we mean that data generated about subjects, hypotheses, outcomes, and interim activities are all scientific data. So too are data generated in proof of concept, piloting, scoping, and other development activities.

Of no less importance, all scientific data ought to generate an informational and descriptive layer of metadata (data about the data) answering:

- How, where and when the data were collected?
- What are the conditions of scientific enterprise and the constituencies of the research hypothesis?
- Were the data curated, tested, audited, and/or reviewed? By whom and with what results?
- How were the data secured, coded, and standardized?
- How may the data be transmitted for data sharing purposes?

II. The requirements for Data Management and Sharing Plans

Data management is a complex issue, and while related to data standards, we won't speak to it as it is not our area of expertise. Data sharing requires a level of data standardization, which is our area of expertise. There are many misconceptions about data standardization; our hope is

to dispel some of these misconceptions, while underscoring the critical role of data standardization for any data sharing plan.

Data sharing requires content standards, which can be used for organizing and formatting data to streamline processes in collection, management, analysis and reporting of data. Data sharing also requires technical standards, which facilitate messaging between computer systems and performance metrics used to measure various operational benchmarks.

One major misconception of data standardization is that its use limits the scope of questions or research in which the investigator can engage. Another misconception is that standardization imposes value judgments on the quality of research. Neither concern is true for content standards. CDISC includes both content standards and technical standards for transporting and archiving data.

CDISC builds quality content standards that are required for submission of electronic data by the US Food and Drug Administration, the Japan Pharmaceutical and Medical Devices Agency, and used by the China National Medical Products Agency (formerly China FDA), and a host of other pharmaceutical and medical device regulatory agencies around the globe.

When correctly deployed, content standards support FAIR data principles(1) and consistently help to:

- Find the data. Standardized data generally appears in the same place in data sets, the same format, and with the same contextual clues.
- Access the data, help users determine the completeness of the data, and note any missing components.
- Understand the data and the research question or hypothesis that the investigator is attempting to answer.
- Provide statistical analyses in a standardized view.
- Make the data interoperable by linking to machine-actionable data and metadata, and through the use of shared vocabularies and ontologies.
- Facilitate data re-use by linking data across studies and from a variety of origins to engage in analysis across multiple data sources, a process called meta-analysis.

Content standards are thus indispensable for any effort to share data.

References

- 1) Wilkinson, M, Dumontier, M, et. al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Retrieved from <https://www.nature.com/articles/sdata201618>

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

We can only speak to the standardization component of this timing question. We believe all divisions of the NIH should begin standardization of internal and extramural research data sooner rather than later. The NIH mission is currently constrained by an inability to engage in meta-analyses of research funded by the NIH. NIH researchers are not required to report back data nor is there a universal content standard for NIH data. Until these two requirements are implemented: (1) a requirement to standardize research data and (2) a requirement for researchers to provide data to the NIH in this standardized format, meta-analysis will remain elusive.

At CDISC, our belief is that US NIH should pilot a data standardization effort using a quality data standards content system. CDISC is the best option for this standardization effort due to the global community of adopters of the standards and our long standing and successful collaborations with the National Cancer Institute (NCI), the National Institute of Allergy and Infectious Diseases (NIAID), National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), and the National Institute of Child Health and Human Development (NICHD) that have resulted in the development of CDISC standards for specific therapeutic areas. Use cases of successful implementation of these standards and their ability to lead researchers to discover new biomarkers and treatments can be found on the CDISC website:

<https://www.cdisc.org/use-cases-for-clear-data>.

We estimate a pilot would take approximately 18 months and would, building upon work already completed, primarily focus on standardizing large bodies of scientific data in both human and animal subjects, adjusting the CDISC data model to fully meet the needs of NIH researchers and funded research community. After this pilot phase, deploying standardization will likely take an additional 24-36 months, depending on resources. The NIH would then begin to see the benefits of meta-analysis.

Submission #52

Date: 11/29/2018

Name: Yongjian Liu

Name of Organization: Washington University School of Medicine

Type of Organization: University

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

Diagnosis and treatment of cardiovascular diseases

I. The definition of Scientific Data

Scientific data is typically defined as information collected using specific methods for a specific purpose of studying or analyzing. However, it is important to have appropriate controls in the data collection process. Otherwise, it may have false or misleading information.

Submission #53**Date:** 11/30/2018**Name:** Jerry Power**Name of Organization:** USC**Type of Organization:** University**Role:** Institutional Official**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Data Governance

II. The requirements for Data Management and Sharing Plans

There are 3 key aspects that need to be considered when seeking data. (1) the institutions overriding data policy (policies toward sharing, data security measures in place, notification policies, etc). (2) the project specific data (the need for the data, retention periods, etc), and (3) incentives etc that would be provided to the data source in exchange for the data.

Most data privacy policy policies are written by lawyers for other lawyers. The the lay person they are often not understood, long, and complicated. Ideally NIH will develop templates and examples to define how these policies are explained to the public so they can understand what is being done.

People should be able to rescind any approvals they have provided in the past. They should also be able to ask that any information collected from them is completely deleted. A stop action and a delete action should be considered as two distinct operations where one does not necessarily imply the other.

Permission records should be stored by an independent third party. If the user asks an organization to stop collecting data because they no longer trust that authority, that authority cannot be trusted to properly maintain a record of the stop request. Incentive distribution records should also be managed by an independent third party so there is a means of validating incentives were properly managed.

Data can be sourced directly from human users. Data can also be sourced from computing/IOT devices. The assumption is that the owner of any such automated devices is 1) responsible for the device, 2) responsible for the data from the devices they own. If the device is owned by a

commercial or government agency, they should be an individual within that organization that is identified as being responsible for the data from that device.

Data is a malleable entity. If a user provides their data to an entity and the entity changes some aspect of the data, it is now different from the original data. While it is clear that the original data is sourced and presumably owned by the data generator, it is not clear who owns the modified data. The modification of the data might be as simple as masking of the data source or it could be a complicated averaging of a much larger data set. Tracking the provenance of any data set to its original source is at best complicated, but more likely impossible. In the end, whether a user releases their data to any entity comes down to a matter of trust. Therefore it becomes important that users also have a crowd sourced means to evaluate the trust of any entity asking for their data; users need a way to ask others whether the permission documents the requesting organization has sent them can be trusted (or not).

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

I am a member of the I3 consortium. The Consortium is working to build an open source system that manages/governs the flow of information between IOT devices and organizations (government, commercial, etc). While our efforts have been initially targeted to Smart-Cities applications, the same technology might be of value in a NIH setting. The system is primarily focused on IOT environments but we have included the concept that data analytic experts will want to apply value added processing to data sets and offer the embellished data sets in an open market place. Such a data broker concept has the potential to be reapplied to a more human data collection environment.

If NIH is interested in the efforts of the I3 Consortium, information can be found at i3.usc.edu. NIH is welcome to join the consortium and participate in the process as we move the I3 concept from a proof-of-concept system to a public open source system.

Submission #54**Date:** 12/03/2018**Name:** Denise Sturdy**Name of Organization:** Duke Clinical Research Institute**Type of Organization:** University**Role:** Member of the Public**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Clinical research across all disciplines and therapeutic areas

II. The requirements for Data Management and Sharing Plans

As a Regulatory Professional working at a large Academic Research Organization, I am commenting generally to state support for the “Proposed Provisions for a Draft NIH Data Management and Sharing Policy.” In my experience, data sharing allows the best opportunity to advance medical research and fosters higher standards and accountability for research. Research that is supported by federal funds was never intended to be solely for the original researcher’s edification: The results from all medical research, regardless of outcome, can be beneficial to future studies. The proposed requirement is also in line with the International Committee of Medical Journal Editors requirement for a data sharing plan in a trial’s registration, effective for clinical trials enrolling participants on or after January 1, 2019.

The faculty at this institution are universally strong proponents of open data and support including the “Data Management and Sharing Plan” as part of the funding/support application process. We would encourage NIH to implement training to all program officers on the requirements so that the data sharing and management plans submitted with applications are monitored for compliance and enforced. Data sharing should be seen as a condition to any funding that must be honored by all recipients.

Regarding the proposal’s requirement to indicate what software/computer codes will be used to process or analyze the scientific data, and if not an open code, describe alternative free and open source software/codes that may be used in subsequent analyses, we recognize the importance of ensuring that data is at least actually readable by future researchers; however, our biostatisticians have indicated that in the clinical trial arena, the logistics for meeting this

requirement would benefit by having additional detail around this requirement in the final policy.

Consideration of the effects of global privacy laws and regulations may also be merited. Patient privacy should not be allowed to serve as an excuse for refusal to share research data; rather, data sharing plans with global implications should incorporate plans to remove identifying information such that future researchers cannot re-identify participants in compliance with such laws.

The Patient-Centered Outcomes Research Institute (PCORI) recently adopted a “Policy for Data Management and Data Sharing”, requiring that PCORI-funded research data and findings be shared with other researchers for future studies. Data management and sharing plans must be included in PCORI funded research proposals. The PCORI plan, linked here:

(<https://www.pcori.org/sites/default/files/PCORI-Policy-for-Data-Management-and-Data-Sharing.pdf>) may be a valuable resource for other researchers creating their own plans.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

The sooner the better.

Submission #55**Date:** 12/03/2018**Name:** Helen**Name of Organization:** Berman**Type of Organization:** University**Role:** Scientific Researcher**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

My primary research has been in structural biology and structural bioinformatics. I am the former head of the RCSB Protein Data Bank. I have a long standing interest and involvement in data management and data sharing.

I. The definition of Scientific Data

Although every experiment done in a laboratory results in the production of data, the data that result in a publication or in a deposition in a data repository need to be preserved in order to better ensure reproducibility. Reproducibility is enabled via the preservation of the analysis code that lead to the generation of published findings, often from the data created by the experiment. In other words, all digital artifacts that support published findings should be preserved. Data may include images, spectra, other experimental measurements as well as the metadata that define the conditions surrounding the data collection. The software and workflows used to analyze the data must be preserved. The proposed guidelines exclude physical samples. If a repository already exists for these samples, it should be declared.

II. The requirements for Data Management and Sharing Plans

The key provisions as described in the RFI seem to be on target. However, to make data management plans useful and enforceable, they must be fully computer readable. Controlled vocabularies need to be used so that the DMP's are themselves searchable. By doing this, program officers will be able to determine the exactly how data are archived including: the names of the repositories for data and software, the names of the software packages, the names of the standards. It will also make it possible to determine if there is a need for a new repository for certain types of data.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

There is an urgent need for the adoption of a data management and sharing policy. Data management tools are currently available to make this possible although they do not use controlled vocabularies. To correct this, Kerstin Lehnert (Columbia), Victoria Stodden (University of Illinois) and I are working on an NSF funded project to create a tool that contains most of the features described in the proposed NIH policy. It is called ezDMP and an alpha version of the tool is accessible here:

<http://dev.ezdmp.org/index>

Although the tool was created to meet the requirements of NSF DMP's, its design is such that it could be tailored to meet NIH requirements.

Submission #56

Date: 12/04/2018

Name: Elisa A. Hurley

Name of Organization: Public Responsibility in Medicine and Research (PRIM&R)

Type of Organization: Other

Other Type of Organization: non-profit

Role: Bioethicist/Social Science Researcher

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

Research ethics

I. The definition of Scientific Data

Please see attached document.

II. The requirements for Data Management and Sharing Plans

Please see attached document.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

Please see attached document.

Attachment:

Chair

Heather H. Pierce, JD, MPH

Vice Chair

Christine Grady, MSN, PhD

Secretary

David H. Strauss, MD

Treasurer

Christian E. Newcomer,
VMD, MS, DACLAM

Board of Directors

Albert J. Allen, MD, PhD

A. Cornelius Baker

Barbara E. Bierer, MD

Elizabeth A. Buchanan, PhD

Alexander Capron, LLB

Owen Garrick, MD, MBA

Bruce Gordon, MD

Mary L. Gray, PhD

F. Claire Hankenson,
DVM, MS, DACLAM

Karen M. Hansen

Martha Jones, MA, CIP

Natalie L. Mays, BA, LATG, CPIA

Robert Nobles, DrPH, MPH, CIP

Sally Okun, RN, MMHS

Suzanne Rivera, PhD, MSW

Stephen Rosenfeld, MD

Walter L. Strauss, MD, MPH

Ex Officio

Elisa A. Hurley, PhD
Executive Director

December 3, 2018

Submitted electronically at <https://osp.od.nih.gov/provisions-data-management-sharing/>

Francis S. Collins, MD, PhD
National Institutes of Health
6705 Rockledge Drive, Suite 750
Bethesda, MD 20892-7985

RE: Request for Information on Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research

Dear Dr. Collins:

Public Responsibility in Medicine and Research (PRIM&R) appreciates the opportunity to comment on the National Institutes of Health (NIH)'s Request for Information on Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research, published October 10, 2018.

PRIM&R is a nonprofit organization dedicated to advancing the highest ethical standards in the conduct of research. Since 1974, PRIM&R has served as a professional home and trusted thought leader for the research protections community, including members and staff of human research protection programs and institutional review boards (IRBs), investigators, and their institutions. Through educational programming, professional development opportunities, and public policy initiatives, PRIM&R seeks to ensure that all stakeholders in the research enterprise understand the central importance of ethics to the advancement of science.

PRIM&R fully supports initiatives that seek to promote broad data sharing. We agree with the NIH that data sharing can optimize the use of scarce research resources and has the potential to accelerate science and its application to human health. Furthermore, the sharing

of scientific data from clinical trials and other research involving humans honors those subjects' contributions by maximizing the value of their involvement. Relevant to the NIH's point that data sharing can inform "future research pathways," data sharing may also lead to better designed, and safer, future research, serving the ethical imperative to minimize risks to research subjects.

While data sharing has significant benefits, it also involves inherent risks, most notably privacy and confidentiality risks particularly when it comes to certain types of genetic data. These risks are magnified by the fact that research often involves accessing and aggregating multiple primary data sets. Though each of these data sets may include only "de-identified" personal data, their aggregation increases the chances that individuals will be inadvertently identified and their privacy breached. Indeed, advances in technology and the proliferation of data sources mean that no data can be considered permanently de-identified.

These risks, most agree, are, unavoidable, but there are strategies to mitigate them. **We believe the draft provisions fall short on acknowledging and grappling with these realities of data sharing. As the NIH considers updating its data sharing policies, it has a unique and important opportunity to lead the way on responsible data sharing by articulating the tradeoffs between maximizing the value of scientific data and protecting the rights and interests of research subjects, and by providing guidance on best practices for responsible data sharing given those tradeoffs. Both of these measures will, in turn, enhance public trust in the data sharing enterprise.**

In what follows we elaborate on these points, highlighting more specific areas we urge the NIH to address in future iterations of data sharing and management policies. In response to the specific areas of focus in the RFI, we begin by addressing, in sections I and II, the definition of scientific data and the proposed requirements for data management and sharing plans. Section III raises several additional points for the NIH to consider.

I. The Definition of Scientific Data

The definition of "scientific data" provided requires clarification. According to the proposal, scientific data is "the recorded factual material commonly accepted in the scientific community as necessary to validate and replicate research findings including, but not limited to, data used to support scholarly publications," and "may include certain individual-level and summary or aggregate data, as well as metadata." The definition offered is very broad but, at the same time, excludes specific sources of information, which could lead to confusion about how best to interpret and implement it. For example, the

definition of scientific data explicitly excludes laboratory notebooks. However, the information contained in laboratory notebooks might reasonably be seen as “necessary to validate and replicate research findings.” Indeed, given its breadth, the proposed definition of scientific data seems open to the interpretation that almost *all* data collected for a study counts, and therefore should be shared. We are concerned that a definition of scientific data that is open in this way to multiple local interpretations increases the likelihood that more identifiable information will be shared than is intended by the policy.

Furthermore, we note that the proposed definition of “scientific data” is presumably meant to include data from both quantitative and qualitative research. If this is the case, then the draft provisions fail to acknowledge that there are key differences between qualitative and quantitative research methods and data. For instance, qualitative research data often contains more identifiable information than quantitative research data. PRIM&R is unaware of any established standards for making qualitative data widely available in a way that protects the rights and interests of research subjects. Indeed, qualitative researchers would argue that there are strong ethical reasons *not* to share primary datasets. The one-size-fits-all model proposed may not be appropriate for all research data. The NIH should address this concern, at the very least clarifying whether and when its policies apply to both quantitative and qualitative data, and if so, acknowledging the unique challenges associated with the latter.

Finally, we urge the NIH to consider and clarify whether and how its definition of scientific data applies to data that is generated after a grant ends. We can imagine circumstances in which a researcher re-analyzes a project’s data after the end of the grant that initially funded the project, and finds something that fits the proposed definition of scientific data. Will that data be covered by the NIH’s policy? What is NIH’s “reach through” in such circumstances? The agency should address how it plans to oversee any sharing requirements when new data is generated after the funding period has ended.

II. The Requirements for Data Management and Sharing Plans

We believe the NIH’s proposed requirements for data management and sharing plans cover many important elements of such plans. The agency notes several times that data management and sharing plans should provide for the broadest use of data, “consistent with privacy, security, informed consent, and proprietary issues.” However, the agency provides very little guidance about what those issues are or how they should be addressed in data sharing plans. We urge the NIH to more fully acknowledge and address the risks and complexities associated with data sharing. Below we provide several examples.

First, the proposed provisions suggest that use of “persistent unique identifiers” for scientific data would be acceptable as an indexing tool for making shared data discoverable. However, persistent unique identifiers may actually facilitate reidentification. For example, some government projects require the use of the Global Unique Identifier (GUID Tool). In these circumstances, a given subject retains the same GUID, which enables the triangulation of data from unrelated studies and poses privacy and confidentiality risks. We suggest the NIH reconsider whether it should endorse use of persistent unique identifiers or instead suggest other indexing tools that might be more appropriate and less susceptible to re-identification efforts.

Second, we are also concerned about the requirement that broad sharing be consistent with informed consent. The revised Common Rule requires consent documents to include a statement about what will happen to any identifiable private information collected during the course of research—specifically, whether or not the information might be stripped of identifiers and distributed or used for future research without consent. This means that for data collected under the new rule, it will be relatively clear what it means to share it “consistent with consent.” However, information collected prior to the January 21, 2019 compliance date for the revised rule is not subject to those new consent requirements, and existing consent forms vary with respect to how, or even whether, they address data sharing. In these circumstances, it is not clear what it means to say that sharing should be done as broadly as possible, consistent with consent.

We again encourage the NIH to lead by providing guidance for IRBs and other stakeholders on how to determine what retrospective uses of existing data, including data sharing, would be ethically appropriate when consent is not specific about, or is silent on, future uses. We urge the NIH to be explicit, in its own policies, about the series of considerations that come into play when making these decisions, such as the characteristics of the study population, the sensitivity of the data, the likelihood of reidentification, and the scientific utility and value of the data itself. The NIH should also remind stakeholders that these issues may need to be reviewed on a case by case basis to reach a decision about how best to share data while protecting research subjects’ rights.

More generally, the **NIH should provide more guidance not just on how appropriate informed consent can facilitate responsible data sharing, but also on best practices for sharing data in ways that are consistent with privacy and confidentiality standards.** For example, the draft policy currently does not mention the HIPAA Privacy Rule, with which much data sharing must, of course, be consistent for certain health related research. The research community would also benefit from guidance on when it is reasonable to place restrictions on data use and sharing.

Finally, **the NIH should encourage, if not require, data management and sharing plans to include provisions about how research subjects will be informed about the limitations of current technologies to completely de-identify or anonymize their data while preserving that data's utility for research.** This sort of transparency about data sharing and the tradeoffs involved demonstrates respect for research subjects and may enhance the public's trust in the data sharing enterprise. In the same spirit of transparency and fostering trust, we further urge the NIH to encourage the creators of data management and sharing plans to incorporate input from research subjects and/or the public on those plans' assessment of risks and benefits. This may not necessarily require soliciting input on each project's plan, but rather institutions seeking input on the risks and benefits associated with particular categories of data or types of research.

Given the inherent complexity of all of these issues, we suggest eliminating the proposed two-page cap for data management and sharing plans.

III. Other considerations

We encourage the NIH to consider how it can ensure that institutions who have promised to share data have the resources to do it well. The proposed provisions focus almost exclusively on the requirement to create a data management and sharing plan, and on what should be included in the plan. But plans are only as effective as their implementation. We are concerned that institutions with fewer resources to dedicate to data sharing, and/or less experience with data sharing, may write good plans, but may be unable to execute them successfully, leaving people and their data vulnerable. This concern is particularly acute given the agency's expectation that, where possible, scientific data be digitized, a resource intensive process.

These concerns may be partially addressed by the NIH encouraging grantees to request the appropriate amount of resources to facilitate data sharing in a safe and ethical manner—though less experienced institutions may need help from the NIH understanding what the costs are. Furthermore, the NIH should consider other ways it can support institutions with fewer resources for safe data sharing, so that this policy does not for them constitute an unfunded mandate.

Relatedly, we suggest that any future policy expand on the compliance and enforcement provisions proposed. Although other sections of the proposal emphasize that data sharing must be consistent with privacy and confidentiality considerations and informed consent, there is no discussion of what penalties might be levied if research subjects' rights are

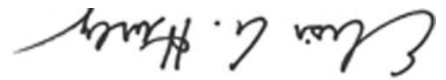
violated in the course of data sharing—for instance, in the event that private information about them gets in the wrong hands—and how to determine who should be held responsible for such violations. As data sharing becomes more prevalent, the public will increasingly demand consequences when their data are not shared with adequate attention to protections. **The NIH can demonstrate its commitment to the public’s interest by detailing the consequences when data is shared inappropriately, beyond just rescinding funding.**

In addition, although the proposed policy mentions the utility of data repositories, it doesn’t address the current proliferation of repositories, each with their own rules and procedures. Not only does this state of affairs lead to confusion, it weakens the overall utility of the data sharing enterprise. Effective use of existing data to advance science requires an accessible set of data repositories structured in a rational and coherent way. The agency endorses use of repositories that meet “community-based standards,” but it is unclear, without further explication, what the NIH has in mind—for instance, whether the agency means the FAIR data principles. **We urge the NIH to use its policies to encourage standardization across data repositories, and to articulate a gold standard for how data should be managed and shared to maximize utility.**

Ultimately, **it will be important for the NIH to justify the relative risks and benefits of its data sharing policies**, and we hope we have provided some useful input accordingly. But it is worth pointing out that we, collectively, still have a lot to learn about data sharing and its risks and benefits. We are at the early stages of broad data sharing efforts, and technologies for both sharing and protecting data are evolving rapidly. Until we fully understand the risks and benefits of data sharing, we urge the NIH, in its leadership role, to continue to monitor both the utilization of data sharing strategies and the barriers to their use, to learn from the successes and failures of methods used to protect people’s privacy and enhance their welfare, and to incorporate what is learned into its communications with the research community, and into its own policies.

Thank you again for the opportunity to comment on this important issue. My PRIM&R colleagues and I are available to discuss our comments further, should that be of interest. We look forward to the next stage of policymaking in this area. Please feel free to contact me at 617.303.1872 or ehurley@primr.org.

Respectfully submitted,



Elisa A. Hurley, PhD
Executive Director

cc: PRIM&R Public Policy Committee, PRIM&R Board of Directors

Submission #57

Date: 12/04/2018

Name: Rebecca H. Li

Name of Organization: Vivli, Inc.

Type of Organization: Nonprofit Research Organization

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

Clinical research data sharing

Attachment:

Vivli Center for Global Clinical Research Data

Submitted electronically via <https://osp.od.nih.gov/provisions-data-managment-sharing/>.

Comment re: **Request for Information on Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research**

As a non-profit platform for storing and sharing individual participant-level data (IPD) from completed clinical trials, Vivli commends and supports the NIH in promoting and requiring the submission of data sharing plans. If instituted, this proposed provision will significantly expand the promotion of data sharing and transparency.

Comment

Definition of Scientific Data:

Regarding the first aspect of this request for comment, the definition of Scientific Data, Vivli particularly supports the inclusion of the following sentences:

- For the purposes of a possible Policy, **scientific data** may include certain individual level and summary or aggregate data, as well as **metadata**. NIH expects that reasonable efforts should be made to digitize all scientific data.
- In order to maximize the value of shared clinical data, the NIH should encourage or require that investigators and sponsors share both IPD and aggregate data where appropriate. Also, digitized data are easier to store and share, and allow for greater utility for secondary or meta-analysis.

Although the culture is changing, and aggregate data are shared in a structured format on clinicaltrials.gov, IPD are not routinely shared for all NIH sponsored studies. There are some organizations (NHLBI's BioLINCC repository for example) that share the aggregate and IPD-level datasets in a well-organized and easy to retrieve manner.

General Requirement and Plan Elements:

Regarding the general requirement for Data Management and Sharing plans:

- (1) Vivli strongly supports the allowance of data sharing costs in grant budgets.
 - Given that some data require additional levels of security or control (for example, IPD) to minimize the risk to participants of re-identification, NIH should allow for some data-sharing options or plans that deviate from open access standards provided there are additional safeguards. See *Limitations on Access* below.
 - While requiring and evaluating proposals on their statement of *plans* is a prerequisite to achieving data sharing, it is critical for NIH to provide or point researchers to actual capacity for data sharing. In order to move into implementation of these plans, researchers should have easy, reasonably-priced access to trusted FAIR (Findable, Accessible, Interoperable,

Reusable) data sharing¹ without each research group having to manage the entire end-end complex processes.

- (2) Regarding the specific plan elements suggested by the NIH: “Plans could have a two-page limit and address the following research elements: (i) data types, (ii) related tools and software, (iii) data standards, (iv) data preservation, access (including timelines) and discoverability, (v) terms for re-use and redistribution, (vi) limitations on access, and (vii) oversight of data management.”
- All seven of these items represent key elements of a useful data sharing effort or plan.
 - Data type, related tools, and software and standards are important information for any secondary research or meta-analysis. Regarding data standards, requiring a consistent standard for data increases future utility and interoperability of shared data.
 - Data preservation, access, and discoverability must be included in order to prevent data silos, data destruction, and lack of accessibility.
 - Terms for re-use and redistribution, as well as limitations on access, can serve to protect sensitive data while still promoting data sharing and transparency. Allowing for some limitations on access does not necessarily run contrary to the spirit or concept of data sharing and transparency. Rather, some data can only be shared appropriately while allowing for certain limitations on access. In the interests of maximizing the amount of data that is shared, NIH should allow and accept a range of access control levels.
 - For example, IPD-level clinical data may require managed access, in order to protect the privacy of the individual participants and prevent any attempts at participant re-identification. IPD-level clinical data also require managed access in order to make sure that the participants’ consent for sharing has been obtained where applicable. Given the special protections afforded to data derived from human subjects research under the Common rule and other regulations and policies, it is appropriate to allow for managed, rather than open, access to these data. Managed access allows for data sharing while still respecting the sensitive nature of IPD-level clinical data.
 - Oversight of data management will be of particular importance as these plans are implemented.

Phasing and Implementation

There are already significant resources and infrastructure available to facilitate the implementation of data sharing plans. The NIH should ensure that sponsors and investigators have a means to store and share their IPD from completed clinical trials. In addition to satisfying the requirements of a data sharing plan such as the NIH proposed policy, such access is also necessary for purposes of supporting publication, meta-analysis, and other secondary uses including data aggregation.

¹ See generally: Wilkinson, Mark D, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. “The FAIR Guiding Principles for Scientific Data Management and Stewardship.” *Scientific Data* 3 (March 15, 2016): 160018. <http://dx.doi.org/10.1038/sdata.2016.18>.

Recommendations

General Requirement and Plan Elements:

- We strongly recommend that NIH involve the community to define and establish clear criteria for NIH approval of data repositories, manage the repository approval process, and ensure that the research community is aware of and has easy access to approved repositories.

Phasing and Implementation:

- Depending on the nature of the data, an investigator should be required to provide either managed or open access to a digitized, de-identified data package.
- Because there are costs to de-identification, consideration should be given to allowing de-identification to be deferred unless and until a request is made for that IPD.
- Data repositories should also adopt the assignment of unique Digital Object Identifiers for all data sets. This improves tracking of future use / access, as well as giving researchers a means to show the utility and accessibility of their data while preserving the integrity of the original data set. Additional community expectations for data repositories include the ability to search stored data, track data access, and manage access as needed.

In conclusion, Vivli strongly supports the adoption of a future requirement for data sharing plans. The plans must be adequately specific so as to be meaningful, but sufficiently flexible to allow for managed access, in order to balance data security with utility and transparency. The NIH should also actively promote the actual capacity and practicality of data sharing to facilitate researchers meeting the requirements. A critical role for NIH is to solicit additional assistance from the community to define and establish clear criteria for NIH approval of data repositories to guide both repositories and researchers towards best practices that will materially advance data sharing.

.....

Vivli thanks the NIH for the “Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research”; we appreciate the opportunity to provide comments for your further consideration. We hope the agency finds these comments helpful as you finalize the policy. We look forward to providing additional comments regarding the policy provisions.

Respectfully submitted,

Rebecca Li, PhD

Ida Sim, MD, PhD

On behalf of Vivli Center for Global Clinical Research Data

Submission #58**Date:** 12/04/2018**Name:** Sarah Wright**Name of Organization:** Cornell University**Type of Organization:** University**Role:** Scientific Researcher**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Veterinary Medicine, Engineering, Human Ecology and Nutrition, molecular biology and genetics, chemistry, etc.

II. The requirements for Data Management and Sharing Plans

NSF sections are somewhat confusing, with overlap between sections. More clarity around formatting and content of DMP would be helpful. Researchers ask consistently for a template, so it's worth considering whether the plan could have a more defined structure.

Researchers often prefer a longer time-line, for example: suggest that the data are public one year from the END DATE of the grant, or one year from the publication date.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

Enforcement in US is difficult due to a very different culture than that in Europe. Education and outreach about importance of sharing will be key, unless get much more stringent around enforcement.

Submission #59

Date: 12/04/2018

Name: C. Titus Brown

Name of Organization: University of California Davis

Type of Organization: University

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

Genomics

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

Data management and sharing is a diverse and challenging sociotechnical problem that is resistant to top-down mandates. I believe the NIH should consider investing in organic development of infrastructure, resources, and especially standards, with the goal of slowly gaining community adoption rather than defining a single "right" approach and imposing it.

Submission #60**Date:** 12/05/2018**Name:** Jonathan Petters**Name of Organization:** Data Services, Virginia Tech University Libraries**Type of Organization:** University**Role:** Institutional Official**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Virginia Tech supports the (new) Carilion School of Medicine and Research Institute, as well as the College of Science and the College of Agricultural and Life Sciences. These colleges (and others) submit funding proposals to a variety of NIH Institutes.

I. The definition of Scientific Data

This definition closely mirrors the federal government's definition of research data, and as such promotes

harmony between the NIH's definition and other US research funding agencies. I support this choice, and this

policy should therefore cite uniform guidance within the Federal Register as its origin

(<https://www.federalregister.gov/documents/2013/12/26/2013-30465/uniform-administrative-requirements-costprinciples->

[and-audit-requirements-for-federal-awards#sec-200-315](#))

One addition I strongly recommend is that it be made clear that metadata is part of scientific data. Without

appropriate metadata, scientific data that is shared is far less useful (if useful at all).

Metadata could alternatively be defined as "documentation that describes data, providing the context under which it was collected, processed and interpreted". I am not wedded to any particular definition, but a definition of metadata that will be meaningful to the researchers and data managers who create metadata will be optimal

if this policy is to have a positive impact on NIH-funded research. I am not sure the definition of metadata

currently in the policy accomplishes this.

II. The requirements for Data Management and Sharing Plans

I recommend a separate data management plan element section entitled “Ethical and Legal Compliance”

focusing on ethical and legal issues that are associated with the data types to be generated within the

proposed research. Data Preservation and Access, Data Sharing Agreements, Licensing and Intellectual

Property, and Oversight of Data Management (Elements 4 to 7 under Section IV.) are all dependent on what

ethical or legal issues surround the data. It would behoove the proposal writer to ponder these issues before

determining how they can store, secure and share their data.

Additionally, I encourage this policy to be modified to encourage NIH-funded researchers to share datasets

and software be in repositories that provide data and software citations. Further, NIH-funded researchers

should be encouraged to cite datasets and software they use, including in their proposals for NIH funding.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

For the first phase of implementation, scientific data could be initially defined more narrowly as data directly underlying a research publication, coupled with the understanding that this definition will be expanded in the future.

What data and software assets are generated by NIH funds is to a large extent currently unknown. Therefore, this policy can not yet be prescriptive about in which repositories research data generated by NIH-funded are to be deposited. After the NIH understands what data are generated through NIH-funds and where repository infrastructure exists and is lacking, NIH can be more prescriptive.

In the meantime, institutional repositories administered by university libraries like that I administer (VTechData, <https://data.lib.vt.edu>) can play a helpful role in curating this data for now (and in the future).

It would be useful for policy implementers and policy followers to understand that the policy will be an instrument by which NIH will learn about what data is produced through its funds, and in concert with the development of the NIH Data Commons model the NIH will be in a better position to recommend other data sharing infrastructure development.

Weaving the public sharing of research data and software into the current academic credit system is vital to incentivize this change of academic practices. Perhaps NIH can not currently mandate the use of data repositories that provide citations by which credit can be given to data creators and contributors, but NIH should mandate this as soon as possible. Further, in the future NIH should mandate NIH-funded researchers to cite datasets and software they use.

Attachment:

Comment on the [Request for Information on Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research](#)

This comment is informed by my experiences as

- a AAAS Science and Technology Policy fellow in DOE-Science when it developed and released its own statement on digital data management (<https://science.energy.gov/funding-opportunities/digital-data-management/>), and
- as the primary content creator for the Scholarly Publishing and Academic Resources Coalition data sharing resource (<http://datasharing.sparcopen.org/data>) while a Data Management Consultant in Johns Hopkins University Libraries

About the draft policy as a whole:

In general this policy is reasonable and in line with the data management and sharing policies that NSF, DOE-Science and USDA currently have. Again, this promotes harmony between the NIH's policies and other US research funding agencies and should lower the administrative burden for US funded researchers who are familiar with data management and sharing policies from these other agencies.

Training around how researchers will meet these new requirements and how NIH program officers will evaluate compliance towards this policy is critical. This will include frequently asked question collections (e.g. <https://science.energy.gov/funding-opportunities/digital-data-management/faqs/>), tutorials, and in-person training sessions for researchers and program officers alike.

I especially emphasize motivating NIH program officers about the reasons for this policy enactment. If they do not understand the purpose of this policy it could quickly become a check-the-box activity for NIH researchers. In this case this policy could raise the administrative burden across the NIH research enterprise without adding value to that enterprises. This scenario is to be avoided.

Regarding the definition of scientific data (this section copied into section I):

This definition closely mirrors the federal government's definition of research data, and as such promotes harmony between the NIH's definition and other US research funding agencies. I support this choice, and this policy should therefore cite uniform guidance within the Federal Register as its origin (<https://www.federalregister.gov/documents/2013/12/26/2013-30465/uniform-administrative-requirements-cost-principles-and-audit-requirements-for-federal-awards#sec-200-315>)

One addition I strongly recommend is that it be made clear that metadata is part of scientific data. Without appropriate metadata, scientific data that is shared is far less useful (if useful at all).

Metadata could alternatively be defined as "documentation that describes data, providing the context under which it was collected, processed and interpreted". I am not wedded to any particular definition, but a definition of metadata that will be meaningful to the researchers and data managers who create metadata will be optimal if this policy is to have a positive impact on NIH-funded research. I am not sure the definition of metadata currently in the policy accomplishes this.

Regarding the requirements for data management and sharing plans (this section copied into section II):

I recommend a separate data management plan element section entitled “Ethical and Legal Compliance” focusing on ethical and legal issues that are associated with the data types to be generated within the proposed research. Data Preservation and Access, Data Sharing Agreements, Licensing and Intellectual Property, and Oversight of Data Management (Elements 4 to 7 under Section IV.) are all dependent on what ethical or legal issues surround the data. It would behoove the proposal writer to ponder these issues before determining how they can store, secure and share their data.

Additionally, I encourage this policy to be modified to encourage NIH-funded researchers to share datasets and software be in repositories that provide data and software citations. Further, NIH-funded researchers should be encouraged to cite datasets and software they use, including in their proposals for NIH funding.

Regarding the timing and phasing of implementation (this section copied into section III):

For the first phase of implementation, scientific data could be initially defined more narrowly as data directly underlying a research publication, coupled with the understanding that this definition will be expanded in the future.

What data and software assets are generated by NIH funds is to a large extent currently unknown. Therefore, this policy can not yet be prescriptive about in which repositories research data generated by NIH-funded are to be deposited. After the NIH understands what data are generated through NIH-funds and where repository infrastructure exists and is lacking, NIH can be more prescriptive.

In the meantime, institutional repositories administered by university libraries like that I administer (VTechData, <https://data.lib.vt.edu>) can play a helpful role in curating this data for now (and in the future).

It would be useful for policy implementers and policy followers to understand that the policy will be an instrument by which NIH will learn about what data is produced through its funds, and in concert with the development of the NIH Data Commons model the NIH will be in a better position to recommend other data sharing infrastructure development.

Weaving the public sharing of research data and software into the current academic credit system is vital to incentivize this change of academic practices. Perhaps NIH can not currently mandate the use of data repositories that provide citations by which credit can be given to data creators and contributors, but NIH should mandate this as soon as possible. Further, in the future NIH should mandate NIH-funded researchers to cite datasets and software they use.

Regards,

Jonathan Petters Ph.D.

Data Management Consultant and Curation Services Coordinator

Data Services, University Libraries

Virginia Tech

jpeters@vt.edu

(540) 232-8682

<https://www.lib.vt.edu/research-learning/ResearchDataManagementAndCuration.html>

ORCID: 0000-0002-0853-5814

Submission #61**Date:** 12/05/2018**Name:** Mary Ellen K. Davis, Executive Director ACRL**Name of Organization:** Association of College and Research Libraries**Type of Organization:** Professional Org/Association**Role:** Institutional Official**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

The Association of College and Research Libraries (ACRL) is the division in the American Library Association that serves more than 10,000 academic and research librarians and interested individuals working in institutions of higher education. ACRL develops programs, products, and services to help academic and research librarians learn, innovate, and lead within the academic community. We enhance the ability of academic library and information professionals to serve the information needs of students and researchers. For example, through a one-day workshop, ACRL presenters travel to campuses across the U.S. and train liaison librarians to enhance their skills with research data management. As reflected in our previous support for governmental policies and legislation that facilitate open access and open education -- including the NIH Open Access Policy, the Office of Science and Technology Policy mandate, and the Fair Access to Science & Technology Research Act and Federal Research Public Access Act bills -- ACRL is fundamentally committed to the open exchange of information to empower individuals and facilitate scientific discovery.

I. The definition of Scientific Data

NIH's definition is generally consistent with the 2013 OSTP Memo "Increasing Access to the Results of Federally Funded Scientific Research"

(https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf), with OMB circular A-110 (<https://www.whitehouse.gov/wp-content/uploads/2017/11/Circular-110.pdf>) and 2 CFR § 200.315 - Intangible property (<https://www.gpo.gov/fdsys/granule/CFR-2014-title2-vol1/CFR-2014-title2-vol1-sec200-315>), which is helpful to bring alignment across the federal landscape.

The definition in the Proposed Provisions specifically excludes laboratory notebooks and case reports, which would be in agreement with the previous definitions. ACRL believes that case report forms should not be excluded even though they may contain personnel and medical

information of which a disclosure would be an unwarranted invasion of personal privacy. Instead, ACRL encourages NIH to include case reports, other medical records, or data containing PII in the definition of scientific data and clearly note that researchers should share them in accordance with federal policy and other best practices (e.g., HIPAA, restricted sharing, aggregation to a level that will reduce the possibility of disclosure).

ACRL also requests that NIH reconsider the exclusion of laboratory notebooks, as their exclusion is in tension with Section V, Part 1.2 of the Proposed Provisions, which states that the DMP must:

"Describe any other information that is anticipated to be shared along with the scientific data, such as relevant associated data, and any other information necessary to interpret the data (e.g., study protocols and data collection instruments)."

Laboratory notebooks include recorded information that is "necessary to interpret the data." NIH should consider requiring that the Data Management Plan address how laboratory notebooks will be managed and how the information contained within them will be shared.

II. The requirements for Data Management and Sharing Plans

An NIH requirement for a Data Management and Sharing Plans at all funding levels would be a new requirement, presumably overriding what is set out in NIH's Data Sharing Policy and Implementation Guidance (https://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm). The expansion to include all funding levels, wholly or partially funded by NIH, helps bring NIH in closer alignment with other federal agencies and creates a more comprehensive treatment of data in the funding landscape. A new NIH Data Management and Sharing Policy based on the proposed revisions has the potential to clarify the importance of the data management and sharing by creating mechanisms to ensure that researchers follow it.

Part V provides the potential for stronger compliance and enforcement mechanisms, although it may be worthwhile to consider how the data management and sharing plan compliance could be integrated into eRA Commons and MyNCBI, to create a similar workflow as exists for publication public access compliance via PubMed Central. Moreover, ACRL encourages NIH to provide the guidance for making data shareable via the NIH Data Catalog, or available via PMC and linked to any published articles. Providing guidance and low-burden interfaces to researchers will help adoption of NIH-supported public access methods, which should reinforce the parameters laid out in this proposal.

In Section II, NIH proposes that scientific data should be "made accessible in a timely manner for appropriate use by the research community and broader public." It goes on to state that any new NIH policy would establish requirements for responsible management and sharing. We suggest that any policy NIH creates should have a clear definition for what "timely" and "appropriate" mean. Given the diversity of domain engagement with NIH, "timely" may have

very different interpretations by the community. Looking to other federal agencies for precedent, directorates across NSF have dictated the embargo periods in the data management plan guidance.

Within the Proposed Provisions (Section IV), NIH suggests that Data Management Plans (DMPs) remain an Additional Review Consideration. Although this is one method for considering DMPs, because Additional Review Considerations are not individually scored and do not influence the overall score, ACRL encourages NIH to consider designating the DMP as Additional Review Criteria and incorporating review of the DMP in the overall impact score. Failing that designation, ACRL encourages NIH to expand upon when and to what degree this integration would be appropriate.

A well-conceived and well-described DMP requires significant investment of time for grant applicants and conveying such may well require more than the proposed limit of two pages. Although this could be required at the time of submission, it would be more reasonable to require the detailed DMP as a condition and term of the award. A detailed DMP required at the time of award would outline specifics that would be incorporated into the terms and conditions, and NIH could provide support to ensure that investigators' plans are appropriate and actionable.

Relatedly, it is impossible to predict changes in technology standards over the life of a research grant. ACRL suggest that NIH explicitly allow the DMP to be revised as part of the annual report process. This would ensure that researchers are following the most up-to-date standards and increase the appropriate and successful preservation of data.

Section IV part 2 adds that, "the inclusion of scripts may be helpful." ACRL encourages NIH to include a stronger statement requiring the inclusion of scripts and require a justification from the researchers as a decision to use non-open source software and code. Access to scripts (which would include having access to open source software used to create and run them) is necessary for research reproducibility.

Section IV part 4 states that, "If an existing repository will not be used, indicate why not and how scientific data preservation will be assured (e.g., in a newly created repository or by the investigator's organization)." ACRL encourages NIH to offer more explicit guidance to the researcher explaining what minimally adequate preservation (e.g., exhibit a sustainable funding model, provide a succession plan) is acceptable, as long-term preservation requirements are not common knowledge across all researchers. Section V should include a requirement for long-term planning that "meets community-based standards at the time of deposition."

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

With robust guidance and infrastructure in place from NIH, a year of community preparation could be sufficient to bring about adoption of this proposal. Researchers seeking NIH funding may have some experience in planning for the sharing of data from funded research, but a new policy as proposed in the Proposed Provisions would represent a significant change for grant applicants. We recognize the need for additional clarification and support for investigators seeking funding due to the inherent difficulties in writing a thorough and actionable DMP. NIH should provide clear guidelines and recommendations for researchers, including working with their research support partners, such as the library, on campus.

Scientific data standards are an area in which researchers may need additional information. In support of Section IV part 3, NIH could provide more assistance to proposal authors to help them better understand existing data standards, which common data elements would be appropriate, and how they should be applied. Providing tutorials or other learning objects in the call for proposals could help disseminate information to researchers. Providing embeddable learning objects also allows for librarians and other research supporting offices to reinforce these standards through other delivery avenues.

Section III of the proposed provisions states that, “[r]easonable costs associated with data management and sharing could be requested under the budget for the proposed project.” There may be significant costs associated with implementing a quality data management and sharing plan, and ACRL applauds the NIH acknowledging this in the budget allowance.

Attachment:

Association of College & Research Libraries
50 E. Huron St. Chicago, IL 60611
800-545-2433, ext. 2523
acr@ala.org, <http://www.acrl.org>



TO: National Institutes of Health
DATE: Wednesday, December 5, 2018
RE: Response to Proposed Provisions for a draft NIH Data Management and Sharing Policy

Submitted online at <https://osp.od.nih.gov/provisions-data-managment-sharing/>

Name: Mary Ellen K. Davis, Executive Director ACRL
Name of Organization: Association of College and Research Libraries
Type of Organization: Professional Org/Association
Role: Institutional Official

Research Area Most Important to You or Your Organization (e.g., clinical, genomics, neuroscience, infectious disease, epidemiology)

The Association of College and Research Libraries (ACRL) is the division in the American Library Association that serves more than 10,000 academic and research librarians and interested individuals working in institutions of higher education. ACRL develops programs, products, and services to help academic and research librarians learn, innovate, and lead within the academic community. We enhance the ability of academic library and information professionals to serve the information needs of students and researchers. For example, through a one-day workshop, ACRL presenters travel to campuses across the U.S. and train liaison librarians to enhance their skills with research data management. As reflected in our previous support for governmental policies and legislation that facilitate open access and open education -- including the NIH Open Access Policy, the Office of Science and Technology Policy mandate, and the Fair Access to Science & Technology Research Act and Federal Research Public Access Act bills -- ACRL is fundamentally committed to the open exchange of information to empower individuals and facilitate scientific discovery.

NIH welcomes comments on any aspect of the issues presented, it is particularly interested in comments on

I. The definition of Scientific Data.

NIH's definition is generally consistent with the 2013 OSTP Memo "Increasing Access to the Results of Federally Funded Scientific Research" (https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf), with OMB circular A-110 (<https://www.whitehouse.gov/wp-content/uploads/2017/11/Circular-110.pdf>) and 2 CFR § 200.315 - Intangible property (<https://www.gpo.gov/fdsys/granule/CFR-2014-title2-vol1/CFR-2014-title2-vol1-sec200-315>), which is helpful to bring alignment across the federal landscape.

The definition in the Proposed Provisions specifically excludes laboratory notebooks and case reports, which would be in agreement with the previous definitions. ACRL believes that case report

forms should not be excluded even though they may contain personnel and medical information of which a disclosure would be an unwarranted invasion of personal privacy. Instead, ACRL encourages NIH to include case reports, other medical records, or data containing PII in the definition of scientific data and clearly note that researchers should share them in accordance with federal policy and other best practices (e.g., HIPAA, restricted sharing, aggregation to a level that will reduce the possibility of disclosure).

ACRL also requests that NIH reconsider the exclusion of laboratory notebooks, as their exclusion is in tension with Section V, Part 1.2 of the Proposed Provisions, which states that the DMP must:

Describe any other information that is anticipated to be shared along with the scientific data, such as relevant associated data, and any other information necessary to interpret the data (e.g., study protocols and data collection instruments).

Laboratory notebooks include recorded information that is “necessary to interpret the data.” NIH should consider requiring that the Data Management Plan address how laboratory notebooks will be managed and how the information contained within them will be shared.

II. The requirements for Data Management and Sharing Plans.

An NIH requirement for a Data Management and Sharing Plans at all funding levels would be a new requirement, presumably overriding what is set out in NIH’s Data Sharing Policy and Implementation Guidance

(https://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm). The expansion to include all funding levels, wholly or partially funded by NIH, helps bring NIH in closer alignment with other federal agencies and creates a more comprehensive treatment of data in the funding landscape. A new NIH Data Management and Sharing Policy based on the proposed revisions has the potential to clarify the importance of the data management and sharing by creating mechanisms to ensure that researchers follow it.

Part V provides the potential for stronger compliance and enforcement mechanisms, although it may be worthwhile to consider how the data management and sharing plan compliance could be integrated into eRA Commons and MyNCBI, to create a similar workflow as exists for publication public access compliance via PubMed Central. Moreover, ACRL encourages NIH to provide the guidance for making data shareable via the NIH Data Catalog, or available via PMC and linked to any published articles. Providing guidance and low-burden interfaces to researchers will help adoption of NIH-supported public access methods, which should reinforce the parameters laid out in this proposal.

In Section II, NIH proposes that scientific data should be “made accessible in a timely manner for appropriate use by the research community and broader public.” It goes on to state that any new NIH policy would establish requirements for responsible management and sharing. We suggest that any policy NIH creates should have a clear definition for what “timely” and “appropriate” mean. Given the diversity of domain engagement with NIH, “timely” may have very different interpretations by the community. Looking to other federal agencies for precedent, directorates across NSF have dictated the embargo periods in the data management plan guidance.

Within the Proposed Provisions (Section IV), NIH suggests that Data Management Plans (DMPs) remain an Additional Review Consideration. Although this is one method for considering DMPs, because Additional Review Considerations are not individually scored and do not influence the overall score, ACRL encourages NIH to consider designating the DMP as Additional Review Criteria and incorporating review of the DMP in the overall impact score. Failing that designation, ACRL encourages NIH to expand upon when and to what degree this integration would be appropriate.

A well-conceived and well-described DMP requires significant investment of time for grant applicants and conveying such may well require more than the proposed limit of two pages. Although this could be required at the time of submission, it would be more reasonable to require the detailed DMP as a condition and term of the award. A detailed DMP required at the time of award would outline specifics that would be incorporated into the terms and conditions, and NIH could provide support to ensure that investigators' plans are appropriate and actionable.

Relatedly, it is impossible to predict changes in technology standards over the life of a research grant. ACRL suggest that NIH explicitly allow the DMP to be revised as part of the annual report process. This would ensure that researchers are following the most up-to-date standards and increase the appropriate and successful preservation of data.

Section IV part 2 adds that, "the inclusion of scripts may be helpful." ACRL encourages NIH to include a stronger statement requiring the inclusion of scripts and require a justification from the researchers as a decision to use non-open source software and code. Access to scripts (which would include having access to open source software used to create and run them) is necessary for research reproducibility.

Section IV part 4 states that, "If an existing repository will not be used, indicate why not and how scientific data preservation will be assured (e.g., in a newly created repository or by the investigator's organization)." ACRL encourages NIH to offer more explicit guidance to the researcher explaining what minimally adequate preservation (e.g., exhibit a sustainable funding model, provide a succession plan) is acceptable, as long-term preservation requirements are not common knowledge across all researchers. Section V should include a requirement for long-term planning that "meets community-based standards at the time of deposition."

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards.

With robust guidance and infrastructure in place from NIH, a year of community preparation could be sufficient to bring about adoption of this proposal. Researchers seeking NIH funding may have some experience in planning for the sharing of data from funded research, but a new policy as proposed in the Proposed Provisions would represent a significant change for grant applicants. We recognize the need for additional clarification and support for investigators seeking funding due to the inherent difficulties in writing a thorough and actionable DMP. NIH should provide clear guidelines and recommendations for researchers, including working with their research support partners, such as the library, on campus.

Scientific data standards are an area in which researchers may need additional information. In support of Section IV part 3, NIH could provide more assistance to proposal authors to help them better understand existing data standards, which common data elements would be appropriate, and how they should be applied. Providing tutorials or other learning objects in the call for proposals could help disseminate information to researchers. Providing embeddable learning objects also allows for librarians and other research supporting offices to reinforce these standards through other delivery avenues.

Section III of the proposed provisions states that, “[r]easonable costs associated with data management and sharing could be requested under the budget for the proposed project.” There may be significant costs associated with implementing a quality data management and sharing plan, and ACRL applauds the NIH acknowledging this in the budget allowance.

Submission #62

Date: 12/05/2018

Name: Paul Anderson

Name of Organization: Brigham and Women's Hospital

Type of Organization: Other

Other Type of Organization: Hospital

Role: Institutional Official

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

Clinical

Attachment:

DRAFT – 12.5.18

Francis S. Collins, MD, PhD
National Institutes of Health
Bethesda, MD

Submitted electronically: <https://osp.od.nih.gov/provisions-data-managment-sharing/>

Re: RFI on Proposed Provisions for Draft Data Management and Sharing Policy

Dear Dr. Collins:

Thank you very much for providing the research community with an opportunity to comment on the NIH proposed data management/sharing policy. I am writing on behalf of the Brigham and Women's Hospital (Brigham), a principal teaching affiliate of Harvard Medical School. The Brigham was founded in 1980 with the merger of three of Boston's oldest and most prestigious Harvard teaching hospitals: the Peter Bent Brigham Hospital, the Robert Breck Brigham Hospital, and the Boston Hospital for Women. A founding member of Partners HealthCare System, the Brigham is known for its clinical, translational, bench and population-based research and is consistently ranked among the top two hospital recipients of NIH funding. In FY 19, the Brigham received approximately \$382 million in NIH/HHS research support. Thus, any proposed change in NIH data management and sharing requirements is of vital interest to us.

Let me begin by stating my colleagues and I conceptually support data sharing as a means of enabling researchers to test the validity of scientific findings, explore new scientific pathways, and shorten the time for ideas to move from the bench to the bedside. Yet, the devil is in the details for data sharing to be successful. The proposed policy is so broad and all-encompassing, we believe if implemented it would be extremely difficult for the NIH to achieve its objective of enhancing science, let alone for Principal Investigators (PI) and institutions to meet their compliance requirements.

Some of our investigators have suggested that the proposed policy appears to be an extension of data sharing requirements for genetic data to scientific data more generally. Genetic data sharing through dbGaP and similar repositories works because genetic data can be supported with standard file formats for data submission. We find it difficult to envision how the many possible experimental designs for laboratory-based experiments would be submitted and archived in a way that could be interpreted by an outside user.

We strongly recommend that the NIH revise the proposed policy to scale back its requirements, add clarity to definitions, and provide meaningful examples for investigators. We also recommend that the NIH consider convening a group of NIH-funded investigators to work with NIH research and administrative leadership to develop a policy that is more realistic and achievable from an investigator's perspective.

Please see below for our comments on specific sections of the proposed policy.

1. Section I

- a. Definitions: The definition of Scientific Data is extremely broad and confusing. We recommend considering the definition of Research Data in OMB Circular A-110 as a substitute. This definition would already be familiar to most of the research community.
 - b. Lab notebooks: Throughout the policy there is confusion about lab notebooks and whether they should be shared. Their role/purpose in a “data sharing policy” should be clarified. We maintain that lab notebooks, while critical to the scientific process, are not Scientific Data; they are a means for recording experiments and the Scientific Data generated.
 - c. Reasonable effort to digitize scientific data: While institutionally we are requiring our investigators to transition to digital recordkeeping, we do not recommend including a statement about digitizing scientific data within the current policy. Not all Scientific Data can be digitized; this makes the data no less valuable to research.
2. Section II. Purpose: Making Scientific Data accessible in a “timely manner:”
 Researchers generate data daily. We recommend clarifying this section by adding timelines for posting/sharing published and unpublished data. We recommend adding a section to the Progress Report where the PI can inform the NIH of data accessibility. The policy should be flexible. Not all data will be ready for sharing or posting in a repository at the same time. Investigators may want to refrain from posting/sharing unpublished data until it has been published. These situations should be taken into consideration in this section.
3. Section III. Scope and Requirements:
- a. We are concerned that requiring a data management/sharing plan for each application/proposal submission, when the overall funding success rate hovers at 20% or less, creates a significant administrative burden for PIs submitting applications. We recommend the NIH consider requiring the plan as part of the first progress report. These plans will not have the benefit of peer review, but is peer review necessary if the strength or weaknesses of the plan will not be considered in the impact score? Continuation funding for year 2 could be delayed until a plan acceptable to the Program Officer is submitted.
 - b. In the general statement that data management/sharing plans will be required regardless of mechanism, we recommend that the NIH review the different funding mechanisms for appropriateness. For example, a data sharing plan would not be appropriate for a shared instrumentation grant; nor would it be appropriate for a conference grant. We also recommend that the NIH consider eliminating the requirement for institutional training grant applications. We recognize that Scientific Data are generated under training grants, but the management and sharing of the data will vary across the training grant based on requirements of each trainee’s mentor who often come from different departments/research labs with different data management/sharing requirements.
 - c. The policy states, “Reasonable costs associated with data management and sharing could be requested under the budget for the proposed project.” Will supplemental funds be available for these costs? If not, the data

management/sharing requirement will only reduce the amount of funding available for the actual research project. We can envision situations where institutions with limited resources will have to provide their investigators with institutional funds or create local data repositories because the NIH funds were simply not enough to complete the project and pay for the costs associated with external data repositories essentially creating yet another unfunded mandate for grantee institutions.

4. Section IV. Requirements for Data Management and Sharing Plans

- a. General comment: We do not believe PIs will be able to provide all the information the NIH is requiring within a two-page limit.
- b. Scoring/Peer Review Process: If the NIH continues to require plans as part of the grant application/contract proposal, we agree whether a plan is acceptable or unacceptable to reviewers should not be included in the overall impact score.
- c. Plan Elements: We recommend that the NIH create a form with drop down boxes for the PI to identify the plan elements relevant for his/her research. The elements should be minimal and allow for PI flexibility.
- d. Describe type and amount of scientific data to be collected and used in the project: This may be difficult for some types of projects. The example provided is for a specific type of project in which the number of cases/patients/individuals may be known at submission. In many lab-based projects, investigators may improvise and adjust the work making use of techniques that may not have been envisaged initially. We are concerned that PIs may feel providing this type of information will restrict their ability to modify the research as they move forward.
- e. Related Tools, Software and/or Code: Please clarify what the NIH is expecting. For example, would the PI have to justify use of a specific image analysis software product?
- f. 4.1 Indicate where Scientific Data will be archived to ensure long-term preservation: We recommend that the NIH create data repositories to meet this new mandate. As we indicated above, many institutions do not have the resources to develop and maintain repositories for their NIH-funded investigators. Grantee institutions cannot continue to absorb unfunded mandates. Moreover, we are concerned at the possible development of numerous and heterogeneous and possibly rogue repositories.
- g. 4.4 Describe alternative plans for maintaining, preserving and providing access to scientific data should the original plan not be achieved: If the NIH is truly interested in this information, we recommend not requiring submission of a “Plan B” as part of the data management/sharing plan in their application/contract proposal. We recommend adding a section to the data management/sharing reporting section of the annual progress report to describe any changes necessary because the original plan could not be achieved.
- h. 5. Data Preservation and Access Timeline: We question the usefulness of requiring this information in the data management/sharing plans. It may be impossible at the beginning of the project to estimate timelines. This may lead PIs to develop meaningless timelines which become a compliance requirement if the application is funded. We recommend removing this requirement.

- i. 6. Data Sharing Agreements, Licensing and Intellectual Property:
 - i. “NIH encourages terms that provide for the broadest use of data resulting from NIH-funded or -supported research.” Please confirm/clarify that this statement applies to data generated as part of the study, i.e., the data would not exist if not for the study; and does NOT include any additional, pre-existing clinical data, e.g., annotated, longitudinal data pulled from a patient’s medical record.
 - ii. 6.1 “Describe any relevant data sharing agreements outlining...how scientific data can and cannot be used.” Please confirm/clarify that this applies only to Scientific Data generated as part of the study. In a situation where the project is supported by NIH and industry or a foundation, the non-NIH sponsors may limit data sharing. Would an SBIR grant be relevant here?
 - iii. 6.3 “[I]ndicate how intellectual property...will be managed in a way to maximize sharing of scientific data.” While Scientific Data do not constitute IP, any plan to maximize sharing should not infringe upon the nature of the IP and should preserve ownership rights.
5. Compliance and Enforcement
- a. Community-based Standards: The NIH should specify these standards within the policy or at a minimum provide examples. When we consulted our investigators to develop our response, they were unsure what the standards were and where they might find them.
 - b. I/C Monitoring Plans: The policy should include information on how I/Cs will monitor plans, reporting requirements, how to modify plans during the lifetime of the grant. If an I/C determined non-compliance, what would be the enforcement mechanism?
 - c. We are very concerned about compliance/enforcement requirements extending beyond the end of the grant’s performance period. If this requirement continues in the policy, the NIH should identify the authority that allows the requirement to continue in perpetuity. Comments made during the NIH webinar on the RFI seemed to suggest that the NIH does not consider data sharing requirement as continuing beyond the project end date. The proposed policy contradicts this point and should be clarified. How will the NIH monitor? How will a grantee know if a former award is out of compliance?

Once again, thank you for providing an opportunity for the research community to submit comments. Please do not hesitate to contact me for any additional information.

Yours sincerely,

Paul Anderson, MD, PhD
Chief Academic Officer
Brigham and Women’s Hospital

s is of vital interest to us.

Submission #63**Date:** 12/05/2018**Name:** David Mellor**Name of Organization:** Center for Open Science**Type of Organization:** Nonprofit Research Organization**Role:** Institutional Official**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Neuroscience, social psychology, cognitive psychology, social science, pre-clinical disease research, education research.

I. The definition of Scientific Data

There are generally two distinctions made when considering “the data” in response to funder mandates for increased data transparency: 1) The digitally shareable data that underlie the findings reported in a scientific, peer reviewed article and 2) All data collected over the course of a project supported by a funding agency.

Our recommendation is to focus on the first working definition of “Scientific Data” because of its simplicity, its potential for widespread understanding among most of the research community, and because it is the most concrete and enforceable standard with specific points in time where expectations can be met. That point is the point of publication.

Barriers to implementing a better policy that covers all of the data collected over the course of a project include: A) creating a timeline that can be achievable by many different communities or modes of research (e.g. should data be shared by the time the grant is finished even prior to publication? If not, what timeline is reasonable will depend on the norms and standards of a particular community, and enforcement will be challenging), and B) providing additional support and guidance for dissemination of sharing unpublished data. Implementing such an ideal policy may be beyond current expectations, and achieving the first goal (sharing data that underlie findings reported in an article) would set the stage for future improvement into sharing all data. To be clear: We would strongly support a policy that states “By the end of the grant period, all data collected for the supported work must be shared in a repository,” however, we also abide by the mantra “Don’t let the perfect be the enemy of the good” and believe that the first strategy is more achievable today.

This strategy does risk continuation of known biases against reporting null results. We believe that the best way to address this shortcoming is not by implementing difficult to enforce mandates, but by addressing the incentive to only share “significant” findings, which is that the decision to share or not share (or publish or stick in the file drawer) only occurs after results are known. Though outside the scope of this current RFI, we encourage policy makers to work with journals to determine what is publishable based on the importance of the research question and the rigor of the proposed methods and analyses, before results are known. This model, the Registered Reports initiative, directly addresses the incentive to not share underlying research data and to bias the research literature against null results.

Finally, we strongly discourage the continuation of a double standard between different specific research data types. There is a widespread culture of sharing genomics and proteomics data, for example, and many funder and journal policies point to such requirements. However, there is no philosophical justification for making such a distinction, and expanding the community norms into all underlying, digital data can only happen as entities such as the NIH take reasonable steps, like this proposed plan, to improve the culture by making science more transparent by default.

II. The requirements for Data Management and Sharing Plans

We have four main recommendations for consideration of the proposed requirements:

First, the proposed policy states that “Plans could be evaluated as an Additional Review Consideration.” We strongly discourage this approach. Data management plans should be included as “Scored Review Criteria.” Researchers submitting DMPs will have little incentive to consider the details of these plans if they are not expecting them to be evaluated by the reviewer team and scored alongside the rest of their proposal. Simply giving the DMP a “pass/no pass” evaluation after the grant decision has been made will ensure that minimal thought or effort will be put into the plan. On the other hand, knowing that the DMP will be scored, grant submitters will work to offer the best plan that they are able to write.

Second, we further recommend that the “Plan Elements” section include ethical considerations to ensure later shareability. For example, a simple statement about including plans to anonymize human subjects data could lead authors to expect that solutions should be considered for sharing data. A statement such as “If you are collecting data from human research participants, please include a strategy for anonymizing identifications so that data are publicly shareable. If the nature of your project is to collect data that could be reasonably expected to lead to re-identification even after anonymization, please include a plan to share parts of the datasets that do not suffer that risk, or use a repository that includes protected access for sensitive datasets.” This sets expectations for finding a solution to sharing data.

Third, we encourage the NIH to consider how the accessibility of these data management plans can be increased. Often the grant submissions are treated as intellectual property of the author, not to be made publicly available. The rationale for that decision does not apply to these proposed data management plans, which should themselves be discoverable through the NIH search portal, so that other members of the research community or the general public are able to find where data is expected to be shared. This will increase transparency and accountability.

Finally, the proposed plan states that "Non-compliance with the NIH-approved Plan would be taken into account by the funding IC for future funding or support decisions" It is unclear if this is non-compliance with the proposed plan, which is to include a data management plan and can be evaluated at the time of grant submission (therefore post award evaluation is unnecessary) or if this implies that failure to adhere to one's own data management plan would be taken into account in future funding support decisions. We support the later interpretation. Furthermore, we encourage you to put a mechanism in place to make such enforcement easier. For example, grant applications can include the following line "have you satisfied the spirit of your data management plans for previously funded work? Y/N/na. If yes, state how. If no, please justify."

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

It is our opinion that the infrastructure largely exists for widespread sharing of digital data today. Even for those that believe this is not the case, the scope of this proposed policy, which is to include a data management plan and not a mandate for what needs to be included in that plan, requires no additional infrastructure. Grant applicants who have not written a DMP yet will surely have additional questions about how to make such a plan and what to include in that, but such educational resources are widely available online, through university libraries, and through NIH's existing online resources.

Submission #64

Date: 12/05/2018

Name: James Kent

Name of Organization: University of Iowa

Type of Organization: University

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

Neuroscience

Submission #65

Date: 12/05/2018

Name: Todd Constable

Name of Organization: Yale University

Type of Organization: University

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

Brain imaging, neuroscience, translational medicine.

I. The definition of Scientific Data

I would consider behavioral, genetic, clinical and imaging information data.

II. The requirements for Data Management and Sharing Plans

Raw data (or minimally processed data) should be released annually starting after the 2nd year of a grant. Too many investigators delay data release citing preprocessing or some other complication - raw data should be released asap. Sufficient supporting documentation for other to make sense of the data must be included. There should be no strings attached to using the data. Any and all investigators should have access to the data.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

New grant submissions that follow the announcement of a new data sharing policy should be required to agree with the data sharing policy and when awarded the grants must comply with the sharing.

Submission #66

Date: 12/05/2018

Name: Brooke N. Macnamara

Name of Organization: Case Western Reserve University

Type of Organization: University

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

Behavioral Psychology

II. The requirements for Data Management and Sharing Plans

Open data reported in cases of federal funding if it can be identified. In most cases, made openly available at the time of research product (i.e., publication), but required by the end of the grant period with exceptions allowed (e.g., for data from human subjects that cannot be de-identified) and extensions allowed (e.g., when the dataset will lead to future publications that have not yet been accepted).

Submission #67

Date: 12/06/2018

Name: Kerry Ressler

Name of Organization: McLean Hospital

Type of Organization: Other

Other Type of Organization: Hospital

Role: Institutional Official

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

neuroscience

Attachment:

Francis S. Collins, MD, PhD
National Institutes of Health
Bethesda, MD

Submitted electronically: <https://osp.od.nih.gov/provisions-data-managment-sharing/>

Re: RFI on Proposed Provisions for Draft Data Management and Sharing Policy

Dear Dr. Collins:

Thank you very much for providing the research community with an opportunity to comment on the NIH proposed data management/sharing policy. I am writing on behalf of McLean Hospital, a member of the Partners HealthCare System and a major teaching facility of the Harvard Medical School. McLean maintains the largest program of research in neuroscience and psychiatry of any private psychiatric hospital in the U.S. In FY 18, McLean received approximately \$24 million in NIH/HHS research support. Thus, any proposed change in NIH data management and sharing requirements is of vital interest to us.

Let me begin by stating my colleagues and I conceptually support data sharing as a means of enabling researchers to test the validity of scientific findings, explore new scientific pathways, and shorten the time for ideas to move from the bench to the bedside. Yet, the devil is in the details for data sharing to be successful. The proposed policy is so broad and all-encompassing, we believe if implemented it would be extremely difficult for the NIH to achieve its objective of enhancing science, let alone for Principal Investigators (PI) and institutions to meet their compliance requirements.

Some of our investigators have suggested that the proposed policy appears to be an extension of data sharing requirements for genetic data to scientific data more generally. Genetic data sharing through dbGaP and similar repositories works because genetic data can be supported with standard file formats for data submission. We find it difficult to envision how the many possible experimental designs for laboratory-based experiments would be submitted and archived in a way that could be interpreted by an outside user.

We strongly recommend that the NIH revise the proposed policy to scale back its requirements, add clarity to definitions, and provide meaningful examples for investigators. We also recommend that the NIH consider convening a group of NIH-funded investigators to work with NIH research and administrative leadership to develop a policy that is more realistic and achievable from an investigator's perspective.

Please see below for our comments on specific sections of the proposed policy.

1. Section I

- a. Definitions: The definition of Scientific Data is extremely broad and confusing. We recommend considering the definition of Research Data in OMB Circular A-110 as a substitute. This definition would already be familiar to most of the research community.

- b. Lab notebooks: Throughout the policy there is confusion about lab notebooks and whether they should be shared. Their role/purpose in a “data sharing policy” should be clarified. We maintain that lab notebooks, while critical to the scientific process, are not Scientific Data; they are a means for recording experiments and the Scientific Data generated.
 - c. Reasonable effort to digitize scientific data: While institutionally we are requiring our investigators to transition to digital recordkeeping, we do not recommend including a statement about digitizing scientific data within the current policy. Not all Scientific Data can be digitized; this makes the data no less valuable to research.
2. Section II. Purpose: Making Scientific Data accessible in a “timely manner:”
Researchers generate data daily. We recommend clarifying this section by adding timelines for posting/sharing published and unpublished data. We recommend adding a section to the Progress Report where the PI can inform the NIH of data accessibility. The policy should be flexible. Not all data will be ready for sharing or posting in a repository at the same time. Investigators may want to refrain from posting/sharing unpublished data until it has been published. These situations should be taken into consideration in this section.
3. Section III. Scope and Requirements:
- a. We are concerned that requiring a data management/sharing plan for each application/proposal submission, when the overall funding success rate hovers at 20% or less, creates a significant administrative burden for PIs submitting applications. We recommend the NIH consider requiring the plan as part of the first progress report. These plans will not have the benefit of peer review, but is peer review necessary if the strength or weaknesses of the plan will not be considered in the impact score? Continuation funding for year 2 could be delayed until a plan acceptable to the Program Officer is submitted.
 - b. In the general statement that data management/sharing plans will be required regardless of mechanism, we recommend that the NIH review the different funding mechanisms for appropriateness. For example, a data sharing plan would not be appropriate for a shared instrumentation grant; nor would it be appropriate for a conference grant. We also recommend that the NIH consider eliminating the requirement for institutional training grant applications. We recognize that Scientific Data are generated under training grants, but the management and sharing of the data will vary across the training grant based on requirements of each trainee’s mentor who often come from different departments/research labs with different data management/sharing requirements.
 - c. The policy states, “Reasonable costs associated with data management and sharing could be requested under the budget for the proposed project.” Will supplemental funds be available for these costs? If not, the data management/sharing requirement will only reduce the amount of funding available for the actual research project. We can envision situations where institutions with limited resources will have to provide their investigators with institutional funds or create local data repositories because the NIH funds were

simply not enough to complete the project and pay for the costs associated with external data repositories essentially creating yet another unfunded mandate for grantee institutions.

4. Section IV. Requirements for Data Management and Sharing Plans

- a. General comment: We do not believe PIs will be able to provide all the information the NIH is requiring within a two-page limit.
- b. Scoring/Peer Review Process: If the NIH continues to require plans as part of the grant application/contract proposal, we agree whether a plan is acceptable or unacceptable to reviewers should not be included in the overall impact score.
- c. Plan Elements: We recommend that the NIH create a form with drop down boxes for the PI to identify the plan elements relevant for his/her research. The elements should be minimal and allow for PI flexibility.
- d. Describe type and amount of scientific data to be collected and used in the project: This may be difficult for some types of projects. The example provided is for a specific type of project in which the number of cases/patients/individuals may be known at submission. In many lab-based projects, investigators may improvise and adjust the work making use of techniques that may not have been envisaged initially. We are concerned that PIs may feel providing this type of information will restrict their ability to modify the research as they move forward.
- e. Related Tools, Software and/or Code: Please clarify what the NIH is expecting. For example, would the PI have to justify use of a specific image analysis software product?
- f. 4.1 Indicate where Scientific Data will be archived to ensure long-term preservation: We recommend that the NIH create data repositories to meet this new mandate. As we indicated above, many institutions do not have the resources to develop and maintain repositories for their NIH-funded investigators. Grantee institutions cannot continue to absorb unfunded mandates. Moreover, we are concerned at the possible development of numerous and heterogeneous and possibly rogue repositories.
- g. 4.4 Describe alternative plans for maintaining, preserving and providing access to scientific data should the original plan not be achieved: If the NIH is truly interested in this information, we recommend not requiring submission of a “Plan B” as part of the data management/sharing plan in their application/contract proposal. We recommend adding a section to the data management/sharing reporting section of the annual progress report to describe any changes necessary because the original plan could not be achieved.
- h. 5. Data Preservation and Access Timeline: We question the usefulness of requiring this information in the data management/sharing plans. It may be impossible at the beginning of the project to estimate timelines. This may lead PIs to develop meaningless timelines which become a compliance requirement if the application is funded. We recommend removing this requirement.
- i. 6. Data Sharing Agreements, Licensing and Intellectual Property:
 - i. “NIH encourages terms that provide for the broadest use of data resulting from NIH-funded or -supported research.” Please confirm/clarify that this statement applies to data generated as part of the study, i.e., the data would

not exist if not for the study; and does NOT include any additional, pre-existing clinical data, e.g., annotated, longitudinal data pulled from a patient's medical record.

- ii. 6.1 “Describe any relevant data sharing agreements outlining...how scientific data can and cannot be used.” Please confirm/clarify that this applies only to Scientific Data generated as part of the study. In a situation where the project is supported by NIH and industry or a foundation, the non-NIH sponsors may limit data sharing. Would an SBIR grant be relevant here?
- iii. 6.3 “[I]ndicate how intellectual property...will be managed in a way to maximize sharing of scientific data.” While Scientific Data do not constitute IP, any plan to maximize sharing should not infringe upon the nature of the IP and should preserve ownership rights.

5. Compliance and Enforcement

- a. Community-based Standards: The NIH should specify these standards within the policy or at a minimum provide examples. When we consulted our investigators to develop our response, they were unsure what the standards were and where they might find them.
- b. I/C Monitoring Plans: The policy should include information on how I/Cs will monitor plans, reporting requirements, how to modify plans during the lifetime of the grant. If an I/C determined non-compliance, what would be the enforcement mechanism?
- c. We are very concerned about compliance/enforcement requirements extending beyond the end of the grant's performance period. If this requirement continues in the policy, the NIH should identify the authority that allows the requirement to continue in perpetuity. Comments made during the NIH webinar on the RFI seemed to suggest that the NIH does not consider data sharing requirement as continuing beyond the project end date. The proposed policy contradicts this point and should be clarified. How will the NIH monitor? How will a grantee know if a former award is out of compliance?

Once again, thank you for providing an opportunity for the research community to submit comments. Please do not hesitate to contact me for any additional information.

Yours sincerely,



Kerry Ressler, MD, PhD
Chief Scientific Officer
James and Patricia Poitras Chair in Psychiatry
Chief, Division of Depression & Anxiety Disorders
McLean Hospital
Professor of Psychiatry, Harvard Medical School

Submission #68**Date:** 12/06/2018**Name:** Wade Harper**Name of Organization:** Harvard Medical School**Type of Organization:** University**Role:** Scientific Researcher**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Cell biology

II. The requirements for Data Management and Sharing Plans

(1) Many researchers in Cell Biology generate imaging data sets that are large with respect to both number of files and data size per file. This includes researchers who are working with light or electron microscopy. While some institutions have invested substantially in systems for storing and managing these data, many institutions have not and currently have no available budget for this. Additionally, the pace of data acquisition may make long-term storage so cost-prohibitive as to be unsustainable. Although the draft policy states that “Reasonable costs associated with data management and sharing could be requested under the budget for the proposed project,” without new investment by NIH into total grant dollars available, these costs will quickly erode funds available for doing actual experiments and for the staff/trainees who perform these experiments. Alternatively, these costs could be covered by indirect costs provided by NIH to institutions, but calculations of these rates would need to adjust accordingly. To our knowledge, there are no repositories for large-scale image storage to the community.

(2) Re: data preservation and access

For some sets of data from our department (e.g. mass spec/proteomics), currently available repositories are not under the control of the submitting PI/lab, so security (e.g. encryption), stability and reliability (e.g. backups, day/time stamping), and dissemination capabilities are not guaranteed. For these sites, PIs may not be able to modify the search-ability or discoverability of these databases. PIs also cannot guarantee the maintenance or longevity of these databases. Ideally, NIH would vet these repositories and do periodic reviews to ensure they are continuing to maintain NIH’s guidelines and standards. NIH could provide a recommended list of vetted repositories to researchers. Critically, researchers will also need guidance on using metadata

and indexing standards so that similar types of data from different labs stored in different repositories would be searchable.

Another issue is that there may also be specific classes of data that are largely impossible to share with the community through typical internet-based interfaces. For example, some very large-scale proteomic data sets may reach 4 Tb of data with tens of thousands of individual files. Simply trying to transfer such data through standard interfaces has been known to crash servers. Currently, in many cases, such data sets are provided to third parties that provide a suitably large hard drive to the investigator's lab. For data sets of this size, standard proteomics data repositories are not apparently able to store and disseminate such data sets routinely. It is not clear how such data will fit into the NIH's scheme.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

We would like the process well planned out with sufficient and specific guidelines to avoid ambiguity and unnecessary administrative burden on PIs and labs.

Submission #69**Date:** 12/06/2018**Name:** Andrew Reimer**Name of Organization:** Case Western Reserve University**Type of Organization:** University**Role:** Scientific Researcher**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

clinical

II. The requirements for Data Management and Sharing Plans

I support this effort to move forward with developing the policy and resources to make publicly funded research data available for additional use. One aspect not addressed in the current documentation is related to the secondary use of large amounts of data - think data science approaches. Specifically the use of electronic medical record data, which contains thousands of variables on hundreds of thousands of patients. The data use agreements between healthsystems and research teams are quite restrictive as healthsystems view the data as proprietary as many things can be discerned from analyzing the data. Requirements to share publicly the data that drives the research or store the data beyond the project period could hinder the willingness to participate in future efforts. Therefore, considerations should be given to a priori criteria that differentiate the need for long-term storage, partial or full sharing requirements.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

A phased approach would be necessary. For those not adept at data management and processing (including developing associated metadata) the additional workload will be significant - potentially requiring the addition of a team member to specifically perform these tasks, thus requiring increased budget allocations. Data standards - including minimum necessary data to be shared - should be adopted and provided with resources on how to apply those standards to the specific types of datasets that will be generated. Additionally, infrastructure at the local Institutional and National level should be developed and provided. At

the local level, Institutions receiving funding - supported via the project indirects - should be required to develop the necessary digital infrastructure to support storage of all project related data during the project period and long-term storage thereafter to support those studies that cannot share all or only partial datasets - at no cost to the investigators after the project period. Nationally, data repositories should continue to be developed and available for studies that can share data. These efforts will require significant financial investments to achieve.

Submission #70

Date: 12/06/2018

Name: Kevin McGhee

Name of Organization: New York Genome Center

Type of Organization: Nonprofit Research Organization

Role: Institutional Official

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

genomics

II. The requirements for Data Management and Sharing Plans

We generally agree with the Proposed Provisions for a Draft NIH Data Management and Sharing Policy. The proposed policy reinforces and clarifies existing requirements for management and sharing of data under NIH-funded awards. In particular, we believe that the proposal to consolidate provisions for data management and sharing into a separate two-page Plan, rather than having various elements of this Plan scattered throughout the award proposal, will be beneficial.

NIH should provide clarification or additional guidance regarding the requirement in the Plan to “describe alternative plans for maintaining, preserving, and providing access to alternative data should the original Plan not be achieved.” This requirement is vague and could lead to a wide range of interpretations, creating an unreasonable burden for the awardee at one extreme, and providing no meaningful alternative solution to the failure of the original Plan at the other extreme.

Submission #71**Date:** 12/06/2018**Name:** Maryrose Franko**Name of Organization:** Health Research Alliance**Type of Organization:** Professional Org/Association**Role:** Member of the Public**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

All areas of Biomedical Science

I. The definition of Scientific Data

Scientific Data and Metadata are defined separately in the proposed provisions. The provisions also state that “scientific data MAY include ...metadata.” However, metadata is usually necessary to achieve the goals of data sharing. There should be a stated requirement or at least the explicit expectation that metadata needed to interpret the data be included as a part of sharing scientific data.

The provisions state “NIH expects that reasonable efforts should be made to digitize all scientific data.” This doesn’t go far enough in pushing for digitizing data. Only by digitizing data can it be fully integrated and analyzed among researchers and across a wide variety of platforms. There should be a requirement that the data be digitized and if not, a justification as to why this can’t be done. If the goal is transparency, reproducibility, and reusability of data, the NIH needs to push harder for digitized data.

Additional Comments:

NIH should explicitly state that data that is collected but not used for scholarly publication should also be shared. This data can be combined with other data sets, for instance, or eliminate duplication of unsuccessful lines of inquiry. This explicit wording would be more powerful than passively stating that scientific data is not limited to data used to support scholarly publications.

II. The requirements for Data Management and Sharing Plans

The scope states that the new policy would apply to all intramural and extramural research, funded or supported in whole or in part by grants, ...or other agreements. By stating

“RESEARCH funded” does that exclude individuals supported via an F or T mechanism? If so, this is confusing as the next sentence states “regardless of NIH funding level or mechanism.”

The removal of the minimum dollar amount received by the grantee that triggers the data sharing requirement is a welcome change to the NIH policy.

Section 4.2 under Data Preservation and Access, asks grantees to indicate how data will be made discoverable and “whether a persistent unique identifier or other standard indexing tools will be used.” There should be a REQUIREMENT to use a unique identifier and standard indexing tools. It would be more effective to say “Indicate how the scientific data will be made discoverable. If a persistent identifier and other standard indexing tools will not be used, provide justification as to why not.”

This philosophy should be carried through the whole policy. Require the most effective data sharing practice but incorporate flexibility to request a waiver with justification.

Another example is section 4.1. There should be a requirement that scientific data be stored in an NIH-supported repository or another repository that makes data accessible for reuse, and that meets community-based standards. If an existing repository is not used, (e.g., a newly created repository or the investigator’s organizational repository is used) the investigator must provide evidence this repository meets community-based standards.

With respect to standards, NIH needs to take the lead on moving toward international data standards. These data standards would also help define which repositories are acceptable and set community-based standards for data repositories. Data that conform to international standards is more FAIR. Agreed upon international standards also have the potential to reduce the burden on the investigator – in both curating his/her own data as well as using others’ data.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

There must be an explicit statement that data needs to be shared at the earliest time possible.

Instead of just asking for a time frame from the investigator with no expectations, at the very least there should be a requirement that data that is used to support scholarly publications be shared within a specified (and limited) time frame. Ideally, this would be at the time of publication. However, the NIH could phase this in and require sharing within 12 months initially, then hopefully move to sharing data at the time of publication.

Other data that is collected but not used to support a scholarly publication should ideally be shared 6-12 months after the end of the grant. However, the policy should include a request for

an embargo period. For instance, this could be for 12 months following the end of the award, if the researcher needs this time to publish his/her results or for another justifiable reason.

Submission #72

Date: 12/06/2018

Name: Anonymous

Name of Organization:

Type of Organization: Other

Other Type of Organization: Hospital

Role: Institutional Official

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

Physical Medicine and Rehabilitation

Submission #73**Date:** 12/06/2018**Name:** Sirimon O'Charoen**Name of Organization:** Crohn's & Colitis Foundation**Type of Organization:** Nonprofit Research Organization**Role:** Patient Advocate**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Finding a cure and improve patient life in Inflammatory Bowel Disease (Crohn's Disease and Ulcerative colitis)

I. The definition of Scientific Data

We have sponsored clinical studies with case report forms are used, so we have a particular interest in individual-level subject data like those collected by case report forms. Since completed case report forms are NOT included in the scope of scientific data, the metadata could use more rigorous requirements that at the “minimum” sample characteristics, outcome measures, and other variables used in data analysis according to the study plan and publication need to be included.

II. The requirements for Data Management and Sharing Plans

Comparing “Data Management and Sharing Plan (Plan)” in this proposal with data submission requirements for several major publications: Since majority of researchers will also publish their research, it would be ideal to ensure that the content needed to submit the data sharing plan as a part of NIH application align/cover publication requirements to prevent repetitive effort.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

Data standard: The proposal only mentions encourage the use of common data elements (CDEs). It would be great to broaden the encouragement to include other data standards depending on the type of data and technology platforms such as CDISC for clinical variables and other “minimum metadata reporting standards” (i.e. MIAME for microarray and MINSEQE for sequencing).

Submission #74

Date: 12/07/2018

Name: Finlay Macrae

Name of Organization: University of Melbourne

Type of Organization: University

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

Cancer prevention and early diagnosis. Clinical genomics research. Epidemiology

Attachment:

Request for Information on Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research

Centralization of variant and particularly annotated phenotypic information is challenged by issues relating to privacy and confidentiality. Pedigrees are particularly contentious, as information on family members provided by other family members may be apparently without their consent.

Variant information without annotated phenotype or familial information we consider quite protected, especially if divorced from data relating to the submitter. This could be challenged but we hold this position.

Beyond this, there may be more bioethical challenges.

Consent for submission of information of all sorts to central databases we believe covers issues relating to privacy and confidentiality.

However, there is a wealth of information held in local registries that can be critical to the advancement of science. Take, for example, the need to assemble as much information as possible associated with Variants of Uncertain Significance - in silico, family history, pedigree (for segregation analyses), tumour characteristics (eg MSI status in Lynch Syndrome), functional assays through a Bayesian approach.

The Royal Melbourne Hospital Clinical Ethics Committee pointed to the ethical acceptability of centralizing legacy (unconsented) clinical and other data where the difficulties of gaining such consent are extraordinarily arduous or not possible, and/or may invoke undue harm on individuals, as there is undoubted good for the common wealth with centralization of data. This has been already helpful to a number of ethical opinions inside and outside Australia. InSiGHT also has an ethical position on this on its website.

The phenotypic information potentially available through dbGAP can be very helpful in the task of classifying variants accurately. Guidance to submitters relating the privacy and confidential aspects of information submission to dbGAP would be not only very important to guide dbGAP submitters who have access to rich and important information, but to the broader communities who are faced with the tasks of variant interpretation through InSiGHT, ENIGMA and the ClinGen expert committees.

So, sharing and publishing the deliberations of the dbGAP bioethicists will be helpful, very helpful, to guide groups around the world. Notwithstanding the local authority of ethics committees, the dbGAP bioethical debates on this will inform their own ethics committees.

I attach the conclusions from our own Clinical Ethics Committee which was asked to address the question for InSiGHT.

I hope this guidance will assist potential submitters in their decision to submit to dbGAP and other important central databases - for the important benefit of the common good.

Finlay Macrae AO
Secretary, InSiGHT
Head, Colorectal Medicine and Genetics
The Royal Melbourne Hospital
Victoria, Australia
Ph: +61 3 8559 7232
Fax: +61 3 9348 2004
Email: finlay.macrae@mh.org.au<<mailto:finlay.macrae@mh.org.au>>

Submission #75

Date: 12/04/2018

Name: Bruce R. Thomadsen, PhD, President

Name of Organization: American Association of Physicists in Medicine

Type of Organization: Professional Org/Association

Role: Institutional Official

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

Clinical Practice of Medical Physicists

Attachment:



December 7, 2018

Office of Science Policy
National Institutes of Health
6705 Rockledge Drive, Suite 750
Bethesda, MD 20892

VIA <https://osp.od.nih.gov/provisions-data-managment-sharing/>

RE: Request for Comment: Proposed Provisions for a Draft NIH Data Management and Sharing Policy

Dear Sir or Madam:

The American Association of Physicists in Medicine (AAPM)¹ is pleased to submit comments to the National Institutes of Health (NIH) regarding its “Proposed Provisions for a Draft Data Management and Sharing Policy” that would implement measures to update NIH's 2003 Data Sharing Policy. The AAPM commends the NIH on its work in advancing data sharing to maximize benefits from research efforts funded by the NIH.

General Comments

The AAPM believes this is an important NIH initiative and the draft proposal is a good first step in what likely will be a well-thought-out process addressing the NIH’s movement toward further sharing of grant-generated data. We feel, however, that there are costs and barriers to implementation that will need to be addressed to fairly balance the differing needs and interests of competing stakeholders. We

¹ The American Association of Physicists in Medicine (AAPM) is the premier organization in medical physics, a broadly-based scientific and professional discipline encompassing physics principles and applications in biology and medicine whose mission is to advance the science, education and professional practice of medical physics. Medical physicists contribute to the effectiveness of radiological imaging procedures by assuring radiation safety and helping to develop improved imaging techniques (e.g., mammography, CT, MR, ultrasound). They contribute to development of therapeutic techniques (e.g., prostate implants, stereotactic radiosurgery), collaborate with radiation oncologists to design treatment plans, and monitor equipment and procedures to ensure that cancer patients receive the prescribed dose of radiation to the correct location. Medical physicists are responsible for ensuring that imaging and treatment facilities meet the rules and regulations of the U.S. Nuclear Regulatory Commission (NRC) and various state regulatory agencies. AAPM represents over 8,700 medical physicists.

believe the proposal is too open-ended as to what constitutes data that would need to be shared, and that the parameters that will be used to control access and use of data should be more clearly specified. The AAPM suggests that the NIH guideline should be limited to only data acquired through NIH funding, not all data needed to validate the main findings of the grant.

The AAPM urges the NIH to give careful consideration of “Scope and Requirements” provisions in the proposed policy. We believe uncontrolled access to data--whether patient data or other data--may create a category of researchers without appropriate domain knowledge regarding the data elements in distributed data sets, potentially leading to inappropriate understanding or use of the data and inaccurate research findings. The AAPM questions whether this would be a desirable result. Moreover, the AAPM expresses concern that the guidelines that would regulate distribution of data sets submitted under this proposal could be at odds with institutional guidelines for how individual researchers may share their data with other researchers and with companies.

We also express concern that the proposed guidelines create a significant burden to research institutions by legislating an unfunded mandate to provide the resources and infrastructure needed to properly aggregate, annotate, document, archive, and distribute publicly shareable data sets. Clinicians have to carry out extensive work to ensure accurate aggregation of key data elements. Those efforts currently are largely self-funded. The AAPM urges the NIH to limit its regulatory requirement for dataset sharing to a level consistent with the NIH funding provided to support such data sharing. Further, the NIH should impose data sharing burdens on research institutions only when justified by the potential scientific or clinical impact of the findings associated with the datasets in question, and when paid for by NIH funding. We believe that requiring more extensive and useful data sharing practices can only be achieved by changes in NIH funding models that value data publication infrastructure over the analysis products.

Accordingly, the AAPM recommends further dialogue between the NIH and various stakeholders regarding the implications of the proposal for changing the research landscape. Without more consideration and detail, the proposed policy has the potential to result in outcomes that would be at odds with the positive spirit intended, and that could create an unfunded mandate for researchers that could negatively impact their ability to accomplish the primary objectives of their research.

The AAPM has the following specific comments on questions identified by NIH:

Comments on Questions Posed by NIH

Definition of “Scientific Data”

The proposed provisions define “scientific data” as follows:

“The recorded factual material commonly accepted in the scientific community as necessary to validate and replicate research findings including, but not limited to, data used to support scholarly publications. Scientific data do not include laboratory notebooks, preliminary analyses, completed case report forms, drafts of scientific papers, plans for future research, peer reviews, communications with colleagues, or physical objects, such as laboratory specimens. For the purposes of a possible Policy, scientific data may include certain individual level and summary or aggregate data, as well as metadata. NIH expects that reasonable efforts should be made to digitize all scientific data.”

The AAPM is concerned that the term “scientific data” as defined above is too broad and open ended. While the text lists some exempt classes of information, there are many types of data that satisfy the “necessary to validate and replicate research findings” criterion that are not easily shareable. The AAPM suggests that a narrower, more rigorous set of criteria be provided to define the data that would be expected to be shared. For example, the definition should specifically address the following examples where data sharing seems unfeasible or excessively burdensome:

1. Often hypothesis validation rests upon previously published findings in addition to data collected by the researcher. The researcher has no control of data published by other researchers.
2. Proprietary datasets, e.g., CT sinogram data, to which the researcher has access to only via a non-disclosure agreement, cannot be legally shared and hence should be exempt from data sharing requirements. Manufacturers cannot be required to broadly provide proprietary information.

3. In research studies involving clinical patients, hypothesis validation may rest, directly or indirectly, upon retrospectively aggregated clinical image sets, physician organ segmentations and other image set annotations, clinical treatment plans/dose distributions, patient follow-up data, and/or histopathology slides. Sharing such a complete set of clinically acquired data would be overly burdensome, if not impossible. Further, there is no standardized mechanism for sharing many of these data types. Finally, this could create conflict between NIH-funded researchers, who need clinical data and patient referrals, and clinicians without NIH funding who treat the patients, order the clinical studies, and invest their time in aggregating retrospective clinical data. If the work that these clinicians are performing becomes mandated by the NIH as part of a funded program, it may become necessary for the NIH to fund the clinical effort of many physicians in an academic medical center, which may be cost prohibitive. Not providing funding for “required” effort could negatively impact the willingness of primarily clinical physicians to support or participate in NIH-funded research studies.
4. Medical imaging may be used indirectly to support a study, e.g., to determine study eligibility or to determine tumor response, but may not be part of the data used to directly refute or confirm a hypothesis from an NIH grant. Would institutions be required to share and de-identify such imaging datasets, associated diagnostic interpretations, detailed annotations of pathology type and location? The AAPM believes that NIH should require sharing imaging data sets only if (a) NIH funds support the acquisition, significant processing and/or manipulation of such imaging data and (b) the NIH-supported study directly addresses imaging technology. For example, sharing de-identified image sets used to validate image-registration or reconstruction algorithms developed with NIH support might meet these criteria.
5. Images that are directly involved in the research should be included in scientific data only if relevant to validating NIH-funded study outcomes. In addition, we are concerned about the costs of collection, packaging and documentation of such data and who would be required to pay those costs, especially given the substantial scope of imaging that might be available in the medical record for some study participants. We believe that how the NIH delineates the imaging that may be required to be shared is important.

6. The AAPM believes that it would be overly burdensome to require sharing of de-identified patient-reported outcome, pathology, laboratory, or genetic sequencing data when such data were used only to inform the final patient diagnosis, which was the primary outcome evaluated.
7. The AAPM believes that it would be overly burdensome to require sharing of de-identified patient data that did not constitute the final tested outcome parameters. For example, there is a myriad of recorded data about patients present in their medical records. It would be overly burdensome, if not impossible, for researchers to be required to abstract and share a patient's complete medical record. Further, there are no standardized methods for sharing such data, or for defining what level of detail is required in the abstracting process. Additionally, if certain data elements were not used by the primary researchers in their original study, it would be inappropriate for researchers to be required to abstract that data and share it if requested to do so by an external researcher who believes such data to be critical to a post-hoc analysis of the primary reported data from the study.
8. Medical physicists use test objects to evaluate and calibrate the performance of imaging devices. Biomedical and industry service engineers similarly perform multiple measurements on imaging devices. The AAPM believes that it would be overly burdensome to provide all of the test measurements used to assess equipment performance or benchmark software accuracy, especially if such data did not constitute the primary tested outcome parameters. Further, a standardized method for sharing such data does not exist, nor is it clear what data would be required to be included, some of which might be manufacturer-specific or proprietary.

The AAPM recommends that the requirement for data sharing be limited to datasets that satisfy the following criteria:

1. Data whose acquisition and processing was collected and/or processed by individuals supported by the NIH grant in question.
2. Data used directly in the formulation or validation of testable hypotheses or developmental goals supported by an NIH grant.
3. In cases where data sharing would commit a research institution to substantial costs, the beneficial scientific or clinical impact of sharing such datasets must justify the costs, and NIH funding must be provided to support the work required to prepare, archive and distribute such data.

4. Sharing the data is consistent with protection of research subject confidentially, intellectual property and/or non-disclosure agreements governing the investigators' use of these data.

The AAPM further recommends that the phrase "certain individual level and summary or aggregate data" be clarified. Principal researchers need to know:

- What level of detail is sufficient to meet these requirements?
- What level of aggregation is acceptable? Without clarification, for example, population statistics could be provided in lieu of per-patient data.

Scope and Requirements

The AAPM believes that clarification is needed to determine whether the phrase, "that results in scientific data," includes the pre-existing scientific data being mined by a specific study, or only the resultant generated data.

The AAPM expresses concern that the open-endedness of the NIH definitions makes it difficult to interpret and guide researchers to the actual requirements. This may also pose issues when other researchers or the general public request such data.

Data Management and Sharing Plans

The AAPM identifies some specific issues with the proposed requirements below, but first, the AAPM expresses its concern about the complexity of these plan requirements and the ability of researchers to successfully comply with these requirements.

Accordingly, the AAPM recommends that NIH consider devising a form to guide the researcher through the process and ensure that the researcher includes sufficient information. The form could include checkboxes or selection tools for major classifiers such as: Imaging, Modality, Approximate Number of Subjects, and Sequences.

We believe that the data sharing policy should address liability concerns. Any data shared due to the NIH policy could potentially raise liability concerns, and this issue has the potential to totally derail the free sharing of data if it is not addressed well.

In the special case of patient data, de-identification of the data to be shared is a critical aspect of any data sharing policy. Given the potential for problems in this

area, and the importance of avoiding conflicts with laws governing protected health information, the AAPM believes that the NIH must develop fail-safe tools to fully de-identify data before they are shared, and that any failure of such tools be the sole responsibility of the NIH.

The AAPM sees the following additional issues regarding de-identification:

1. Does the NIH intend to provide these tools or services? We believe that fail-safe de-identification is a critical barrier to data sharing and that it is incumbent on the NIH to provide such tools or services if it is to require the sharing of patient data.
2. We believe NIH should specify requirements for patient/subject consent for sharing their de-identified data with the scientific community, the general public, and industry. For example, if a class of retrospective research studies is approved by the local institutional review board (IRB) without the need for patient consent, then the research can be conducted with the waiver of consent. However, if the NIH requires sharing of such retrospective data, and if patient consent for the sharing of that (de-identified) data is required by the NIH, then there is a potentially insurmountable barrier to conducting that research because of the impossibility of obtaining patient consent for previously acquired data. As an example of an ethically questionable outcome, suppose a company downloading the data creates a highly profitable product out of specific patients' data. Excluding the patients from sharing in this financial success may not be ethical.
3. Use of data from one set of patients to support discovery of new knowledge to improve care for other patients is widely accepted by patients as a good reason for their medical sharing data. Not all patients, however, are open to having their data shared, particularly outside of the healthcare institution where it was acquired or with industry, which may profit from use of patient data. This is particularly true for rare diseases where the population is small.

The AAPM asserts that the requirement for patient consent is critical when it is needed for ethical conduct of research, for example in a prospective trial. However, requiring patient consent when it is not deemed necessary for ethical conduct of research, such as for a retrospective data review study, but is only required for NIH data sharing requirement purposes, is overly burdensome and may severely limit the ability to answer important medical and scientific questions. Accordingly, the

AAPM urges the NIH to carefully consider the circumstances that would require patient informed consent. We ask the NIH to articulate in this data sharing proposal how this issue will be handled.

The AAPM supports the proposed requirement for Plan Review and Evaluation for extramural grants that specifies that the data management and sharing plans could be evaluated by reviewers but not factored into the impact score through peer review. We urge clarification, however, as to what training the reviewers would receive related to commenting on the acceptability of the plan, what criteria would be used to evaluate the plan, and whether an unfavorable evaluation could result in withholding or delaying funding.

The AAPM urges the NIH to consider the burdens imposed on researchers by implementation of an expanded data sharing requirement. Specifically, we ask the NIH to consider cost effectiveness and balance the value to the community and the cost to the researcher as a criterion for assessing data sharing plans. We also recommend that NIH allocate funds that researchers can apply for to support costs of data gathering, digitalization, annotation, de-identification, validation, archival, and distribution, and that NIH consider not requiring extensive data sharing plan efforts if such funding is declined. In addition, the AAPM recommends including an embargo or delay on sharing data that enables researchers to publish their findings before handing data off to competitors.

Data Preservation and Access

The AAPM believes that requirements for “Data Preservation and Access” raise some questions that must be addressed. We identify the following specific questions:

- Section 4.1: What is long-term? Is it straightforward to add data to an existing repository?
- Section 4.2: What is “made discoverable?” How will others be made aware that it is available for use?
- Section 5.0: Will there be criteria for when data are required to be shared? Data should be shared only after the study is completed.

Data Preservation and Scientific Data Archiving

The AAPM has concerns regarding data preservation and archiving. We believe that the NIH requirements should clarify how long data are required to be preserved

and who has the responsibility to maintain the data. As an example, if data are generated by a subcontract, is the subcontract principal investigator or the overall grant principal investigator responsible for managing the data preservation, archival, and sharing? Who would answer questions regarding the data? Further, the efforts required for these activities could far exceed the time period for which the grant is funded. Researchers should not be mandated to provide services for which no NIH funding is made available. Moreover, researchers can change universities and contact information, and that could complicate these processes.

Optimal Timing, Phased Adoption

The AAPM urges flexibility in phased adoption for implementation. We believe the phase-in period would require the preparation and dissemination of educational materials, tools to de-identify data containing protected health information, standardized patient information and consent forms, tools to guide researchers on how to complete relevant forms, and tools to ensure that all required information is provided in a data management and sharing plan.

In summary, while the AAPM supports NIH's efforts to update its Data Management and Sharing Policy, the AAPM urges NIH to implement a data sharing policy that eases the burden imposed upon investigators, helps investigators in navigating the process, and assists in funding infrastructure requirements currently borne by institutions. The AAPM hopes that the NIH will carefully consider AAPM's comments and adopt the AAPM's recommendations when crafting the final policy.

Thank you for the opportunity to comment. If you have any questions or require additional information, please contact Richard J. Martin, JD, Government Relations Project Manager, at 571-298-1227 or Richard@aapm.org

Sincerely,



Bruce R. Thomadsen, PhD, FAAPM, FABS
President, AAPM

Submission #76**Date:** 12/07/2018**Name:** Allen A. DiPalma**Name of Organization:** University of Pittsburgh**Type of Organization:** University**Role:** Institutional Official**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Clinical, Genomics, Neuroscience, Epidemiology, Bioengineering, Artificial Intelligence, Big Data, Nano Technology, Vaccine Research, Biotechnology

I. The definition of Scientific Data

At a minimum, data supporting graphics and statistics presented in scientific articles with PMcIDs should be shared. Data could be locked at the time of publication (e.g., patients would not be required to be followed for survival after the publication) and made available when the article is published on PMC.

II. The requirements for Data Management and Sharing Plans

Software necessary to export or view the data must be described (e.g., releases of statistical or database packages) and conceivably be containerized with data.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

If resources are allocated in grant awards for sharing data, then data should be made available at the end of the grant period or at the time of publication of data products, whichever comes last (some data are not publishable until after the grant support has ended). Phased adoption could be modeled after the HHS approach described in clinicaltrials.gov.

Regarding infrastructure, NIH should implement a policy that is considerate of existing institutional repositories funded by indirects, including servers (probably cloud-based), appropriate software and consulting experts (e.g., librarians). Investigators should not have to

plan for new or additional infrastructure, nor be required to establish project specific data sharing on a proposal by proposal basis.

Attachment:

University of Pittsburgh Feedback on “Proposed Provisions for a Draft NIH Data Management and Sharing Policy”

Contributions (below) were submitted by members of the University of Pittsburgh’s Data Commons Work Group and School of Computing and Information. Special thanks to Michael Becich, Daniel Normolle, Melissa Ratajeski, Michael Madison and Thomas Hitter for their thoughtful and detailed feedback.

- **NIH definition of scientific data to be covered within these plans:**

At a minimum, data supporting graphics and statistics presented in scientific articles with PMCIDs should be shared. Data could be locked at the time of publication (e.g., patients would not be required to be followed for survival after the publication) and made available when the article is published on PMC.

- **NIH elements of required data management and sharing plans:**

Software necessary to export or view the data must be described (e.g., releases of statistical or database packages) and conceivably be containerized with data.

- **Optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy, as well as how possible phasing could relate to needed improvements in data infrastructure, resources, and standards:**

If resources are allocated in grant awards for sharing data, then data should be made available at the end of the grant period or at the time of publication of data products, whichever comes last (some data are not publishable until after the grant support has ended). Phased adoption could be modeled after the HHS approach described in clinicaltrials.gov.

Regarding infrastructure, NIH should implement a policy that is considerate of existing institutional repositories funded by indirects, including servers (probably cloud-based), appropriate software and consulting experts (e.g., librarians). Investigators should not have to plan for new or additional infrastructure, nor be required to establish project specific data sharing on a proposal by proposal basis.

Specific University of Pittsburgh Comments on draft NIH policy:

Page 1: *Scientific data do not include laboratory notebooks, preliminary analyses, completed case report forms, drafts of scientific papers, plans for future research, peer reviews, communications with colleagues, or physical objects, such as laboratory specimens.* That preliminary analyses and lab notebooks are not considered “Scientific Data” for the purposes of data sharing is an important distinction to make.

Page 1: There is no explicit mention of software code or for either analysis or data management, nor of the data base platform, in the description of the scientific data, although these components might be implied under “Metadata.”

Page 1: Metadata Definition. There are many categories of metadata: descriptive, preservation, technical, structural, admin (<https://confluence.cornell.edu/display/culpublic/Metadata+Design+and+Best+Practices>). Clarification of what is meant by “Metadata” may be needed: For example some may argue that lab notebooks detailing methodology of data collection or scripts written to clean data would be “metadata” required to make the data useable.

Page 2: It is not clear if software produced as a product of NIH funded research as an end product (i.e., not just the software used to produce a specific scientific result) would be considered “Scientific Data.” Generally, if such software is an explicit work product of an NIH grant, then there must be a plan for its dissemination, generally under “Resource Sharing.”

Page 2: The two bullet points under III appear to be redundant. What does the second point cover that the first point does not?

Page 3: Under IV, it is reasonable that, for extramural grants, a data management plan should be an Additional Review Consideration to give NIH program staff flexibility, but there should be a mechanism of appeal for awardees in case they find the program’s requirements onerous.

Pages 3-4: Under IV, Plan Elements, Data Type, a description of the format of the data should be included, e.g., ASCII tabular file, spreadsheet, PNG images, VC file, etc.

Pages 3-4: 1. Data Types. The term “amount” may need to be more specific. Some people may equate this with the number of files, but when thinking of budget and/or repository selection individual file size and total dataset size is important. Repositories such as Zenado and FigShare have limits on size of file upload and total package (extra space available for fee).

Page 4: Under IV, Related Tools, should source code be required. Should data and code be containerized (e.g., Docker)?

Page 4: Item 4.1. The University of Pittsburgh Health Sciences Library System, as well as many other health sciences libraries, curate Data Catalogs (several through funding from the National Network of Libraries of Medicine: <https://www.datacatalogcollaborationproject.org/partners/>). Rather than functioning as data repositories to store data, the catalogs are digital way-finders which includes rich metadata (including: description, keywords, format of dataset, instrumentation or software utilized/required, and information about who can access each dataset and how) to increase findability and usefulness of datasets. If data was not deposited into a repository for whatever reason a metadata record of the datasets should be made available for discoverability.

Page 4: Item 4.2. As pointed to in NLM Strategic Plan:

https://www.nlm.nih.gov/pubs/plan/strategic_planning.html linkages between publications and datasets must occur.

Page 5: Section 6 of the NIH draft speaks to proprietary interests in the context of data management (generally). It's important that NIH policy keeps separate the ideas of (i) contractual arrangements that may affect access to and use of scientific data, (ii) proprietary rights asserted with respect to the data themselves, as forms of copyright or patent or by analogy to copyright or patent, and (iii) proprietary rights asserted with respect to research outputs that are related to but distinct from data (journal articles and other published scholarship, which are subject to copyright law, publishing restrictions, and (often) open access deposit and distribution requirements; and inventions, which may be subject to disclosure and assignment requirements of Bayh-Dole or private or philanthropic research sponsors).

- (i) Contracts – are not uncommon. They should always be scrutinized for overbreadth, and for consistency with underlying goals of data sharing policies, but generally, they are regarded as legally valid.
- (ii) Proprietary rights in data – are extremely rare, especially in scientific and research domains. The current draft does a nice and appropriate job of ***not*** prioritizing policy development based on “ownership” of data.
- (iii) Proprietary rights adjacent to data – are common, so data sharing mandates must be crafted and applied in ways that do not conflict with laws or policies that bear on publication and distribution of scholarship, or with laws that permit research universities to engage in important technology transfer activities.

Current Section 6 could be strengthened to make it clearer that the public interest in broad access to scientific data should not be merely balanced against potential proprietary claims; these are not equivalent values, or equivalent goals. Scientific data are the lifeblood of the modern university and the foundation of the university's contributions to the public good. Accommodation of proprietary interests should be made only where it is clearly necessary because it is required by law, or because accommodating those interests is otherwise clearly consistent with the public interest.

Page 6: Scientific Data Archiving. Regarding the statement: *“Investigators would be encouraged to consider using repositories that make scientific data available at no cost for extended periods of use”*. This statement should be removed or reworked. Cost should not be the main factor. Researchers should first be encouraged to deposit into an established disciplinary repository to increase findability/impact. Also repositories should meet core requirements (<https://www.datasealofapproval.org/en/information/requirements/>) such as the: “repository guarantees the integrity and authenticity of the data” and “has a continuity plan to ensure ongoing access to and preservation of its holdings”.

Page 6: Compliance and Enforcement. Regarding the statement: *“Many repositories offer data preservation and access for free.”* Many repositories have data size limits on what is free. These considerations need to be factored into the proposal and are important for budgeting, especially for institutions that do not have a data repository infrastructure in place.

General University of Pittsburgh Comments on draft NIH policy:

1. NIH should implement an appropriate level of budget support (i.e., advice on budgeting) for grant applicants to support data management and data sharing. Appropriate amounts of direct project costs should be allowable to support data management and sharing.
2. NIH policy should be sufficiently flexible to include or be considerate of institution-level policies. Policies that promote the use of professional resources are likely to be most effective.
3. Regulations in place are likely sufficient to protect protected health information (PHI); no additional restrictions should be required at this time.
4. Standardization will be extremely costly and complex. Rather, publication of metadata should provide context for data, and metadata should comprise the scientific publication and its supplements (potentially, detailed laboratory protocols). Standardizing data products to the extent that current Big Data methods can meaningfully amalgamate the products of hundreds or thousands of independent research projects could effectively make the entire scientific community data slaves and impede innovation.
5. US Department of Health and Human Services has determined how to enforce the clinicaltrials.gov standards, NIH can certainly rely on their expertise and influence to enforce data sharing.

Submission #77**Date:** 12/07/2018**Name:** Andre Noel Porter**Name of Organization:** American Society of Biochemistry and Molecular Biology**Type of Organization:** Professional Org/Association**Role:** Member of the Public**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Basic biomedical research

I. The definition of Scientific Data

With respect to sharing scientific research products with maximal likelihood for reuse, and with an eye toward maximizing the effectiveness of research funds from NIH, scientific data should be defined as refined observations and analyses utilized to support published conclusions. Data that should be required for sharing will vary from field to field, however the fields of genomics and structural biology provide suitable examples for paths forward. Within each field, the data shared is not true raw data, but rather processed data that provides a clear launch point for specific analyses. For example, within structural biology, sharing of X-ray crystallographic coordinates by deposition in the Protein Data Bank (PDB) along with a solved structure are nearly universally required upon publication. The sharing of structure factors allows for similar rapid analyses by other investigators without requiring the sharing or dissemination of the large, gigabyte-to-terabyte size diffraction image datasets. We advocate similar approaches for other fields within the biomedical research enterprise and encourage them to develop consensus approaches for determining what data formats, if shared widely, would lead to the best advances.

Additionally, while we support the motives of requiring data to be Findable, Accessible, Interoperable, and Reusable (FAIR), we are concerned with potential burdens that would be placed on researchers by overly restrictive policies. We are also concerned with the ability for the biomedical research enterprise to produce a single data sharing standard that will be appropriate for most data types. Data sharing should abide by or attempt to abide by as many FAIR principles as is feasible, though we do not think it is appropriate to require all data to be shared in the same format. Discipline-specific data sharing approaches, for example the PDB and CIF formats used for structural biology provide data in a way that is findable, accessible,

and reusable. We recommend that in addition to requiring data to be deposited in accordance with FAIR standards, NIH should lead the way in organizing taskforces to determine data sharing approaches for disciplines that currently lack standardized approaches. As part of these efforts, we encourage NIH to continue support of publicly accessible data repositories, similar to the support provided for the PDB.

A central question about the data sharing policy is when data will be shared. The draft policy describes data that should be shared as "including, but not limited to, data used to support scholarly publications". Although prepublication sharing is common, and we agree essential, for many large consortium projects, we strongly feel that the default trigger for data sharing for most investigator-initiated funding mechanisms should be publication. Our reasoning for this conclusion is three-fold. 1) Publication is an unambiguous moment in the life cycle of data; there can be no ambiguity as to whether the research findings supported by particular data is published or not. Therefore, publication is a robust and easily adjudicated trigger for sharing. 2) Publication, per se, demonstrates that both the authors and the reviews consider the data complete and reliable. 3) The publication threshold for data sharing provides PIs with control of when their data will be shared. Therefore, we strongly encourage you to set publication as the default trigger for data sharing, which could be modified in the Data Management and Sharing Plan, as appropriate.

II. The requirements for Data Management and Sharing Plans

Data management plans should attest that data acquired with the support of NIH, in part or in whole, should be freely shared with the public using a Creative Commons Attribution-ShareAlike (CC-BY-SA 4.0) license. The CC-BY-SA 4.0 license best meets the spirit of public funding for biomedical research, allows the least restricted reuse of data, provides for appropriate attributions, and encourages innovation that builds upon the widest possible base. PIs should abide by appropriate conventions within their field with respect to data sharing, and data management plans should include commentary on best practices in the field and identify any anticipated deviations from these best practices.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

To move the scientific community toward FAIR data sharing, we suggest that NIH identify data repository partners built around FAIR standards. For all fields, data sharing requirements should be phased in over a four-year period to provide a reasonable period of time for investigators to determine appropriate data sharing approaches. For fields without a consensus approach, general requirements should be developed, and the four-year phase-in window should be delayed until such a time that task forces outline accepted data sharing policies. For fields

without a consensus on general requirements, data management plans should attest to and describe the widest possible realistic approach to data sharing.

As the biomedical research enterprise moves towards wider sharing of data, the evermore decentralized mode of data sharing is an issue with respect to whether data is findable. We recommend encouraging or requiring PIs to report the digital object identifiers (DOIs) for shared data directly in publications. These DOIs are commonly available from publicly accessible repositories and provide a direct link to the shared data. The use of DOIs shared within publications allows researchers to deposit data in any of the publicly accessible data repositories while providing a facile access route for consumers of the data. This route takes advantage of the community's existing paradigm for reporting and disseminating research products as DOIs are commonly used as a short and rapid format for linking to publications. An alternative approach could be for NIH to require principal investigators (PIs) to report shared data and DOIs as a part of annual progress reports. However, communicating this shared data to the public would then require substantial efforts by NIH to create an internet portal for access by the public. This NIH-hosted option is less desirable as it (1) increases workload for NIH, (2) requires expenditure of NIH funds to create and maintain a shared data internet portal, and (3) would require data consumers to search yet another entirely separate domain for shared data. Furthermore, we implore NIH to consider approaches that will limit the administrative burdens that data sharing may place on investigators. Allowing investigators to share data in publicly accessible repositories in formats determined by each individual field will prevent investigators from having to convert their data into formats that may not be consistent with those required for publication.

Attachment:

The American Society for Biochemistry and Molecular Biology's response to NOT-OD-19-014 "Request for Information (RFI) on Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research"

Comments submitted electronically on December 7, 2018

I. The definition of Scientific Data

With respect to sharing scientific research products with maximal likelihood for reuse, and with an eye toward maximizing the effectiveness of research funds from NIH, scientific data should be defined as refined observations and analyses utilized to support published conclusions. Data that should be required for sharing will vary from field to field, however the fields of genomics and structural biology provide suitable examples for paths forward. Within each field, the data shared is not true raw data, but rather processed data that provides a clear launch point for specific analyses. For example, within structural biology, sharing of X-ray crystallographic coordinates by deposition in the Protein Data Bank (PDB) along with a solved structure are nearly universally required upon publication. The sharing of structure factors allows for similar rapid analyses by other investigators without requiring the sharing or dissemination of the large, gigabyte-to-terabyte size diffraction image datasets. We advocate similar approaches for other fields within the biomedical research enterprise and encourage them to develop consensus approaches for determining what data formats, if shared widely, would lead to the best advances.

Additionally, while we support the motives of requiring data to be Findable, Accessible, Interoperable, and Reusable (FAIR), we are concerned with potential burdens that would be placed on researchers by overly restrictive policies. We are also concerned with the ability for the biomedical research enterprise to produce a single data sharing standard that will be appropriate for most data types. Data sharing should abide by or attempt to abide by as many FAIR principles as is feasible, though we do not think it is appropriate to require all data to be shared in the same format. Discipline-specific data sharing approaches, for example the PDB and CIF formats used for structural biology provide data in a way that is findable, accessible, and reusable. We recommend that in addition to requiring data to be deposited in accordance with FAIR standards, NIH should lead the way in organizing taskforces to determine data sharing approaches for disciplines that currently lack standardized approaches. As part of these efforts, we encourage NIH to continue support of publicly accessible data repositories, similar to the support provided for the PDB.

A central question about the data sharing policy is when data will be shared. The draft policy describes data that should be shared as "including, but not limited to, data used to support scholarly publications". Although prepublication sharing is common, and we agree essential, for many large consortium projects, we strongly feel that the default trigger for data sharing for most investigator-initiated funding mechanisms should be publication. Our reasoning for this conclusion is three-fold. 1) Publication is an unambiguous moment in the life cycle of data; there can be no ambiguity as to whether the research

findings supported by particular data is published or not. Therefore, publication is a robust and easily adjudicated trigger for sharing. 2) Publication, *per se*, demonstrates that both the authors and the reviews consider the data complete and reliable. 3) The publication threshold for data sharing provides PIs with control of when their data will be shared. Therefore, we strongly encourage you to set publication as the default trigger for data sharing, which could be modified in the Data Management and Sharing Plan, as appropriate.

II. The requirements for Data Management and Sharing Plans

Data management plans should attest that data acquired with the support of NIH, in part or in whole, should be freely shared with the public using a Creative Commons Attribution-ShareAlike ([CC-BY-SA 4.0](https://creativecommons.org/licenses/by-sa/4.0/)) license. The CC-BY-SA 4.0 license best meets the spirit of public funding for biomedical research, allows the least restricted reuse of data, provides for appropriate attributions, and encourages innovation that builds upon the widest possible base. PIs should abide by appropriate conventions within their field with respect to data sharing, and data management plans should include commentary on best practices in the field and identify any anticipated deviations from these best practices.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards.

To move the scientific community toward FAIR data sharing, we suggest that NIH identify data repository partners built around FAIR standards. For all fields, data sharing requirements should be phased in over a four-year period to provide a reasonable period of time for investigators to determine appropriate data sharing approaches. For fields without a consensus approach, general requirements should be developed, and the four-year phase-in window should be delayed until such a time that task forces outline accepted data sharing policies. For fields without a consensus on general requirements, data management plans should attest to and describe the widest possible realistic approach to data sharing.

As the biomedical research enterprise moves towards wider sharing of data, the evermore decentralized mode of data sharing is an issue with respect to whether data is findable. We recommend encouraging or requiring PIs to report the digital object identifiers (DOIs) for shared data directly in publications. These DOIs are commonly available from publicly accessible repositories and provide a direct link to the shared data. The use of DOIs shared within publications allows researchers to deposit data in any of the publicly accessible data repositories while providing a facile access route for consumers of the data. This route takes advantage of the community's existing paradigm for reporting and disseminating research products as DOIs are commonly used as a short and rapid format for linking to publications. An alternative approach could be for NIH to require principal investigators (PIs) to report shared data and DOIs as a part of annual progress reports. However, communicating this shared data to the public would then require substantial efforts by NIH to create an internet portal for access by the public. This NIH-hosted option is less desirable as it (1) increases workload for NIH, (2) requires expenditure of NIH funds to create and maintain a shared data internet portal, and (3) would require data consumers to search yet another entirely separate domain for shared data. Furthermore, we implore NIH to consider

approaches that will limit the administrative burdens that data sharing may place on investigators. Allowing investigators to share data in publicly accessible repositories in formats determined by each individual field will prevent investigators from having to convert their data into formats that may not be consistent with those required for publication.

The ASBMB appreciates the opportunity to weigh in on the National Institutes of Health's Data Management policy and we welcome any further discussions on this important topic.

André Porter
Science Policy Analyst
American Society for Biochemistry and Molecular Biology
11200 Rockville Pike, Suite 302
Rockville, MD 20852-3110
Office: 240-283-6621
Email: aporter@asbmb.org

Submission #78**Date:** 12/07/2018**Name:** Meir Stampfer**Name of Organization:** Brigham and Women's Hospital**Type of Organization:** Healthcare Delivery Organization**Role:** Scientific Researcher**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Epidemiology, genomics, biomarkers

II. The requirements for Data Management and Sharing Plans

The first suggestion is to evaluate the current data sharing systems to identify whether they are functioning in a way that allows the desired access to external investigators and whether there are problems that should be fixed. I.e., see what is broken before applying fixes. As PI of the Nurses Health Study, I have overseen over 300 data sharing requests over the past four years; virtually all are approved if we have the data.

My colleagues and I strongly support the principles of data sharing and the value of maximizing the knowledge that can be gained from existing data.

Posting epidemiologic data on widely available data bases may lead to bad science, wrong results, misuse of data. Epidemiologic data variables are more complicated than genetic variables. I have had personal experience where mistakes were made due to lack of familiarity with the nature of the data. Some people believe science is self-correcting (I do not think this happens automatically, especially for wrong null results) but even when it works, it is often at great cost of energy and trust.

There is a difference between new cohort studies that will obtain written informed consent for data to be widely shared on publicly available databases and existing cohorts that began long ago without this kind of consent. Requiring existing cohort studies to re-consent participants to allow wider sharing will likely result in only a fraction of the cohort giving written consent, while some participants may withdraw altogether out of concern that their privacy will be jeopardized. To preserve continued active follow-up in the existing cohorts, it may be best to permit the cohorts to continue sharing through current methods, which work remarkably well.

Epidemiologic studies often rely in part on data from Medicare and from state cancer registries. These organizations do not allow the study to share the data with others except under very specified conditions. These issues should be solved before mandating data sharing through a central database.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

See above - I suggest a delay until there is a review of the extent of the "problem" and resolution of the issues around confidentiality.

Submission #79**Date:** 12/07/2018**Name:** Holly J Falk-Krzesinski, PhD**Name of Organization:** Elsevier**Type of Organization:** Other**Other Type of Organization:** Research Information Analytics; Publisher**Role:** Institutional Official**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

All types of research data

I. The definition of Scientific Data

The proposed definition is consistent with those from the OSTP Public Access Memo and OMB circular A-110 and it is sufficiently flexible to allow for discipline-specific data standards setting. Having the benefit of being able to build on those earlier definitions, we propose a slightly amended definition for greater clarity to the research community:

“Research Data: The recorded factual material commonly accepted in the scientific community as necessary to validate and replicate research findings, including, but not limited to, the underlying primary data that support the central findings of a scholarly publication. Research data do not include laboratory notebooks, preliminary analyses, completed case report forms, drafts of scientific papers, plans for future research, peer reviews, communications with colleagues, or physical objects, such as laboratory specimens. For the purposes of a possible Policy, research data may include certain individual level and summary or aggregate data, as well as metadata. NIH expects that reasonable efforts should be made to digitize all research data.”

The proposed revised definition above replaces “scientific data” with “research data” throughout. While “scientific data” and “research data” are often used interchangeably, “scientific data” may be perceived as exclusive of the social and behavioral sciences and medical humanities domains, whereas “research data” is a more inclusive term encompassing all disciplines. The addition of “underlying primary data” to the definition provides further clarity that research data are distinct from the text in a manuscript or a final published article and affiliated supplementary materials published as part of a journal article, a distinction

between research data and interpretations or presentations of research data. Once this policy is updated, we encourage consistency of definitions in other related data sharing policies across the NIH.

Within Provision I, we recommend adding a definition for “Research Data Repository” (a digital platform where research data is stored for the purposes of publishing, sharing, re-use, linking, and preservation). Along with the addition of a Research Data Repository definition, we also recommend that the policy include basic guidance to researchers on criteria that constitute a trustworthy repository. Some very useful resources about trustworthy repositories include:

- CoreTrustSeal (<https://www.coretrustseal.org/>): Offers certification based on the Core Trustworthy Data Repositories Requirements catalogue and procedures. This universal catalogue of requirements reflects the core characteristics of trustworthy data repositories and is the culmination of a cooperative effort under the umbrella of the Research Data Alliance (RDA).
- Repository Finder (<https://repositoryfinder.datacite.org/about>; <https://www.re3data.org/>): A project of the Enabling FAIR Data Project in partnership with DataCite that queries the re3data registry of research data repositories.
- Scientific Data Recommended Data Repositories (<https://www.nature.com/sdata/policies/repositories#general>): The journal Scientific Data has compiled a comprehensive list of trusted discipline-specific, community-recognized, and generalist research data repositories.
- Recommended versus Certified Repositories: (<http://doi.org/10.5334/dsj-2017-042>) A research article that examines both recommended and certified repository characteristics. Husen, S.E., de Wilde, Z.G., de Waard, A. and Cousijn, H., 2017. Recommended versus Certified Repositories: Mind the Gap. *Data Science Journal*, 16, p.42.

Elsevier’s data repository, Mendeley Data, which is free to researchers globally, is recognized as a trusted research data repository in all of the directories above and has received the CoreTrustSeal.

II. The requirements for Data Management and Sharing Plans

The list of elements of a data management plan (Plan) is quite comprehensive. OSP may also wish to review the Elements of a Data Management Plan at North Carolina State University, an abbreviated compilation of data management plan elements from several sources. There are some additional elements in that compilation that could be considered for inclusion by OSP for its revised policy: Roles and Responsibilities; Data Formats and Metadata; Privacy; and Costs.

Given that some projects are data-intensive and some are complex research programs and/or multi-institutional proposals, the two-page limit may be too constraining and not allow for researchers to provide sufficient detail necessary for review by peer reviewers and program staff.

OSP might consider working with the California Digital Library to add an NIH template to the DMPTool, which is used by many research universities and institutions to prepare quality data management plans.

Beyond human-readable Plans, a revised policy should also address the use of machine-readable data management plans (DMPs; a.k.a., machine-actionable DMPs and active DMPs) in the not-too-distant future. Machine-readable DMPs focus on assigning identifiers and machine-actionable components of a plan. It is premature to require researchers to develop machine-readable DMPs at this time, but the revised policy could encourage researchers to develop them when possible. For additional information on machine-readable tools and standards, we recommend the following resources, as well as others available from the Research Data Alliance (RDA) web site:

- Miksa, T., Rauber, A., Ganguly, R., & Budroni, P. (2017). Information Integration for Machine Actionable Data Management Plans. *International Journal of Digital Curation*, 12(1), 22. <https://doi.org/10.2218/ijdc.v12i1.529>
- Miksa, T., Simms, S., Mietchen, D., & Jones, S. (2018). Ten simple rules for machine-actionable data management plans (preprint). <https://doi.org/10.5281/ZENODO.1172673>
- Research Data Alliance (RDA). (2017). DMP Common Standards WG | RDA. Retrieved from <https://www.rd-alliance.org/groups/dmp-common-standards-wg>

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

We posit that a phased implementation isn't as critical as a commitment by OSP to review and revise the policy on a more regular cycle, perhaps every 2-3 years. A shorter review/revise cycle will allow OSP to be nimble and keep the policy up to date with advances in both technical and technology capabilities, such as machine-readable DMPs. A shorter cycle will also allow for timely revisions should unforeseen negative consequences result, or if previously unconsidered limitations are brought to light. Moreover, since the NLM has recently commissioned the National Academies of Sciences, Engineering, and Medicine (NASEM) to conduct a study on forecasting the long-term costs for preserving, archiving, and promoting access to biomedical data, it will be important to review this policy in consideration of the findings from that study once complete in mid-2020.

In addition, we strongly encourage OSP to set a schedule for collecting data about research data sharing practices, evaluating the impact of sharing research data on both research and researchers, and work with RDA and other community partners to develop and establish research data sharing metrics—sharing the findings with the community. These efforts underpin an evidence-based approach to science policy consistent with the science of science policy and will provide data to inform future policy changes and revisions.

Attachment:

Name:

Holly J. Falk-Krzesinski, PhD

Name of Organization:

Elsevier

Type of Organization:

Other

Other Type of Organization:

Research Information Analytics; Publisher

Role:

Vice President, Research Intelligence

Research Area Most Important to You or Your Organization (e.g., clinical, genomics, neuroscience, infectious disease, epidemiology)

All types of research data

Response Introduction

Elsevier is a global information analytics business that helps institutions and professionals advance healthcare, open science, and improve performance for the benefit of humanity. To unlock the full potential of research data, Elsevier offers a Research Data Management portfolio that supports researchers by integrating workflow tools throughout the data lifecycle and supports institutional stakeholders by integrating these tools with institutional workflow solutions. As both a research information analytics provider and a publisher, Elsevier embeds research data in the workflow and makes it Findable, Accessible, Interoperable, and Reusable. Elsevier also applies metrics that enable both researchers and institutions to gauge progress toward compliance and performance goals. We appreciate the opportunity to share feedback during the process of OSP revising NIH's data management and sharing policy. Data sharing enables researchers to reuse the results of experiments and supports the creation of new science that is built upon previous findings, making the research process more efficient. Data sharing also supports transparency and reproducibility, building trust in science. Elsevier is committed to supporting researchers to store, share, discover, and reuse research data and we are committed to working with other stakeholders to address challenges in making research data more effective. Our response to the Request for Information (RFI) on Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research is below. In addition to this current RFI response, Elsevier has submitted responses to all research data-related NIH RFIs over the last four years, including:

- NOT-OD-17-015, Strategies for NIH Data Management, Sharing, and Citation
- NOT-OD-16-133, Metrics to Assess Value of Biomedical Digital Repositories

- NOT-ES-15-011, Input on Sustaining Biomedical Data Repositories
- NOT-AI-15-045, Input on NIAID Data Sharing Repository, Immunology Database and Analysis Portal (ImmPort), and Services
- NOT-OD-15-067, Soliciting Input into the Deliberations of the Advisory Committee to the NIH Director (ACD) Working Group on the National Library of Medicine (NLM), Comment 5

Response

I. The definition of Scientific Data (Provisions I, II, and III)

The proposed definition is consistent with those from the [OSTP Public Access Memo](#) and [OMB circular A-110](#) and it is sufficiently flexible to allow for discipline-specific data standards setting. Having the benefit of being able to build on those earlier definitions, we propose a slightly amended definition for greater clarity to the research community (changes notes in red below):

“Research Data: The recorded factual material commonly accepted in the scientific community as necessary to validate and replicate research findings, including, but not limited to, **the underlying primary data that support the central findings of a** scholarly publication. **Research** data do not include laboratory notebooks, preliminary analyses, completed case report forms, drafts of scientific papers, plans for future research, peer reviews, communications with colleagues, or physical objects, such as laboratory specimens. For the purposes of a possible Policy, **research** data may include certain individual level and summary or aggregate data, as well as meta data. NIH expects that reasonable efforts should be made to digitize all **research** data.”

The proposed revised definition above replaces “scientific data” with “research data” throughout. While “scientific data” and “research data” are often used interchangeably, “scientific data” may be perceived as exclusive of the social and behavioral sciences and medical humanities domains, whereas “research data” is a more inclusive term encompassing all disciplines. The addition of “underlying primary data” to the definition provides further clarity that research data are distinct from the text in a manuscript or a final published article and affiliated supplementary materials published as part of a journal article, a distinction between research data and interpretations or presentations of research data. Once this policy is updated, we encourage consistency of definitions in other related data sharing policies across the NIH.

Within Provision I, we recommend adding a definition for “Research Data Repository” (a digital platform where research data is stored for the purposes of publishing, sharing, re-use, linking, and preservation). Along with the addition of a Research Data Repository definition, we also recommend that the policy include basic guidance to researchers on criteria that constitute a trustworthy repository. Some very useful resources about trustworthy repositories include:

- **CoreTrustSeal** (<https://www.coretrustseal.org/>): Offers certification based on the Core Trustworthy Data Repositories Requirements catalogue and procedures. This universal catalogue

of requirements reflects the core characteristics of trustworthy data repositories and is the culmination of a cooperative effort under the umbrella of the Research Data Alliance (RDA).

- **Repository Finder** (<https://repositoryfinder.datacite.org/about>; <https://www.re3data.org/>): A project of the Enabling FAIR Data Project in partnership with DataCite that queries the re3data registry of research data repositories.
- **Scientific Data Recommended Data Repositories** (<https://www.nature.com/sdata/policies/repositories#general>): The journal *Scientific Data* has compiled a comprehensive list of trusted discipline-specific, community-recognized, and generalist research data repositories.
- **Recommended versus Certified Repositories:** (<http://doi.org/10.5334/dsj-2017-042>) A research article that examines both recommended and certified repository characteristics. Husen, S.E., de Wilde, Z.G., de Waard, A. and Cousijn, H., 2017. Recommended versus Certified Repositories: Mind the Gap. *Data Science Journal*, 16, p.42.

Elsevier's data repository, [Mendeley Data](#), which is free to researchers globally, is recognized as a trusted research data repository in all of the directories above and has received the CoreTrustSeal.

II. The requirements for Data Management and Sharing Plans (Provision IV)

The list of elements of a data management plan (Plan) is quite comprehensive. OSP may also wish to review the [Elements of a Data Management Plan](#) at North Carolina State University, an abbreviated compilation of data management plan elements from several sources. There are some additional elements in that compilation that could be considered for inclusion by OSP for its revised policy: Roles and Responsibilities; Data Formats and Metadata; Privacy; and Costs.

Given that some projects are data-intensive and some are complex research programs and/or multi-institutional proposals, the two-page limit may be too constraining and not allow for researchers to provide sufficient detail necessary for review by peer reviewers and program staff.

OSP might consider working with the California Digital Library to add an NIH template to the [DMPTool](#), which is used by many research universities and institutions to prepare quality data management plans.

Beyond human-readable Plans, a revised policy should also address the use of machine-readable data management plans (DMPs; a.k.a., machine-actionable DMPs and active DMPs) in the not-too-distant future. Machine-readable DMPs focus on assigning identifiers and machine-actionable components of a plan. It is premature to require researchers to develop machine-readable DMPs at this time, but the revised policy could encourage researchers to develop them when possible. For additional information

on machine-readable tools and standards, we recommend the following resources, as well as others available from the Research Data Alliance (RDA) web site:

- Miksa, T., Rauber, A., Ganguly, R., & Budroni, P. (2017). Information Integration for Machine Actionable Data Management Plans. *International Journal of Digital Curation*, 12(1), 22. <https://doi.org/10.2218/ijdc.v12i1.529>
- Miksa, T., Simms, S., Mietchen, D., & Jones, S. (2018). Ten simple rules for machine-actionable data management plans (preprint). <https://doi.org/10.5281/ZENODO.1172673>
- Research Data Alliance (RDA). (2017). DMP Common Standards WG | RDA. Retrieved from <https://www.rd-alliance.org/groups/dmp-common-standards-wg>

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards. (Provision V)

We posit that a phased implementation isn't as critical as a commitment by OSP to review and revise the policy on a more regular cycle, perhaps every 2-3 years. A shorter review/revise cycle will allow OSP to be nimble and keep the policy up to date with advances in both technical and technology capabilities, such as machine-readable DMPs. A shorter cycle will also allow for timely revisions should unforeseen negative consequences result, or if previously unconsidered limitations are brought to light. Moreover, since the NLM has recently commissioned the National Academies of Sciences, Engineering, and Medicine (NASEM) to conduct a study on forecasting the long-term costs for preserving, archiving, and promoting access to biomedical data, it will be important to review this policy in consideration of the findings from that study once complete in mid-2020.

In addition, we strongly encourage OSP to set a schedule for collecting data about research data sharing practices, evaluating the impact of sharing research data on both research and researchers, and work with RDA and other community partners to develop and establish research data sharing metrics—sharing the findings with the community. These efforts underpin an evidence-based approach to science policy consistent with [the science of science policy](#) and will provide data to inform future policy changes and revisions.

Submission #80

Date: 12/07/2018

Name: Harry W. Orf

Name of Organization: Massachusetts General Hospital

Type of Organization: Other

Other Type of Organization: Research Hospital/Academic Medical Center

Role: Institutional Official

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

biomedical research

Attachment:

7 December 2018

Francis S. Collins, MD, PhD
National Institutes of Health
Bethesda, MD

Submitted electronically: <https://osp.od.nih.gov/provisions-data-managment-sharing/>

Re: RFI on Proposed Provisions for Draft Data Management and Sharing Policy

Dear Dr. Collins:

Thank you very much for providing the research community with an opportunity to comment on the NIH proposed data management/sharing policy. I am writing on behalf of Massachusetts General Hospital (MGH). The MGH is the third oldest general hospital in the United States and the original and largest teaching hospital of Harvard Medical School. A founding member of Partners HealthCare System, the MGH conducts the largest hospital-based research program in the U.S, encompassing both basic science and clinical research, and is ranked first among hospitals nationally receiving NIH funding. In FY18, MGH received approximately \$387 million in NIH/HHS research support. Thus, any proposed change in NIH data management and sharing requirements is of vital interest to us.

Let me begin by stating my colleagues and I conceptually support data sharing as a means of enabling researchers to test the validity of scientific findings, explore new scientific pathways, and shorten the time for ideas to move from the bench to the bedside. Yet, the devil is in the details for data sharing to be successful. The proposed policy is so broad and all-encompassing, we believe if implemented it would be extremely difficult for the NIH to achieve its objective of enhancing science, let alone for Principal Investigators (PI) and institutions to meet their compliance requirements.

Some of our investigators have suggested that the proposed policy appears to be an extension of data sharing requirements for genetic data to scientific data more generally. Genetic data sharing through dbGaP and similar repositories works because genetic data can be supported with standard file formats for data submission. We find it difficult to envision how the many possible experimental designs for laboratory-based experiments would be submitted and archived in a way that could be interpreted by an outside user.

We strongly recommend that the NIH revise the proposed policy to scale back its requirements, add clarity to definitions, and provide meaningful examples for investigators. We also recommend that the NIH consider convening a group of NIH-funded investigators to work with NIH research and administrative leadership to develop a policy that is more realistic and achievable from an investigator's perspective.

Please see below for our comments on specific sections of the proposed policy.

1. Section I

- a. Definitions: The definition of Scientific Data is extremely broad and confusing. We recommend considering the definition of Research Data in OMB Circular A-110 as a substitute. This definition would already be familiar to most of the research community.
 - b. Lab notebooks: Throughout the policy there is confusion about lab notebooks and whether they should be shared. Their role/purpose in a “data sharing policy” should be clarified. We maintain that lab notebooks, while critical to the scientific process, are not Scientific Data; they are a means for recording experiments and the Scientific Data generated.
 - c. Reasonable effort to digitize scientific data: While institutionally we are requiring our investigators to transition to digital recordkeeping, we do not recommend including a statement about digitizing scientific data within the current policy. Not all Scientific Data can be digitized; this makes the data no less valuable to research.
2. Section II. Purpose: Making Scientific Data accessible in a “timely manner:”
 Researchers generate data daily. We recommend clarifying this section by adding timelines for posting/sharing published and unpublished data. We recommend adding a section to the Progress Report where the PI can inform the NIH of data accessibility. The policy should be flexible. Not all data will be ready for sharing or posting in a repository at the same time. Investigators may want to refrain from posting/sharing unpublished data until it has been published. These situations should be taken into consideration in this section.
3. Section III. Scope and Requirements:
- a. We are concerned that requiring a data management/sharing plan for each application/proposal submission, when the overall funding success rate hovers at 20% or less, creates a significant administrative burden for PIs submitting applications. We recommend the NIH consider requiring the plan as part of the first progress report. These plans will not have the benefit of peer review, but is peer review necessary if the strength or weaknesses of the plan will not be considered in the impact score? Continuation funding for year 2 could be delayed until a plan acceptable to the Program Officer is submitted.
 - b. In the general statement that data management/sharing plans will be required regardless of mechanism, we recommend that the NIH review the different funding mechanisms for appropriateness. For example, a data sharing plan would not be appropriate for a shared instrumentation grant; nor would it be appropriate for a conference grant. We also recommend that the NIH consider eliminating the requirement for institutional training grant applications. We recognize that Scientific Data are generated under training grants, but the management and sharing of the data will vary across the training grant based on requirements of each trainee’s mentor who often come from different departments/research labs with different data management/sharing requirements.
 - c. The policy states, “Reasonable costs associated with data management and sharing could be requested under the budget for the proposed project.” Will supplemental funds be available for these costs? If not, the data

management/sharing requirement will only reduce the amount of funding available for the actual research project. We can envision situations where institutions with limited resources will have to provide their investigators with institutional funds or create local data repositories because the NIH funds were simply not enough to complete the project and pay for the costs associated with external data repositories essentially creating yet another unfunded mandate for grantee institutions.

4. Section IV. Requirements for Data Management and Sharing Plans

- a. General comment: We do not believe PIs will be able to provide all the information the NIH is requiring within a two-page limit.
- b. Scoring/Peer Review Process: If the NIH continues to require plans as part of the grant application/contract proposal, we agree whether a plan is acceptable or unacceptable to reviewers should not be included in the overall impact score.
- c. Plan Elements: We recommend that the NIH create a form with drop down boxes for the PI to identify the plan elements relevant for his/her research. The elements should be minimal and allow for PI flexibility.
- d. Describe type and amount of scientific data to be collected and used in the project: This may be difficult for some types of projects. The example provided is for a specific type of project in which the number of cases/patients/individuals may be known at submission. In many lab-based projects, investigators may improvise and adjust the work making use of techniques that may not have been envisaged initially. We are concerned that PIs may feel providing this type of information will restrict their ability to modify the research as they move forward.
- e. Related Tools, Software and/or Code: Please clarify what the NIH is expecting. For example, would the PI have to justify use of a specific image analysis software product?
- f. 4.1 Indicate where Scientific Data will be archived to ensure long-term preservation: We recommend that the NIH create data repositories to meet this new mandate. As we indicated above, many institutions do not have the resources to develop and maintain repositories for their NIH-funded investigators. Grantee institutions cannot continue to absorb unfunded mandates. Moreover, we are concerned at the possible development of numerous and heterogeneous and possibly rogue repositories.
- g. 4.4 Describe alternative plans for maintaining, preserving and providing access to scientific data should the original plan not be achieved: If the NIH is truly interested in this information, we recommend not requiring submission of a “Plan B” as part of the data management/sharing plan in their application/contract proposal. We recommend adding a section to the data management/sharing reporting section of the annual progress report to describe any changes necessary because the original plan could not be achieved.
- h. 5. Data Preservation and Access Timeline: We question the usefulness of requiring this information in the data management/sharing plans. It may be impossible at the beginning of the project to estimate timelines. This may lead PIs to develop meaningless timelines which become a compliance requirement if the application is funded. We recommend removing this requirement.

- i. 6. Data Sharing Agreements, Licensing and Intellectual Property:
 - i. “NIH encourages terms that provide for the broadest use of data resulting from NIH-funded or -supported research.” Please confirm/clarify that this statement applies to data generated as part of the study, i.e., the data would not exist if not for the study; and does NOT include any additional, pre-existing clinical data, e.g., annotated, longitudinal data pulled from a patient’s medical record.
 - ii. 6.1 “Describe any relevant data sharing agreements outlining...how scientific data can and cannot be used.” Please confirm/clarify that this applies only to Scientific Data generated as part of the study. In a situation where the project is supported by NIH and industry or a foundation, the non-NIH sponsors may limit data sharing. Would an SBIR grant be relevant here?
 - iii. 6.3 “[I]ndicate how intellectual property...will be managed in a way to maximize sharing of scientific data.” While Scientific Data do not constitute IP, any plan to maximize sharing should not infringe upon the nature of the IP and should preserve ownership rights.
5. Compliance and Enforcement
- a. Community-based Standards: The NIH should specify these standards within the policy or at a minimum provide examples. When we consulted our investigators to develop our response, they were unsure what the standards were and where they might find them.
 - b. I/C Monitoring Plans: The policy should include information on how I/Cs will monitor plans, reporting requirements, how to modify plans during the lifetime of the grant. If an I/C determined non-compliance, what would be the enforcement mechanism?
 - c. We are very concerned about compliance/enforcement requirements extending beyond the end of the grant’s performance period. If this requirement continues in the policy, the NIH should identify the authority that allows the requirement to continue in perpetuity. Comments made during the NIH webinar on the RFI seemed to suggest that the NIH does not consider data sharing requirement as continuing beyond the project end date. The proposed policy contradicts this point and should be clarified. How will the NIH monitor? How will a grantee know if a former award is out of compliance?

Once again, thank you for providing an opportunity for the research community to submit comments. Please do not hesitate to contact me for any additional information.

Yours sincerely,

Harry W. Orf

Harry W. Orf, PhD

Senior Vice President for Research, MGH

Submission #81

Date: 12/07/2018

Name: Xia Jing

Name of Organization: Ohio University

Type of Organization: University

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

Health informatics

Attachment:

Comments on: Request for Information on Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research

2018-12-08

Xia Jing

I first saw the RFI announcement in the NIGMS newsletter in October. Then, at AMIA 2018 in San Francisco, Dr. Valerie Florance from NLM emphasized the RFI in her presentation, which made me mark the RFI deadline in my calendar.

The document is well thought through and well prepared. I have only some minor points to share.

1. The definition of scientific data

In regard to metadata definition, my impression is that many researchers outside of the data science or informatics world, who use the data dictionary a lot, would not necessarily use metadata, per se. Therefore, I wonder whether the data dictionary should be mentioned here, just to provide a broader audience with a connection or a reference point to something that they may not have previously heard of.

In regard to scientific data, I do not know whether it is feasible or even necessary to maintain a list of scientific data examples that should be *included* and a list of scientific data examples that should be *excluded*, considering the broad nature of research funded by NIH. Over the long run, such a list would be helpful for new investigators to make a more accurate judgment about what should be counted as scientific data for sharing purposes. For example, one of my funded studies will involve a collection of video clips of study participants, using an online tool with audio that explains what a participant is doing. Under the current definition of scientific data, I understand that the original videos should not be submitted as scientific data for sharing. I do wonder, however:

- How about the transcript of the audio?
- How about the coded data from the analysis of each video clip?
- How about the intermediate analysis results, which are still more inclusive than are the published results?
- When should we submit the data? As soon as we have the data ready? After the papers are published?

2. The requirements for data management and sharing plans

- How about a basic template for a data-sharing agreement and data usage rules? The applicant can work from there to customize the documents based on an individual project's requirements.
- How about recommending existing data repositories that can meet the data security requirements? With such a recommendation, investigators will have the option to archive the scientific data in one of the designated data repositories without having to consider meeting the preservation, access, and safety requirements of individual groups.

===

Xia Jing, M.D., Ph.D.
Assistant Professor of Clinical Informatics/Health Services Administration
Department of Social and Public Health
College of Health Sciences and Professions

Grover Center W357
Ohio University
Athens, Ohio, USA
45701
Tel: 740-593-0750
Fax: 740-593-0555

Submission #82

Date: 12/07/2018

Name: James P Sluka

Name of Organization: Indiana University

Type of Organization: University

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

Indiana University: biomedical research including a range of Omics, preclinical and clinical research. Biocomplexity Institute: computational biology.

I. The definition of Scientific Data

Please see the Section I in the attached document.

II. The requirements for Data Management and Sharing Plans

Please see the Section II in the attached document.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

Please see the Section III in the attached document.

Attachment:

Section I —

Define Data — Everything that a scientist produces when developing and running an experiment. This includes experimental design, work flow, generated data, and data analysis.

Section II —

Data management and sharing plans must define a standard method of annotating, sharing, facilitating search and reuse of data.

Sharing (FAIR) is widely done by Ecommerce entities. Amazon lists millions of products, all of which are indexed in Google. Scientific FAIR has lagged Ecommerce sharing do to lack of incentives to share and the lack of tools to make sharing easy. The technology exists to share biological data, as a community we have just not taken advantage of those technologies.

Many biotech communities are developing their own data and annotation standards. For example, MIRIAM in the modeling community, OMERO in the microscopy community, and MIAME in the gene expression community. What has been lacking is that often the standards do not define the biology the experiment is exploring. For example, MIRIAM annotation does not require any biological descriptors. What has not been widely recognized is that at the level of biological description the technological details (e.g., the maker of the microscope or of the gene microarray) fall away and the biological description is the same across all biotech domains. As outlined in the **Figure 1**, what has been lacking is the recognition that the biological description of an experiment (or model) is common across the multiple biotechnologies that might be used to study the problem. Efforts are needed to ensure that data generated across biological and biotechnological domains use a consistent method of describing the biology studied in the experiment.

Furthermore, the method of annotating an experiment (or data set) must be compatible with existing web search engines. This aspect has not been widely recognized and is a significant impediment to finding and reuse of data. This in turn reduces the value returned to funding organizations for their research dollars.

Another key aspect of any standard for annotating and sharing of biological data is that it should recognize that this is an international challenge and it is critically important that an international standard is developed.

A final critical point is that this challenge spans more than just the human health domain. It is critical that non-human biological research is just as findable and shareable. Non-human research, such as food plants and animals, engineered organisms and bioremediation, may provide data useful in human health research. The two point above suggest that NIH data sharing standards should be compatible with the data sharing standards in other areas of biological research as well as in other nations.

Long Term Vision: Common annotation across multiple data sources (Somitogenesis Example)

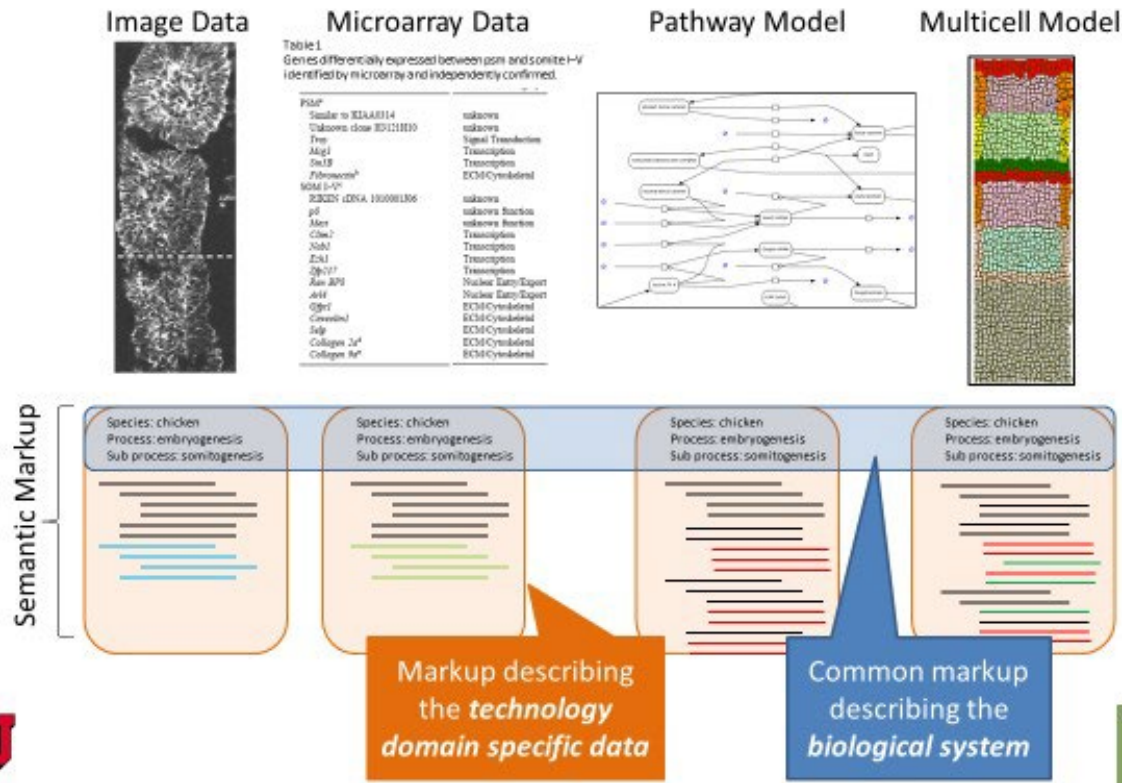


Figure 1: Multiple biotechnologies can be applied to a biological research problem. As an example, we show multiple biotech's applied to the same biological question of somitogenesis in early embryonic development. Biotech's of high resolution imaging, gene expression, pathway and tissue modeling (as well as many others) have been applied to this biological problem. Individual biotech research communities are developing their own domain-specific annotation schemes (OMERO, MIAME, SBML, etc.). What has been lacking is the recognition that the biological annotation should be the same across the biotech domains and that the annotation schemes should be compatible and include the same biological descriptors.

Some desirable features for data sharing and annotation standards include:

- Annotation should not require deep expertise in ontologies and the process of annotation. This suggests there are opportunities to develop intelligent annotation tools. Tools that embed knowledge of the annotation process, the types of concepts that should be annotated (and the relevant bio-ontologies for those concepts), and the proper syntax and reification of the annotations.
- The annotation standard should be syntax independent. Tools are needed that can convert annotations into the syntax needed for a particular application or use. The syntax for web sharing is different from that for mining by automated processes, which may be different from biological domain and/or biotech domain specific usage.

- Data should be findable without the user (the data consumer) knowing where to look.
- Large central repositories are not the only approach. While large repositories of biological data are important, it is equally important to recognize small local repositories as well as the repositories maintained by journal publishers. Publications represent a critical data resource and annotation standards should be applied to publications as well as data repositories.
- A concerted effort is needed to ensure that we leverage existing web search tools and technologies for making massive amounts of data findable via common web search engines.
- A data annotation and sharing standard must include defined methods to test if the data is findable. For example; "Is the data findable and retrievable via web search with queries from controlled vocabularies?"

Annotation minimum requirements — Must include keyword annotations using terms from bio-ontologies. This removes the current ambiguity in entity and process names. All data records should be annotated in a standard way using bio-ontologies. Records can also be annotated with human readable versions (for cases where the ontological term is numeric) and may include alternate names, acronyms etc.

Repository minimum requirements — Must be searchable via standard public web search engines (e.g., Google). Must be indexed using, at least, terms selected from bio-ontologies, including numeric, alphanumeric and human readable versions.

A "standard annotation block" is needed that defines the minimal list of annotations, along with suggestions of suitable bio-ontologies that can supply the needed terms. The standard annotation block might consist of;

Standard Annotation block

- 1) Species (biological research goes beyond human medical applications)
- 2) Sex
- 3) Age
- 4) Biological Big Question, e.g., why was the experiment carried out? (often a term from GO, or a disease or fundamental biological process name)
- 5) Organ
- 6) Tissue
- 7) Cell Type
- 8) List of manipulated entities. What was added to or modified in the experiment? Manipulation includes selection, e.g., when comparing a diseased and a normal population the disease is the manipulated entity even if an intervention is not part of the experiment.
- 9) List of observable and/or measured entities (cells, molecules, processes such as growth, ...)

Note that many of the entries in the standard annotation block shown above can be [linked to specific bio-ontologies](#). This provides coherence and consistency across annotations.

Note that if the standard annotation block is developed as a table then that table can be converted into a set of RDF triples suitable for use in many ontology languages (e.g., OBO or OWL). Therefore, the standard annotation block is a **knowledge construct** compatible with a wide range of reasoning tools.

Examples

Google indexes scientific content. A Google search with a sentence from a scientific article will locate that article if it is in an accessible location. If an article is annotated with ontological terms (be they human readable or alphanumeric) then those terms can be used in a Google search.

A wide range of file types are indexed by Google. Text, PDF, DOC and other standard formats for prose are well indexed. Even non-prose formats, such as computer source code, are indexed by Google. We have shown that web accessible Python code containing ontology terms is indexed by Google and can be found via Google searches.

An important consideration though is that certain file syntaxes are poorly indexed by Google. For example, XML is often poorly indexed by Google. This shows that it is critical that the data annotation standard recognize the importance of using a file format that can be indexed by standard web search engines.

The development of standard annotation schemes and technologies is an enabling technology for the widespread re-use of biological data. This will facilitate both reuse by researchers (by facilitating FAIR) and also enable data mining techniques that will benefit from having access to the massive amounts of data generated in biological research. Currently data mining technologies are severely hampered by their inability to effectively use resources such as journal articles.

Section III —

Phase 1: Require scientific publications have a standard annotation block. This block is similar to the often seen "keywords" and acronym lists but combines those concepts into a single table. The table uses bio-ontologies as the source of the keywords and to define the acronyms. Note that this would also be of value in annotating research proposal and research Funding Announcements (RFA's).

Phase 2: Data repositories (e.g., GenBank, or image repositories) present the same structured, high-level annotation to the world as publications do. Here we can leverage the tools and technologies developed in eCommerce to make the massive amounts of data in these databases available to web search engines.

Needs:

1. Tools to make annotation easy. It must be trivially easy to generate the standard annotation block for a data resource.
2. Tools that embed knowledge of the annotation process, so researchers don't have to be expert in annotation to properly annotate a data resource. Tools can be developed to intelligently guide the user through the annotation process. In the domain of annotating computational models the SBML standard provides the needed functionality but;

- a. Compare the BioModels (<http://www.ebi.ac.uk/biomodels/>) model repository's annotation process; BioModels uses expert annotators and the workflow has not been published. This makes annotation much too difficult.
 - b. Compare COPASI's (<http://copasi.org/>) annotation tools that embed little knowledge of the annotation process, or the type of object being annotated (reaction, molecule, enzyme, ...), and instead simply lists all possible options.
3. Tools to auto-annotate archival resources and generate the standard template. Note that current text mining approaches for e.g., publications do not really know what information is being searched for and, in general, do not attempt to translate prose into terms from bio-ontologies, or create the RDF triple that the standard template creates. Automated tools can be developed to map human readable prose (like journal articles) into the bio-ontology terms to fill the standard annotation table. The same automated tools can also be used to assist researchers in annotating new data resources.
 4. Annotation and search are closely coupled process. Tools are needed that use the same term identification process for both annotation and search purposes.

Finally, it is critical that any data annotation and sharing standard recognizes that this is an international problem that extends beyond human health research. Non-human health related biological research is likely to be relevant to human health and vice versa. OUS research is of great value to US based researchers. Therefore, the NIH should collaborate with OUS efforts such as the International Standards Organization's (ISO) efforts on data standards in the biological sciences. This effort is coordinated in the US by the national Institute of Standards (NIST). Other active OUS efforts in data standards include the EU's CHARME effort (<https://www.cost-charme.eu/home>).

James P. Sluka, PhD
Biocomplexity Institute
Intelligent Systems Engineering
School of Informatics, Computing and Engineering
Indiana University
Bloomington, IN USA
Office: 812-855-2441 Cell: 317-331-7465 JSluka@Indiana.edu

Submission #83**Date:** 12/07/2018**Name:** Raja Mazumder**Name of Organization:** The George Washington University**Type of Organization:** University**Role:** Scientific Researcher**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

genomics, cancer, biomarker discovery, database, data integration; For more information on OncoMX and the relevance of this policy to the project, please see the attached full response pdf below.

I. The definition of Scientific Data

With “Big Data” launching scientific data collected and/or generated by biomedical research into the exciting and challenging realm of data management and data sharing, the National Institutes of Health (NIH) is opportunistically positioned to shape future policy. For the purposes of policy pertaining to the management of such data, we define scientific data to be any information that (1) has been collected, refined, analyzed, or produced using systematic methods, such as through experiments and observations, with the intent of studying and analyzing that information, (2) contributes knowledge to a particular subject, and (3) can be used to validate or replicate previously reported findings. For example, “information” can include experimental input, protocols and procedures, materials and methods, or results; or more specifically, mathematical equations and calculations, graphs, tables, images, audio and video recordings, algorithms, programming code and scripts, variant calls, differentially expressed genes, biomarker evidence, literature mining hits, supplemental data, metadata (clinical and technical), and negative findings. As such, while we generally agree with the proposed definition of scientific data, we would qualify that preliminary and interim analyses are also scientific data and can have beneficial contributions if maintained. We do understand that these data may fall outside of the scope of management covered by NIH policies, but we suggest that a truly comprehensive data management plan would likely make provisions for these data as well.

II. The requirements for Data Management and Sharing Plans

For NIH-funded or supported research, the data management and sharing plan (DMS) should include regulations for submitting, acquiring, validating, storing, protecting, processing, and accessing data which also support the adherence to FAIR principles; data is Findable, Accessible, Interoperable, and Reusable with adequate provenance. The responsibility for viable data management and sharing is two-fold, both belonging to the NIH and to the researchers generating the data subject to the new DMS guidelines. Suggestions for the NIH to facilitate a smoother data transfer process and more clear communication and enforcement of data management and sharing requirements are as follows:

- Standardize the process of scientific data submission from NIH-funded/supported research sources to the NIH via adoption of provenance collection (for example, BioCompute Object model)
- Harmonization of various data types by implementing/updating existing ontologies or creating new ontological systems and databases, which would also assist in data integration between multiple/different platforms
- Simplify access to archived, current, and emerging data
- Developing infrastructure that is both human e.g. txt and machine readable e.g. json and providing simple tools for conversion between the two (where applicable)
- Create incentives for adoption of standardized methods e.g. a submission portal that EITHER requires string capture for various fields OR an option to upload a file of a preferred, specified, standardized format
- Improve interoperability of NIH-generated or supported data with non-NIH-generated or supported data, as possible
- Guidelines/ranges for preservation and access timelines proposed by NIH would facilitate more informed and realistic responses from researchers
- Suggest and/or endorse existing repositories that make scientific data available at no cost for extended periods of use - this could be accomplished through the establishment of a consortia among such repositories

Additional suggestions for the plan requirements that must be adhered to by a researcher generating the data are below:

- Require capture of data provenance and lineage, where applicable. This is extremely important to the reusability of data and their ability to reproduce previous findings

Attachment:

From: Raja Mazumder (GW), Daniel Crichton (NASA-JPL, EDRN), Frederic Bastian (SIB, UNIL), K. Vijay-Shanker (UD), Hayley Dingerdissen (GW)

SUMMARY:

On behalf of the OncoMX team, we generally approve of the proposed definition of scientific data, but would qualify that while preliminary and interim analyses are likely outside the maintenance scope for policies related to management of scientific data, they do, in fact, qualify as scientific data and can be beneficial if maintained. We also feel that organization, management, and availability of so-called negative findings would greatly contribute to future research. Regarding the proposed data management and sharing plan requirements, we strongly suggest that the research element labeled “data standards” be augmented with a well-described provision for capture of provenance, error domain, and usability domain as those described in BioCompute and other ontologies, in addition to the suggestion of using common data elements. We endorse the encouragement of researchers to use repositories that make scientific data available at no cost for extended periods of use, and suggest that establishment of a consortia between the NIH and such repositories could greatly enhance the accessibility and longevity of scientific data. For more information, please find a detailed response below.

DETAILS:

Use case

OncoMX - integrating scientific data and hosting analyzed data via web portal

OncoMX is an integrated cancer data resource facilitating the exploration of multi-faceted cancer data from four perspectives: (1) Exploring biomarkers; (2) Evaluating mutation and expression in an evolutionary context; (3) Side-by-side exploration of published information for gene mutation and expression; and (4) Exploring a specific gene (biomarker) within a pathway context. The project uses data in all stages of the data lifecycle: primary and some pre-analyzed data are retrieved from public repositories with varying levels of access; data are then subjected to quality control (depending on data type) and unification prior to integration; data are analyzed and/or packaged (depending on data type); data (primary and post-analysis) are made available via web portal, as custom exports, or as bulk downloads from the website. The existence and availability of scientific data drives OncoMX research. Appropriate management and sharing policies will ensure the continuity of the project and improve the integrity of future versions.

OncoMX group (PIs, Co-Is, and collaborators)

Principal Investigators

Raja Mazumder (George Washington University)

(Background: Cancer Genomics, NGS standards, Alliance of Glycobiologists, NCBI, UniProt)

Daniel Crichton (NASA Jet Propulsion Laboratory)

(Background: Computational infrastructure for big data, Early Detection Research Network)

Subawardees

Frederic Bastian (Swiss Institute of Bioinformatics, Universite de Lausanne)

(Background: Comparative genomics, Evo-Devo, evolution of gene expression patterns, Bgee, ontologies, biocuration)

Vijay Shanker (University of Delaware)

(Background: Natural language processing, machine learning, DEXTER, DiMeX)

Other key collaborators

William Evan Johnson (Boston University)

(Background: precision genome medicine, tumor heterogeneity, scRNA-seq analysis)

Marc Robinson-Rechavi (Swiss Institute of Bioinformatics, Universite de Lausanne)

(Background: Evo-Devo, comparative functional genomics, Bgee, Professor of Bioinformatics)

Hayley Dingerdissen (George Washington University)

(Background: Cancer genomics, data integration, expression of glycosyltransferases in cancer)

I. The definition of Scientific Data

With “Big Data” launching scientific data collected and/or generated by biomedical research into the exciting and challenging realm of data management and data sharing, the National Institutes of Health (NIH) is opportunistically positioned to shape future policy. For the purposes of policy pertaining to the management of such data, we define scientific data to be any information that (1) has been collected, refined, analyzed, or produced using systematic methods, such as through experiments and observations, with the intent of studying and analyzing that information, (2) contributes knowledge to a particular subject, and (3) can be used to validate or replicate previously reported findings. For example, “information” can include experimental input, protocols and procedures, materials and methods, or results; or more specifically, mathematical equations and calculations, graphs, tables, images, audio and video recordings, algorithms, programming code and scripts, variant calls, differentially expressed genes, biomarker evidence, literature mining hits, supplemental data, metadata (clinical and technical), and negative findings. As such, while we generally agree with the proposed definition of scientific data, we would qualify that preliminary and interim analyses are also scientific data and can have beneficial contributions if maintained. We do understand that these data may fall outside of the scope of management covered by NIH policies, but we suggest that a truly comprehensive data management plan would likely make provisions for these data as well.

II. The requirements for Data Management and Sharing Plans

For NIH-funded or supported research, the data management and sharing plan (DMS) should include regulations for submitting, acquiring, validating, storing, protecting, processing, and accessing data which also support the adherence to FAIR principles; data is Findable, Accessible, Interoperable, and Reusable with adequate provenance. The responsibility for viable data management and sharing is two-fold, both belonging to the NIH and to the researchers generating the data subject to the new DMS guidelines. Suggestions for the NIH to facilitate a smoother data transfer process and more clear communication and enforcement of data management and sharing requirements are as follows:

- Standardize the process of scientific data submission from NIH-funded/supported research sources to the NIH via adoption of provenance collection (for example, BioCompute Object model)
- Harmonization of various data types by implementing/updating existing ontologies or creating new ontological systems and databases, which would also assist in data integration between multiple/different platforms

- Simplify access to archived, current, and emerging data
- Developing infrastructure that is both human e.g. txt and machine readable e.g. json and providing simple tools for conversion between the two (where applicable)
- Create incentives for adoption of standardized methods e.g. a submission portal that EITHER requires string capture for various fields OR an option to upload a file of a preferred, specified, standardized format
- Improve interoperability of NIH-generated or supported data with non-NIH-generated or supported data, as possible
- Guidelines/ranges for preservation and access timelines proposed by NIH would facilitate more informed and realistic responses from researchers
- Suggest and/or endorse existing repositories that make scientific data available at no cost for extended periods of use - this could be accomplished through the establishment of a consortia among such repositories

Additional suggestions for the plan requirements that must be adhered to by a researcher generating the data are below:

- Require capture of data provenance and lineage, where applicable. This is extremely important to the reusability of data and their ability to reproduce previous findings.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards.

We have no comments on this topic at this time.

Best regards,
Raja Mazumder
Daniel Crichton
Frederic Bastian
K. Vijay-Shanker
Hayley Dingerdissen

Submission #84**Date:** 12/07/2018**Name:** Rebecca Osthus**Name of Organization:** American Physiological Society**Type of Organization:** Professional Org/Association**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Physiology

I. The definition of Scientific Data

The American Physiological Society (APS) supports the sharing of scientific data that are used to make conclusions reported in scholarly publications. In addition to data that are used to support conclusions reported in publications, negative data generated using grant funds should also be preserved and made publicly available if at all possible, even if not reported.

As a publisher of 15 scientific journals, the society's publications policies (1) already encourage authors to "make data that underlie the conclusions reported in the article freely available via public repositories or available to readers upon request." In addition, certain specific types of data such as sequences and microarray data must be published in an appropriate repository prior to manuscript submission. Authors may also include a URL linking to data housed on their institutional website.

(1) <https://www.physiology.org/author-info.data-repositories>

II. The requirements for Data Management and Sharing Plans

Data management and sharing plans should address the basic research elements as described in the proposed provisions. It is particularly important that the plans contain adequate information on how any data derived from human participants or biospecimens will be managed, stored, and shared in a way that protects participant privacy and confidentiality and enables its reproducibility.

As new policies are implemented, NIH should provide detailed guidance and examples as researchers prepare data management and sharing plans as part of their grant applications.

The review of data management and sharing plans should not be considered in the overall impact score of a grant application, but rather evaluated to determine whether it adequately addresses how data will be managed, shared and stored. Instructions and training should also be provided to reviewers being asked to evaluate the adequacy of these plans. In order to minimize the administrative burden on applicants APS also recommends that data management and sharing plans be requested with other just-in-time materials.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

While certain types of data (sequence, microarray) may be shared in a standardized format with minimal processing required other types of data would require significant reformatting to be included in currently available data repositories. Requiring deposition of all scientific data in publicly available repositories has the potential to add significant administrative burden to grantees. Funded investigators already face a significant level of administrative and regulatory burden associated with federal grants and imposing additional requirements will further limit the amount of time they can spend focused on engaging in cutting-edge research. NIH should consider these possible consequences as policies are developed and implemented.

Many types of data generated and used in physiology are complex and not easily standardized for deposition in a currently available general data repository. NIH should work with investigator communities to determine what types of repositories, templates and standards are needed to facilitate sharing of data within a particular discipline. These resources should be developed, tested and available before requirements for sharing are fully implemented.

As data sharing policies are implemented, NIH should be prepared to provide necessary resources for compliance, including those needed to minimize the effects of additional administrative burden. This could include supplements to defray costs associated with preparing data for deposition. NIH should also consider increasing the modular R01 budget. The budget has been limited at \$250,000 since it was established in 1999, with no adjustments for inflation, and no increases to accommodate costs associated with considering sex as a biological variable in animal experiments. Additional administrative requirements will further strain research budgets.

Any new policies for data management and sharing should include clear guidance about the deadlines for data deposition and sharing. Investigators should be allowed adequate time to analyze data for their own purposes before being required to share data publicly.

Submission #85**Date:** 12/07/2018**Name:** Greg Raschke, Senior Vice Provost and Director of Libraries**Name of Organization:** North Carolina State University Libraries**Type of Organization:** University**Role:** Institutional Official**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

North Carolina State University's research enterprise is broad and interdisciplinary, encompassing, among other areas, a wide range of genomics, health, and life sciences disciplines such as bioinformatics, environmental health science, genetics and genomics, molecular biology, translational regenerative medicine, and all aspects of veterinary medicine. The university brings together top scholars from diverse backgrounds to collaborate with each other and with public and private sector partners to address the world's grand challenges. As the largest academic institution in the North Carolina, the university enrolls almost 34,000 students, offering bachelor's and master's degrees in more than 120 fields of study and doctoral degrees in 67 disciplines.

Librarians at NC State collaborate intensively with university researchers in all disciplines and on emerging tools and technologies for research and scholarly communication in a changing environment. We offer consultation and guidance during all phases of the research data lifecycle, from developing data management plans for grant proposals, to consulting on best practices and appropriate infrastructure for data storage and preservation, to optimizing the sharing and discovery of data. We also advise on copyright and intellectual property issues.

I. The definition of Scientific Data

This definition of "scientific data" is comprehensive. Because metadata plays such a vital role in the discoverability, reusability, and reproducibility of data, we support the inclusion of the term "metadata" here. However, we want to note that from a legal standpoint, this could present issues with regard to copyright, where data and metadata can have unclear copyright status.

II. The requirements for Data Management and Sharing Plans

General suggestions:

- We recommend that any guidelines or metrics that reviewers would use to evaluate proposals be described or shared with the grant proposers so that they can better understand how to best meet these requirements.

- We suggest that the NIH include language emphasizing the importance of documenting adherence and/or changes to data management and sharing plans, as well as brief guidelines as to the amount of detail and how often the NIH recommends that this be done.

- In Section II, the NIH proposes that scientific data be “made accessible in a timely manner for appropriate use by the research community and broader public.” We suggest that the NIH clearly define “timely.”

Section IV Part 1 (Data Type):

- (1) The proposed language states that the grant proposer should “indicate the rationale for which scientific data will be preserved and shared.” We believe that preserving and sharing data should be the default, and therefore it would be preferable to ask for the rationale if a grant proposer indicates that they will NOT preserve or share data. This would align more closely with the practices of other grant funders.

- We recommend encouraging the use of nonproprietary formats when possible.

- Including language about the importance of versioning of datasets would be helpful.

Section IV Part 2 (Related Tools, Software, and/or Code):

- A grant proposer should also be required to include the versions of software used and the computing environment for better reproducibility. It would also be beneficial to document plugins or modules within the software used to generate or render data.

- We recommend encouraging the use of open software and code for reproducibility.

Section IV Part 3 (Standards):

- It is unclear how flexible the NIH is about the use of non-CDEs. It would be helpful to indicate whether a grant proposer needs to describe the rationale behind using different data elements.

Section IV Part 4 (Data Preservation and Access):

- (4.4) Noting the two-page limit, it would be helpful to clarify the level of detail recommended for describing alternative plans should the original plan not be achieved.

Section IV Part 6 (Data Sharing Agreements, Licensing, and Intellectual Property):

- (6.1) We support and recommend that the NIH consider encouraging the use of non-proprietary third-party data, where possible, in this section.

- (6.2) Acknowledging that Creative Commons is well recognized in the field and a defined standard, we recommend that the NIH include language suggesting the use of Creative Commons licenses, specifically CCO or CCPD, as well as linking to the Creative Commons webpage on Open Data, found at <https://creativecommons.org/about/program-areas/open-data/>

This would align with several federal departments, including the Department of Education, the Department of Labor, and the Department of State.

-(6.3) We find the use of the term “invention” peculiar, and wonder whether “patent” would be more appropriate since the sentence describes a right.

Attachment:

TO: National Institutes of Health

DATE: December 7, 2018

RE: Response to Proposed Provisions for a draft NIH Data Management and Sharing Policy

Submission online at <https://osp.od.nih.gov/provisions-data-managment-sharing/>

Name: Greg Raschke, Senior Vice Provost and Director of Libraries

Name of Organization: North Carolina State University Libraries

Type of Organization: University

Role: Institutional Official

Research Area Most Important to You or Your Organization (e.g., clinical, genomics, neuroscience, infectious disease, epidemiology)

North Carolina State University's research enterprise is broad and interdisciplinary, encompassing, among other areas, a wide range of genomics, health, and life sciences disciplines such as bioinformatics, environmental health science, genetics and genomics, molecular biology, translational regenerative medicine, and all aspects of veterinary medicine. The university brings together top scholars from diverse backgrounds to collaborate with each other and with public and private sector partners to address the world's grand challenges. As the largest academic institution in the North Carolina, the university enrolls almost 34,000 students, offering bachelor's and master's degrees in more than 120 fields of study and doctoral degrees in 67 disciplines.

Librarians at NC State collaborate intensively with university researchers in all disciplines and on emerging tools and technologies for research and scholarly communication in a changing environment. We offer consultation and guidance during all phases of the research data lifecycle, from developing data management plans for grant proposals, to consulting on best practices and appropriate infrastructure for data storage and preservation, to optimizing the sharing and discovery of data. We also advise on copyright and intellectual property issues.

I. Definition of "Scientific Data"

- This definition of "scientific data" is comprehensive. Because metadata plays such a vital role in the discoverability, reusability, and reproducibility of data, we support the inclusion of the term "metadata" here. However, we want to note that from a legal standpoint, this could present issues with regard to copyright, where data and metadata can have unclear copyright status.

II. The requirements for Data Management and Sharing Plans

- **General suggestions**
 - We recommend that any guidelines or metrics that reviewers would use to evaluate proposals be described or shared with the grant proposers so that they can better understand how to best meet these requirements.

- We suggest that the NIH include language emphasizing the importance of documenting adherence and/or changes to data management and sharing plans, as well as brief guidelines as to the amount of detail and how often the NIH recommends that this be done.
 - In Section II, the NIH proposes that scientific data be “made accessible in a timely manner for appropriate use by the research community and broader public.” We suggest that the NIH clearly define “timely.”
- **Section IV Part 1 (Data Type)**
 - (1) The proposed language states that the grant proposer should “indicate the rationale for which scientific data will be preserved and shared.” We believe that preserving and sharing data should be the default, and therefore it would be preferable to ask for the rationale if a grant proposer indicates that they will NOT preserve or share data. This would be align more closely with the practices of other grant funders.
 - We recommend encouraging the use of nonproprietary formats when possible.
 - Including language about the importance of versioning of datasets would be helpful.
- **Section IV Part 2 (Related Tools, Software, and/or Code)**
 - A grant proposer should also be required to include the versions of software used and the computing environment for better reproducibility. It would also be beneficial to document plugins or modules within the software used to generate or render data.
 - We recommend encouraging the use of open software and code for reproducibility.
- **Section IV Part 3 (Standards)**
 - It is unclear how flexible the NIH is about the use of non-CDEs. It would be helpful to indicate whether a grant proposer needs to describe the rationale behind using different data elements.
- **Section IV Part 4 (Data Preservation and Access)**
 - (4.4) Noting the two-page limit, it would be helpful to clarify the level of detail recommended for describing alternative plans should the original plan not be achieved.
- **Section IV Part 6 (Data Sharing Agreements, Licensing, and Intellectual Property)**
 - (6.1) We support and recommend that the NIH consider encouraging the use of non-proprietary third-party data, where possible, in this section.
 - (6.2) Acknowledging that Creative Commons is well recognized in the field and a defined standard, we recommend that the NIH include language suggesting the use of Creative Commons licenses, specifically CC0 or CCPD, as well as linking

<https://creativecommons.org/about/program-areas/open-data/>

This would align with several federal departments, including the Department of Education, the Department of Labor, and the Department of State.

- (6.3) We find the use of the term “invention” peculiar, and wonder whether “patent” would be more appropriate since the sentence describes a right.

Submission #86

Date: 12/07/2018

Name: Anonymous

Name of Organization:

Type of Organization: University

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

Our university is a major research institution and is involved in a broad range of research, including all of the areas cited above.

Attachment:

The notion of data sharing, as implied in the proposed policy, poses several potential concerns:

In sharing data, how will the sharing party track use of the data by the using party in order to measure its impact? This is relevant to applications for tenure, as well as ethical and meaningful use of the data.

At the conclusion of a study, it is neither necessarily practical, nor feasible to make data available right away. This is especially so, as doing so could potentially rob originating researchers of opportunities to pursue publications before competitors potentially use the data to represent the research as their own.

In some cases, removal of PHI before making the data available could significantly reduce the utility of the data for other researchers. Specifically, in some medical research where key variables include PHI and are pertinent to the results of the research.

Given the potentially realistic complexity of an appropriate data management and sharing plan, it seems unrealistic to place a two page limit on them in applications for funding. Rather, there should be no limit on the number of pages in the **Plan**. This is especially so, given the information that is called for in the proposed changes (data code, data types, information on software, number of cases, etc.).

It is unrealistic to ask for computer code that will be used to analyze data, as some research groups may be new and, therefore, do not have a person in place who has the expertise to write requisite code.

It is unrealistic to ask that data sharing agreements be outlined in advance of proposed research, as the researcher may not necessarily know in advance what other parties may take an interest in the data and whether a data use agreement may be necessary.

The same would apply to licensing.

Submission #87

Date: 12/07/2018

Name: Juliet P. Lee

Name of Organization: Prevention Research Center of PIRE

Type of Organization: Nonprofit Research Organization

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

social-behavioral

I. The definition of Scientific Data

The RFI specifies "digital" scientific data; the policy should clarify whether it only applies to data which are maintained in digital (electronic) form, or to any data collected under a funded program of research, which may include data not digitally recorded and maintained.

II. The requirements for Data Management and Sharing Plans

The requirements for data management and sharing are not in keeping with best practices in community data oversight: (1) Tribal sovereignty should be acknowledged and included in data management and sharing plans. (2) Vulnerable and/or protected communities of identity, including communities historically excluded from and/or violated by programs of research, should be acknowledged and included in data management and sharing plans; for example, sexual/gender minorities, People With Disabilities, immigrants/refugees, African Americans, homeless, poor, and working class communities. Use of data (e.g., analysis and interpretation) obtained from these populations taken out of the original data collection contexts entails substantial risks of harm to communities in the forms of stigma, loss of reputation, and loss of vital rights, resources, and social goods. Data management and sharing plans should request investigators note whether their research includes data from (A1) American Indians and/or Alaska Natives, and if yes (A2) Describe plans for Tribal community oversight, e.g., Tribal board review of data sharing requests; and (B1) Vulnerable and/or historically excluded communities of identity, and if yes (B2) Describe plans for community oversight, e.g., community or appointed representative review of data sharing requests.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

NIH should assemble an Expert Panel composed of Tribal leaders or their designated representatives (e.g., NCAI) and representatives from vulnerable/historically excluded populations (may start with referrals from NIMHD) to review and refine the policy and rollout.

Submission #88

Date: 12/07/2018

Name: UC Davis Library

Name of Organization: University of California, Davis

Type of Organization: University

Role: Institutional Official

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

clinical sciences, genomics, neuroscience, population health, animal sciences, translational science, biological sciences

I. The definition of Scientific Data

Please see attached document for our response.

II. The requirements for Data Management and Sharing Plans

Please see attached document for our response.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

Please see attached document for our response.

Attachment:

Summary of Key Points:

- I. Definition of Scientific Data
 - Distinguish between basic research data and human subjects data (e.g., biospecimens, private information)
- II. Plan Requirements
 - Provide metadata templates and examples for users to facilitate compliance
 - Provide training for researchers on topics such as data sharing, reproducibility, interoperability, de-identification, data security
 - Develop criteria-based and actionable timeframes for accessibility and preservation of data, especially with regard to clinical practice guidelines
 - The Common Data Elements portal needs to be updated and made more user-friendly in order to be useful
 - The NIH needs to give more consideration to data repositories' business models as free service is not sustainable in the long run.
- III. Optimal Timing
 - Address if or how oversight and compliance will be provided in more detail
 - Mandate basic science research data sharing first and later phase in human subjects data

Detailed Comment for the “Proposed Provisions for a Draft NIH Data Management and Sharing Policy”:

I. [The definition of Scientific Data](#)

When discussing “Scientific Data” resulting from research sponsored by the NIH as defined in the proposed provisions, we suggest a further distinction to distinguish between basic research data and human-subjects data (e.g., biospecimens, private information). These two types of data require different data management strategies and sharing policies. One missing aspect of the Scientific Data definition is whether this policy applies to raw data and processed data or processed data only. In an ideal world, it would apply to both. However, there may be limitations such as the storage size or the need to integrate the raw data with calibration parameters for them to be suitable for further analysis. The policy should be clear in which cases raw data should be shared and in which they may be exempt.

We welcome the emphasis on metadata to ensure data interoperability and re-usability. Ideally, the NIH will develop templates and examples to help researchers comply with its new requirements. Those can be developed *de novo* in collaboration with NLM and other interested parties or built on ongoing efforts, such as the [Data Curation Network primers](#). We also would like to note that code and scripts can be seen as a fusion of analysis procedures and metadata documenting the analysis and should be addressed in this section.

II. [The requirements for Data Management and Sharing Plans](#)

General Comments:

Given that NIH and FDA can have overlapping projects, we suggest that the current Policy is also discussed between the two agencies to ensure consistency.

Related to Section IV, *Plan Elements*, we find the two pages proposed for Data Management Plans are not enough to sufficiently cover the seven elements requested for the plan. This page limit was established for early data management plan guidelines at other agencies and is not sufficient for a carefully considered plan. The policy should also expand its emphasis from FAIR to include reproducibility (in other words, making data findable, accessible, interoperable, reusable and reproducible). We also encourage the NIH to develop training for researchers to make them more comfortable with sharing data when appropriate. For example, training in de-identification of data for sharing (especially in regard to human subjects data) would be an essential component for a responsible and effective application of the Policy.

There are several sections, in which more guidance and support will accelerate researchers' understanding and ability to comply.

- Provide more guidance on what security is needed for particular types of data. Additional training to researchers and supporting staff would enable them to make the right decisions.
- Provide an actionable timeframe for how long data have to be accessible. Many existing policies are too vague, and keeping **all** data, or even all processed data, FAIR **indefinitely** is unrealistic, unless the federal government is prepared to store and preserve all of said data indefinitely. We surmise that more research is needed in this area and encourage the NIH to task a workforce to establish guidelines based on real-life need. We imagine those guidelines may be based on data uniqueness (research around establishing a screening practice) or data collection costs or data collection time investment (a threshold amount of cost or length of time would be useful guidelines). These guidelines can also include recommendations about decommissioning support for a dataset for research purposes--for example, if a higher-quality dataset is available (according to pre-defined quality metrics). We foresee that there needs to be an overlap between the two datasets to confirm that the older one is replaceable by the newer one. We also recognize that the older dataset may have historical value, but acknowledge the NIH Data Management and Sharing Policy needs to focus on research utility. This distinction should be made explicit and documented.
- Evaluating plans for external grants requires substantial expertise that is not part of the training for scientists. Consideration should be given about who would review those data management and sharing plans and make constructive revisions. We recommend training by NIH experts for reviewers in terms of what would be desirable to see in a data management plan. This type of support in the beginning would help the smooth implementation of the policy.
- We recommend including language in the policy that encourages researchers to explore more options for sharing as they develop project submissions to their IRBs.
- It should be specified that the amount of data requested in the Plan should be reported in orders of magnitude.

- When addressing “Related Tools, Software and/or Code”, emphasis is needed on the documentation of the analysis process (for example, in a readme file). We support language that encourages and promotes using open source options, but we also recognize that there may be barriers to fast adoption of these options, such as interface user-friendliness, context provision, and customer support.
- As noted earlier, we strongly support the emphasis on good metadata. We find the Common Data Elements to be a great idea, but the Portal is confusing. We suggest directing researchers to the repository <https://cde.nlm.nih.gov/cde/search>. We also recommend engaging more researchers to review the repository and make suggestions to improve self-navigation and make the discovery of relevant standards more user-friendly. Tutorials and educational materials will make it easier for scientists and the staff who support them in managing their data to select the right elements.
- The “Data Preservation and Access” section of the plan (Section IV.4) combines requirements about storage of active data (subsection 4.3) and true preservation and access after the project is completed. We suggest combining the storage requirements with the security requirements for active data, and reserving this section for addressing the handling of completed datasets. If subsection 4.3 refers only to cases in which the shared final dataset is going to be hosted in a unique repository created and maintained by the researcher, that should be stated explicitly. We also believe that “scientific data generated from humans or human biospecimens” is a very broad term. It requires a better definition. For example, both data from established human cell lines and from individuals would fall under this category, when the two require separate guidelines on how they need to be handled.
- In the section “Data Preservation and Access Timeline” (Section IV.5), the Policy should provide some guidelines on how long the different data should be kept. Keeping all data indefinitely is unrealistic, not only because of the cost but also because digital environments are still untested by time, and most repositories are relatively young. We strongly suggest reflecting on the duration of time during which data will need to be accessible and interoperable for reuse and reproducibility. The factors we suggest for this decision are the uniqueness, cost and duration requirements around data collection. Data that support current clinical guidelines or practice revisions should remain accessible until at least two subsequent revisions to the applicable guideline or practice have been made. We also suggest that the Policy sets expectations that the data should be available at time of publication if there are no restrictions on data usage. Subsection 5.2 requires further clarification. Does it refer to data that may not be shared publicly and require special permission for reuse? Right now this section reads as if any data resulting from an NIH project is subject for reuse approval by the researchers that produce them, and our understanding is that this is not the intention of the Policy.
- In the section “Data Sharing Agreements, Licensing, and Intellectual Property” (Section IV.6) we recommend stronger encouragement for sharing. Support protecting the data while filing for a patent, but then encourage sharing after the application succeeded (restricted data concerns fall under a different category).
- We have significant concerns about the language in “Oversight of Data Management” (Section IV.7) regarding repositories available to researchers at “no cost for extended

periods of use.” We want to emphasize that responsible data management and preservation has costs both in maintaining infrastructure and in properly-trained staff. Repositories need business models to be sustainable. As part of their license agreements “free” repositories can start charging fees at any time. We have seen that happen with many “free” software options. Other repositories operate on grants and endowment, and their fate is unknown once the grant ends. Responsible data management requires investment. The existing NIH repositories are great options, but with the expansion on data sharing that the Policy postulates, there need to be either more NIH repositories (including staff and resources to support such repositories), or the NIH should work with research universities and other entities to support this function beyond one-off grants. Another concern we want to raise is that even if the grant allows the researcher funds for data management, preservation often takes place and always continues after the project is complete and grant funding is closed. This contradiction needs to be addressed in cases where a long-term funding source like an endowment or recurring budget is not available.

- We recommend highlighting the importance of documenting important decisions such as “how decisions will be made to stop storing the scientific data or change its level of accessibility”
- We suggest being more specific in the “Compliance and Enforcement” section (Section V). For example, clarify
 - Whether the researchers need to provide DOIs or other evidence of compliance in their RPPRs.
 - Whether funding will be withheld for noncompliance.
 - How compliance will be enforced after the grant has ended.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards.

We suggest that the Policy can be applied to data from basic research sponsored by the NIH after a six-month outreach period on the finalized text. In order to apply it to the broader spectrum of data (including human-subjects related data), we suggest first creating educational materials for researchers on metadata elements and on responsible and effective data sharing. We also propose introducing temporary time recommendations (i.e. preserving access to data for 7-10 years after the project completion) while conducting research on what the optimal timeline would be for different types of data.

Submission #89**Date:** 12/07/2018**Name:** Wendy D. Streitz**Name of Organization:** University of California**Type of Organization:** University**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

The University of California believes that the curation and sharing of research data offers benefits to the larger research community and advances public knowledge. UC supports NIH's effort to facilitate data sharing and appreciates NIH's recognition of the challenges that come with regulating data generated by a broad research community. While UC generally agrees with NIH's draft data management and sharing policy, we ask that NIH consider the comments below, particularly around potentially confusing and administratively burdensome requirements, before considering implementation strategies.

I. The definition of Scientific Data

The University of California suggests that NIH adopt the Uniform Guidance definition of research data. Uniform Guidance at § 200.315(e)(3) defines research data as:

(3) Research data means the recorded factual material commonly accepted in the scientific community as necessary to validate research findings, but not any of the following: preliminary analyses, drafts of scientific papers, plans for future research, peer reviews, or communications with colleagues. This "recorded" material excludes physical objects (e.g., laboratory samples). Research data also do not include:

(i) Trade secrets, commercial information, materials necessary to be held confidential by a researcher until they are published, or similar information which is protected under law; and (ii) Personnel and medical information and similar information the disclosure of which would constitute a clearly unwarranted invasion of personal privacy, such as information that could be used to identify a particular person in a research study.

Uniform Guidance is the culmination of a two-year effort by the federal government to harmonize and streamline administration of federal grants and cooperative agreements. This effort was beneficial to the university community because it standardized administrative requirements, cost principles and audit requirements for federal awards. In addition to addressing the goal of a more efficient, effective and transparent government, the Uniform

Guidance also sought to reduce waste, abuse and burden in the administration of grants and other federal assistance awards by promoting consistency across federal agencies. As adoption of the Uniform Guidance illustrates, multiple definitions of research or scientific data not only cause confusion, but create undue administrative burden by requiring consultation and analysis under two separate and distinct processes for essentially the same output. UC recommends that NIH work with other agencies on a consistent data management policy for research data rather than have a piecemeal approach from different federal agencies.

It is also worth noting that the definition of scientific data as proposed by NIH expands beyond the Uniform Guidance definition by stating that scientific data includes recorded information that is “necessary to validate and replicate research findings” [emphasis added]. However, the NIH definition also excludes laboratory notebooks from scientific data. Laboratory notebooks are very likely to be essential in validating and replicating research findings. The Plan Element section 1.2 also references “any other information necessary to interpret the data” as a proposed element of a data management plan. Therefore, managing scientific data will include managing laboratory notebooks; aligning the definition of scientific data with the provided guidance will ensure proper management of scientific data and reduce confusion among the research community.

Lastly, NIH’s expectation that all scientific data be digitized raises concerns to UC. Not all scientific records must be digital to be useful. Moreover, some of the most worthy digitization efforts are of data potentially excluded from the current definition (e.g., 3D scans of rare physical specimens). Digitizing also imposes administrative burden upon institutions, and creates risk, both in loss or error in translation, as well as in of the disclosure or use of sensitive material, including medical information. If external parties are necessary to digitize such data, the risk of loss, error or exposure becomes more pronounced. Because of the potential for heightened risk, the decision to digitize data should be left to the principal investigator’s discretion. UC recommends that this expectation be removed.

II. The requirements for Data Management and Sharing Plans

The proposed requirements for the data management plan are extensive and will increase the amount of time principal investigators need to prepare a proposal. UC recommends that either: (1) the data management plan be required at a later point in the process (e.g., at the Just-In-Time request); or (2) if the plan is required at the time of proposal submission in order for reviewers to have confidence in the investigator’s commitment and ability to share data, the data management plan requirements be bifurcated into only those elements deemed critical for the proposal review with supplemental information provided either at the Just-In-Time request or at time of the Research Performance Progress Report.

Within the proposed plan review and evaluation criteria, the data management plan is proposed as an Additional Review Consideration for extramural grants. As an Additional Review Consideration, the data management plan would not be individually scored nor would it influence the overall score, although there is an expectation that compliance with the plan "would be integrated into terms and conditions as appropriate" and that NIH staff would engage with potential awardees to modify the plan as appropriate prior to award. Given the extent of the proposed data management plan requirements and that, as noted later in the draft policy, that "[f]ailure to comply with the award Terms and Conditions may result in an enforcement action, including additional special terms and

conditions or termination of award," it would appear that the effort and implications of the data management plan are consistent with their apparent value in the review process. UC recommends that if needed at time of proposal, these elements instead be aligned by positioning the data management plan as Additional Review Criteria, which would not be scored individually but would be considered in the overall impact score.

Maintaining restricted access to data should not be the responsibility of the individual researcher. Not only is this administratively burdensome, but it also introduces a dependence on the researcher (and their current contact information) that undermines the goal of long-term data accessibility. NIH should recommend restricted access repositories that provide this level of control for sharing data.

The "Compliance" section sends a strong message that NIH intends to enforce data management requirements. However, data management and sharing is difficult to fully anticipate in detail. Researchers require flexibility to update and change their data management plan as the project progresses. In practice, data management plans embedded in proposals may not be used (or fully used) for several years; repositories and data standards may also evolve over time. Complications may arise. We recommend that some degree of flexibility be built in, as technology and standards evolve quickly and plans may need to change over time to address unforeseen issues. UC recommends that NIH consider allowing researchers to make annual revisions (with explanation) to their data management plan. Such revisions could be included in the project's annual report.

UC appreciates the "Oversight of Data Management" section as an explicit component of the data management plan. This highlights that the management of research data is an active process, which requires the long-term investment of resources, and that these resources should be predetermined and incorporated into budgeting and planning prior to submission. However, UC notes two challenges: (1) funding is already very limited to support a researcher's project, and new "draws" on existing and available funding would have the inevitable consequence of reducing the amount available for the direct costs of science; and (2) some of the costs associated with data storage and sharing cannot reasonably be incurred within the period of the grant (or its closeout period). UC recommends that such oversight costs either be

separately funded by NIH or, at minimum, set aside, particularly if centrally-funded repositories that do not require deposit fees are not available.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

The University of California requests that before considering an implementation plan that NIH address these comments and others provided by the scientific community. The timing to implement will be difficult to assess prior to receiving responses to these comments.

In addition, UC would like to add the following comments on the proposed data sharing policy that did not fit into the allotted text boxes:

1. The scope of the proposed policy would apply to all intramural and extramural research, funded or supported in whole or in part by NIH that results in scientific data, regardless of NIH funding level or mechanism. This is a change from the current policy that applies to all investigator-initiated applications to NIH with direct costs greater than \$500,000. This imposes significant administrative burden on both researchers and research administrators. UC recommends that NIH maintain the original threshold dollar amount for this policy.

2. The “Scope and Requirements” section also states how “reasonable costs associated with data management and sharing” can be budgeted in proposals. The inclusion of new costs in proposals without an increase in expected funding is problematic for the reasons set forth above. In addition, if costs can be included in the award, the types of costs allowed and how researchers should handle costs that may need to be incurred after a grant has ended remains unclear. UC once again recommends that examples be provided as to the types of costs that may be included (e.g., data curation services, web hosting, personnel) and how these costs may be legitimately expensed. UC further recommends that NIH allow these costs over and above the existing modular grant ceiling.

Attachment:



OFFICE OF THE VICE PRESIDENT - RESEARCH AND GRADUATE STUDIES

OFFICE OF THE PRESIDENT
1111 Franklin Street, 11th Floor
Oakland, California 94607-5200

December 7, 2018

Carrie D. Wolinetz, Ph.D.
Office of Science Policy
National Institutes of Health
6705 Rockledge Drive, Suite 750
Rockville, MD 20892

RE: NOT-OD-19-014: Request for Information (RFI) on Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research

Dear Dr. Wolinetz:

I write on behalf of the University of California (UC) system with regard to the Request for Information (RFI) on Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research issued on October 10, 2018.

The UC system comprises ten research-intensive campuses, six medical schools, and three affiliated U.S. Department of Energy national laboratories. As a system, UC receives approximately \$5 billion annually of extramural awards to support research conducted throughout all UC locations. UC generally receives 5 to 6 percent of NIH's annual appropriations for research, making UC the largest single recipient of NIH funding for academic research.¹

The University of California believes that the curation and sharing of research data offers benefits to the larger research community and advances public knowledge. UC supports NIH's effort to facilitate data sharing and appreciates NIH's recognition of the challenges that come with regulating data generated by a broad research community. While UC generally agrees with NIH's draft data management and sharing policy, we ask that NIH consider the comments below, particularly around potentially confusing and administratively burdensome requirements, before considering implementation strategies.

I. The definition of Scientific Data

The University of California suggests that NIH adopt the Uniform Guidance definition of research data. Uniform Guidance at § 200.315(e)(3) defines research data as:

¹ University of California Office of the President. 2018 UC Accountability Report. Available: <https://accountability.universityofcalifornia.edu/2018/chapters/chapter-9.html>

(3) Research data means the recorded factual material commonly accepted in the scientific community as necessary to validate research findings, but not any of the following: preliminary analyses, drafts of scientific papers, plans for future research, peer reviews, or communications with colleagues. This “recorded” material excludes physical objects (e.g., laboratory samples). Research data also do not include:

*(i) Trade secrets, commercial information, materials necessary to be held confidential by a researcher until they are published, or similar information which is protected under law; and
(ii) Personnel and medical information and similar information the disclosure of which would constitute a clearly unwarranted invasion of personal privacy, such as information that could be used to identify a particular person in a research study.*

Uniform Guidance is the culmination of a two-year effort by the federal government to harmonize and streamline administration of federal grants and cooperative agreements. This effort was beneficial to the university community because it standardized administrative requirements, cost principles and audit requirements for federal awards. In addition to addressing the goal of a more efficient, effective and transparent government, the Uniform Guidance also sought to reduce waste, abuse and burden in the administration of grants and other federal assistance awards by promoting consistency across federal agencies. As adoption of the Uniform Guidance illustrates, multiple definitions of research or scientific data not only cause confusion, but create undue administrative burden by requiring consultation and analysis under two separate and distinct processes for essentially the same output. UC recommends that NIH work with other agencies on a consistent data management policy for research data rather than have a piecemeal approach from different federal agencies.

It is also worth noting that the definition of scientific data as proposed by NIH expands beyond the Uniform Guidance definition by stating that scientific data includes recorded information that is “necessary to validate *and replicate* research findings” [emphasis added]. However, the NIH definition also excludes laboratory notebooks from scientific data. Laboratory notebooks are very likely to be essential in validating and replicating research findings. The Plan Element section 1.2 also references “any other information necessary to interpret the data” as a proposed element of a data management plan. Therefore, managing scientific data will include managing laboratory notebooks; aligning the definition of scientific data with the provided guidance will ensure proper management of scientific data and reduce confusion among the research community.

Lastly, NIH’s expectation that all scientific data be digitized raises concerns to UC. Not all scientific records must be digital to be useful. Moreover, some of the most worthy digitization efforts are of data potentially excluded from the current definition (e.g., 3D scans of rare physical specimens). Digitizing also imposes administrative burden upon institutions, and creates risk, both in loss or error in translation, as well as in of the disclosure or use of sensitive material, including medical information. If external parties are necessary to digitize such data, the risk of loss, error or exposure becomes more pronounced. Because of the potential for heightened risk, the decision to digitize data should be left to the principal investigator’s discretion. UC recommends that this expectation be removed.

II. Requirements for Data Management and Sharing Plans

The proposed requirements for the data management plan are extensive and will increase the amount of time principal investigators need to prepare a proposal. UC recommends that either: (1) the data management plan be required at a later point in the process (e.g., at the Just-In-Time request); or (2) if the plan is required at the time of proposal submission in order for reviewers to have confidence in the investigator's commitment and ability to share data, the data management plan requirements be bifurcated into only those elements deemed critical for the proposal review with supplemental information provided either at the Just-In-Time request or at time of the Research Performance Progress Report.

Within the proposed plan review and evaluation criteria, the data management plan is proposed as an Additional Review Consideration for extramural grants. As an Additional Review Consideration, the data management plan would not be individually scored nor would it influence the overall score, although there is an expectation that compliance with the plan "would be integrated into terms and conditions as appropriate" and that NIH staff would engage with potential awardees to modify the plan as appropriate prior to award. Given the extent of the proposed data management plan requirements and that, as noted later in the draft policy, that "[f]ailure to comply with the award Terms and Conditions may result in an enforcement action, including additional special terms and conditions or termination of award," it would appear that the effort and implications of the data management plan are consistent with their apparent value in the review process. **UC recommends that if needed at time of proposal, these elements instead be aligned by positioning the data management plan as Additional Review Criteria, which would not be scored individually but would be considered in the overall impact score.**

Maintaining restricted access to data should not be the responsibility of the individual researcher. Not only is this administratively burdensome, but it also introduces a dependence on the researcher (and their current contact information) that undermines the goal of long-term data accessibility. NIH should recommend restricted access repositories that provide this level of control for sharing data.

The "Compliance" section sends a strong message that NIH intends to enforce data management requirements. However, data management and sharing is difficult to fully anticipate in detail. Researchers require flexibility to update and change their data management plan as the project progresses. In practice, data management plans embedded in proposals may not be used (or fully used) for several years; repositories and data standards may also evolve over time. Complications may arise. We recommend that some degree of flexibility be built in, as technology and standards evolve quickly and plans may need to change over time to address unforeseen issues. **UC recommends that NIH consider allowing researchers to make annual revisions (with explanation) to their data management plan. Such revisions could be included in the project's annual report.**

UC appreciates the "Oversight of Data Management" section as an explicit component of the data management plan. This highlights that the management of research data is an active process, which requires the long-term investment of resources, and that these resources should be

predetermined and incorporated into budgeting and planning prior to submission. However, UC notes two challenges: (1) funding is already very limited to support a researcher's project, and new "draws" on existing and available funding would have the inevitable consequence of reducing the amount available for the direct costs of science; and (2) some of the costs associated with data storage and sharing cannot reasonably be incurred within the period of the grant (or its closeout period). **UC recommends that such oversight costs either be separately funded by NIH or, at minimum, set aside, particularly if centrally-funded repositories that do not require deposit fees are not available.**

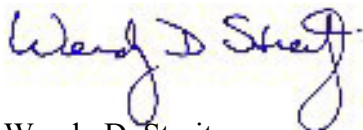
III. Timing for NIH to implement

The University of California requests that before considering an implementation plan that NIH address these comments and others provided by the scientific community. The timing to implement will be difficult to assess prior to receiving responses to these comments.

In addition, UC would like to add the following comments on the proposed data sharing policy that did not fit into the allotted text boxes:

1. The scope of the proposed policy would apply to all intramural and extramural research, funded or supported in whole or in part by NIH that results in scientific data, regardless of NIH funding level or mechanism. This is a change from the current policy that applies to all investigator-initiated applications to NIH with direct costs greater than \$500,000. This imposes significant administrative burden on both researchers and research administrators. **UC recommends that NIH maintain the original threshold dollar amount for this policy.**
2. The "Scope and Requirements" section also states how "reasonable costs associated with data management and sharing" can be budgeted in proposals. The inclusion of new costs in proposals without an increase in expected funding is problematic for the reasons set forth above. In addition, if costs can be included in the award, the types of costs allowed and how researchers should handle costs that may need to be incurred after a grant has ended remains unclear. **UC once again recommends that examples be provided as to the types of costs that may be included (e.g., data curation services, web hosting, personnel) and how these costs may be legitimately expensed. UC further recommends that NIH allow these costs over and above the existing modular grant ceiling.**

Sincerely,



Wendy D. Streit
Executive Director
Research Policy Analysis & Coordination
Office of Research & Graduate Studies

Submission #90**Date:** 12/09/2018**Name:** Mark Musen**Name of Organization:** Stanford University**Type of Organization:** University**Role:** Scientific Researcher**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Biomedical informations, data science

I. The definition of Scientific Data

The NIH should take a permissive approach regarding the definition of scientific data. It is often difficult to know in advance which datasets will have the most value to future generations of researchers. Investigators need to be encouraged to view the primary product of their research as their datasets, rather than their publications, and to manage those datasets with care. The draft NIH policy, which would require investigators to make an explicit, enforceable commitment to data sharing, would be an important step in this direction and could have a transformative effect on how the nation engages in science.

II. The requirements for Data Management and Sharing Plans

A significant limitation of most data-management plans is an absence of consideration of the standards that will be used to represent experiment-related metadata—both to specify the structure of the metadata and to define the value sets that will be used to provide standardized content for metadata fields. The “FAIRness” of scientific data is a function both of the richness of the corresponding metadata (which allows other investigators to understand what experiment actually was performed) and of the standardization of the metadata (which allows for effective data search, retrieval, and integration). Data-management plans need to speak directly to these metadata concerns if the output of scientific research is to be sharable and searchable.

The NIH should closely follow a pilot program that has been initiated by the Health Research Board of Ireland (HRB) and the Netherlands Organization for Health Research and Development (ZonMw). These funders, working with the GO FAIR initiative and the Research Data Alliance, are planning to take a proactive stance to ensure adequate data stewardship from the time that

the agencies first issue RFAs for new scientific programs. New RFAs will include links to templates that will specify the minimal metadata needed to ensure adequate data annotation for the kinds of data that the sponsors anticipate that investigators will collect. The templates, whenever possible, will reflect community-based metadata standards, and the templates will designate standard ontologies and value sets with which the fields of the templates should be filled in.

It will be straightforward for investigators to comply with the metadata requirements specified in the RFAs, as the RFAs will point to metadata templates created using The CEDAR Workbench—a Web-based tool developed by the Center for Expanded Data Annotation and Retrieval under the NIH BD2K program. CEDAR allows users to create libraries of structured templates for defining metadata, to associate the fields of those templates with standardized terms from ontologies, value sets, and common data elements, and to fill in the fields of a metadata template with standard values to enhance the searchability and interpretability of the metadata.

In the HRB/ZonMw pilot, the same electronic metadata templates to which the sponsors will have linked in the original RFAs will be used by the funded investigators to ensure that their datasets are annotated completely and in a standardized fashion. Thus, the funding organizations will use CEDAR to create templates that define the kinds of metadata that they expect to enable comprehensive description of datasets and good data sharing, and investigators will use CEDAR to fill in those templates, assuring compliance with the sponsors' expectations.

The NIH Strategic Plan for Data Science emphasizes the importance of making data FAIR. The Strategic Plan, however, does not discuss the data-management steps needed to achieve FAIRness. The Health Research Board of Ireland and the Netherlands Organization for Health Research and Development have asserted that FAIRness lies in the quality of experimental metadata, and they have committed to ensuring that their grantees will create high-quality metadata by building those expectations directly into their RFAs using CEDAR. The NIH could easily commit to a similar course of action.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

Implementation of a new data management and sharing policy should not wait for new infrastructure, resources, and standards to appear. The NIH has already made substantial investments in necessary infrastructure such as CEDAR and the BioPortal ontology repository, and it would do well to sustain the infrastructure to which it has already made significant commitments. Developing a new data management and sharing policy certainly would

stimulate work to create enhanced infrastructure, resources, and standards. All these activities would be self-reinforcing, and they should not be staged sequentially.

Submission #91**Date:** 12/08/2018**Name:** Mara Blake on behalf of JHU Data Services**Name of Organization:** Johns Hopkins University**Type of Organization:** University**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

We have strong research in many health and medical topic areas.

I. The definition of Scientific Data

In general, anything that is generated from a study and can help others reuse or reproduce this study or results should be counted as scientific data. This includes things that have traditionally been called “documentation.” Specifically, the definition of scientific data should include:

Scientific data should be primarily raw data, unless they contain PHI information. Raw data are usually the best product to share and to be reused by others. Aggregated or summarized tables have little use for others.

Experimental procedures should be included in scientific data. Some procedures are important for others to reproduce a study. Especially for fields that require innovated experimental designs or precise experimental procedures.

The criteria to exclude data points/subjects should be shared, if applicable. Investigators need to explain why they choose to exclude some data points.

Statistical analysis, analysis scripts, and relevant code should be included as well. This is important because others can use the shared data and scripts to re-run the same experiment.

NIH should provide a definition of de-identified data, and if possible, provide resources and training to help investigators properly de-identify human subject data.

A minimum amount of metadata, such as variable names, code books, should be included as scientific data. If there are standards in certain research fields, researchers should apply such standards in their data.

II. The requirements for Data Management and Sharing Plans

We anticipate that investigators will want clarification on the thresholds for "perceived barriers to sharing scientific data." Supplementary guidelines should specify those barriers with a range of examples. These barriers should be balanced by also providing potential solutions, such as finding representative samples to share.

JHU Data Services supports the DMPTool.org as resources for Investigators to write data management plans. We suggest that NIH craft guidance for researchers that fits in the DMPTool format and use it to disseminate the information. We also recommend that NIH, in coordination with grantee institutions, involve local research data services when possible in this process when provided by their libraries or research administration as we can review and provide feedback on drafts to investigators.

We appreciate the emphasis on data security in the policy, given the enhanced needs and consequences for protecting health data. We have found that researchers can find it challenging to both plan and summarize their methods for secure storage and access of data. JHU, however, is among those institutions providing more centralized resources for secure access to sensitive research data. JHU is encouraging use of such facilities as part of IRB and other compliance policies. NIH should encourage institutions, departments, or centers that have standardized secure resources to develop template summary descriptions to include with the Data Management and Sharing Plans.

Extramural grant reviewers will evaluate it as "acceptable or unacceptable by reviewers, but not be factored into the overall impact score through the peer review process. This allows for NIH staff to work with potential awardees to ensure that any reviewer concerns regarding the Plan could be addressed for meritorious applications as a contingency of NIH funding. Plan compliance would be integrated into terms and conditions as appropriate."

The policy should emphasize the overall value and academic credit for making funded data a resource for re-use when relevant and feasible. Investigators should also be encouraged to plan reasonably for scope and costs of data access to be appropriately reflected in the budget.

We recommend that internal IC evaluators and peer reviewers be given written instructions and possibly a brief online tutorial on how to review Plans effectively and efficiently. For example, JHU Data Services developed a content checklist for NSF grant reviewers to evaluate DMP content (<https://doi.org/10.17605/OSF.IO/SNYFB>)

NIH might consider a range between 2 to 4 pages, acknowledging that reviewers may be less inclined to read them as thoroughly. We do find that grant requirements for plans may be an investigator's first efforts at documenting all these topics in one place and can be a useful starting point that they may be encouraged to expand upon when grants are awarded.

1. Data Type:

NIH should clarify if this addresses what data will be produced, preserved, and potentially shared at the end of the project or should investigators list data products through the workflow and managed during the project, including secondary source data if relevant? We suggest providing guidance to investigators on how to describe their data type, as well as clarify if this refers to the state data at the time it is shared. We have used JHU Data Services guidelines to provide this information.

The term "metadata" appears in the definitions but not in the section guides. In our experience, investigators can find it challenging to address. Focusing on describing documentation may be preferable.

We find that many investigators are not familiar with the requirements and resource commitments of preparing data for broader access with adequate disclosure protection. In particular, some may propose "de-identifying" data when intending actions that would fall at, or below, HIPAA's "Safe Harbor" limited dataset, rather than their "expert determination" criteria for de-identification. We recommend supplementary guidance for investigators to determine what level of disclosure protection is feasible. Institutions such as JHU also have compliance policies and preferences, and could be encouraged to provide guidelines and template language for investigators' Plans. For Plans that require more detail in describing disclosure protection for public access, the guidelines could recommend describing them in a separate section, rather than in separate locations under Data Type and Data Access.

2. Related Tools, Software and/or Code:

We find that investigators may not readily know alternative open source software for their applications. NIH, possibly via the NLM, might consider hosting lists or links to resources for common alternative software as well as open access formats.

3. Standards:

Investigators writing NSF DMPs have found the Standards section challenging particularly in identifying its scope. Similarly, NIH's Standards section asks for many potentially detailed components. It is ambiguous about preferences for an investigator to outline their plans for documenting internal protocols and procedures in a standardized consistent way (e.g., having a file-naming convention among collaborators) or emphasizing compliance with community standards. Use of CDEs is an example of the latter, but specialties can vary widely in their development of standards. Similarly, "including terms of use" implies listing licensed proprietary instruments such as the SF-36 mental health scale or Morisky Medication Adherence Scale. Also, should use of standardized metadata be specified if relevant, or discussed in section 1.2 instead? Should investigators emphasize standards and metadata that facilitate access, discovery, and reuse of shared data? Additional guidance may be needed for this section.

4. Data Preservation and Access:

Supplementary guidelines should summarize what constitutes consistency.

We note that the policy does not indicate preference for using NIH-supported repositories when they are appropriate for a study. Would a particular grant solicitation set any such required or recommended repositories?

The policy might also choose to make more explicit whether NIH prefers investigators use managed repositories when available, whether those specific to a field or provided by their institutions. Are repositories preferred in particular to sharing data directly when requested by approved researchers? Investigators can find the choice challenging especially if the Policy is ambiguous about repository use regarding the merit of the proposal.

An investigator's method and resources for long-term preservation sometimes requires disambiguation from a data repository used for data access. It would be helpful if NIH defined the terms "archived" and "long-term preservation". We often find the researchers have a hard time distinguishing these terms from their personal data retention, the retention period of the repository they share data through, and any institutional or governmental retention policies that may apply.

This section might make more explicit whether it is referring to just the security of preserved and archived data from completed projects or might also discuss security during the management of the data, and PII/PHI in particular, during the project. As discussed, the investigators' institutions are eager that overall security is adequately planned. NIH might consider making this topic a separate section to centralize focus on data security throughout the research project.

5. Data Preservation and Access T

More clarity should be provided about the differences for approval for access and approval for re-use. This might be more efficiently discussed as part of the terms of use, perhaps even moving section 5 after section 6, because overall timing may be contingent on a range of policies and compliance criteria from the grant, data repository, associated g

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

The process for implementation should be scaffolded. Subjecting researchers to immediate requirements may lead to risks and reduced delivery of appropriate needs post funding awards. One recommendation is to use three phases, gradually increasing in stringency of requirements and evaluation. The timing of the phases might correspond with grant cycles - at least one cycle before to allow participants to seek out resources and adjust as needed. Prior to each phase,

NIH could release training, web content, and potentially technical infrastructure to supplement expressed needs. After each phase, it will be important to get feedback from institutions, professionals, researchers, and other stakeholders who are impacted in any way by these new requirements.

We would suggest three phases leading to full implementation 1) a trial period that makes the submissions voluntary or as listed for the Extramural Grants “Additional Review Consideration.” This would allow investigators to get feedback and for NIH reviewers to identify areas where further guidance would be helpful. Voluntary submissions might make good exemplars for training and web content. 2) Mandatory submission with a review and resubmit component specifically for those individuals new to writing the plans or to provide some courses or webinars to individuals as they prepare. 3) Full implementation.

Before going to the full implementation, NIH should provide guidance of the available resources and issues for researchers with sharing information that includes PHI and PII. While JHU is fortunate to provide resources and consultation on the topic, formal guidance and standards are lacking on this topic. Institutions will feel nervous and skeptical about approval for researchers to share these data as will the researchers who may not have a level of support to consult. While there are abundant resources available for constructing data management and sharing plans for data that is not sensitive, there is substantially less resources at institutions that is devoted to secure storage, collaboration, and computing processes including de-identification practices - especially as it concerns large – scale and High-performance research operations. These types of resources will need to be taken into consideration to support institutions in guiding their researchers in some of these measures.

Attachment:

Introduction

Since 2011, Johns Hopkins Data Services has provided direct support for Investigators preparing data management and sharing plans for grant proposals across various disciplines and funders, including NIH Data Sharing statements. In addition, they established the JHU Data Archive to accommodate open access data sharing for fields without other data repository options. JHU Data Services also provides support for best practices in research data management more broadly. Our response below draws upon those experiences.

We identified four key themes from preparing this RFI response:

- A definition of scientific data should encompass anything that is generated from a study and can help others reuse or reproduce this study or results.
- Researchers and institutions need more guidance on what and how data with protected health information (PHI) can be shared, including how to safely de-identify data.
- NIH needs to roll out dmp requirements because the methods for securing, collaborating, and sharing data with PHI are not clear and the appropriate infrastructure is not available to everyone or does not exist.
- Units that support research, such as us, play a key role as conduits of information between NIH, the researchers, and the institution.

The definition of Scientific Data

In general, anything that is generated from a study and can help others reuse or reproduce this study or results should be counted as scientific data. This includes things that have traditionally been called "documentation." Specifically, the definition of scientific data should include:

- Scientific data should be primarily raw data, unless they contain PHI information. Raw data are usually the best product to share and to be reused by others. Aggregated or summarized tables have little use for others.
- Experimental procedures should be included in scientific data. Some procedures are important for others to reproduce a study. Especially for fields that require innovated experimental designs or precise experimental procedures.
- The criteria to exclude data points/subjects should be shared, if applicable. Investigators need to explain why they choose to exclude some data points.
- Statistical analysis, analysis scripts, and relevant code should be included as well. This is important because others can use the shared data and scripts to re-run the same experiment.
- NIH should provide a definition of de-identified data, and if possible, provide resources and training to help investigators properly de-identify human subject data.
- A minimum amount of metadata, such as variable names, code books, should be included as scientific data. If there are standards in certain research fields, researchers should apply such standards in their data.

The requirements for Data Management and Sharing Plans

- We anticipate that investigators will want clarification on the thresholds for "perceived barriers to sharing scientific data." Supplementary guidelines should specify those barriers with a range of examples. These barriers should be balanced by also providing potential solutions, such as finding representative samples to share.

- JHU Data Services supports the DMPTool.org as resources for Investigators to write data management plans. We suggest that NIH craft guidance for researchers that fits in the DMPTool format and use it to disseminate the information. We also recommend that NIH, in coordination with grantee institutions, involve local research data services when possible in this process when provided by their libraries or research administration as we can review and provide feedback on drafts to investigators.
- We appreciate the emphasis on data security in the policy, given the enhanced needs and consequences for protecting health data. We have found that researchers can find it challenging to both plan and summarize their methods for secure storage and access of data. JHU, however, is among those institutions providing more centralized resources for secure access to sensitive research data. JHU is encouraging use of such facilities as part of IRB and other compliance policies. NIH should encourage institutions, departments, or centers that have standardized secure resources to develop template summary descriptions to include with the Data Management and Sharing Plans.
- Extramural grant reviewers will evaluate it as "acceptable or unacceptable by reviewers, but not be factored into the overall impact score through the peer review process. This allows for NIH staff to work with potential awardees to ensure that any reviewer concerns regarding the Plan could be addressed for meritorious applications as a contingency of NIH funding. Plan compliance would be integrated into terms and conditions as appropriate."
- The policy should emphasize the overall value and academic credit for making funded data a resource for re-use when relevant and feasible. Investigators should also be encouraged to plan reasonably for scope and costs of data access to be appropriately reflected in the budget.
- We recommend that internal IC evaluators and peer reviewers be given written instructions and possibly a brief online tutorial on how to review Plans effectively and efficiently. For example, JHU Data Services developed a content checklist for NSF grant reviewers to evaluate DMP content (<https://doi.org/10.17605/OSF.IO/SNYFB>)
- NIH might consider a range between 2 to 4 pages, acknowledging that reviewers may be less inclined to read them as thoroughly. We do find that grant requirements for plans may be an investigator's first efforts at documenting all these topics in one place and can be a useful starting point that they may be encouraged to expand upon when grants are awarded.

1. Data Type:

- NIH should clarify if this addresses what data will be produced, preserved, and potentially shared at the end of the project or should investigators list data products through the workflow and managed during the project, including secondary source data if relevant? We suggest providing guidance to investigators on how to describe their data type, as well as clarify if this refers to the state data at the time it is shared. We have used JHU Data Services guidelines to provide this information.
 - The term "metadata" appears in the definitions but not in the section guides. In our experience, investigators can find it challenging to address. Focusing on describing documentation may be preferable.
 - We find that many investigators are not familiar with the requirements and resource commitments of preparing data for broader access with adequate disclosure protection. In particular, some may propose "de-identifying" data when intending actions that would fall at, or below, HIPAA's "Safe Harbor" limited dataset, rather than their "expert determination" criteria for de-identification. We recommend supplementary guidance for investigators to determine what level of disclosure protection is feasible. Institutions such as JHU also have compliance policies and preferences, and could be encouraged to provide guidelines and template language for investigators' Plans. For Plans that require more detail in describing disclosure protection for public

access, the guidelines could recommend describing them in a separate section, rather than in separate locations under Data Type and Data Access.

2. Related Tools, Software and/or Code:

- We find that investigators may not readily know alternative open source software for their applications. NIH, possibly via the NLM, might consider hosting lists or links to resources for common alternative software as well as open access formats.

3. Standards:

- Investigators writing NSF DMPs have found the Standards section challenging particularly in identifying its scope. Similarly, NIH's Standards section asks for many potentially detailed components. It is ambiguous about preferences for an investigator to outline their plans for documenting internal protocols and procedures in a standardized consistent way (e.g., having a file-naming convention among collaborators) or emphasizing compliance with community standards. Use of CDEs is an example of the latter, but specialties can vary widely in their development of standards. Similarly, "including terms of use" implies listing licensed proprietary instruments such as the SF-36 mental health scale or Morisky Medication Adherence Scale. Also, should use of standardized metadata be specified if relevant, or discussed in section 1.2 instead? Should investigators emphasize standards and metadata that facilitate access, discovery, and reuse of shared data? Additional guidance may be needed for this section.

4. Data Preservation and Access:

- Supplementary guidelines should summarize what constitutes consistency.
- We note that the policy does not indicate preference for using NIH-supported repositories when they are appropriate for a study. Would a particular grant solicitation set any such required or recommended repositories?
- The policy might also choose to make more explicit whether NIH prefers investigators use managed repositories when available, whether those specific to a field or provided by their institutions. Are repositories preferred in particular to sharing data directly when requested by approved researchers? Investigators can find the choice challenging especially if the Policy is ambiguous about repository use regarding the merit of the proposal.
- An investigator's method and resources for long-term preservation sometimes requires disambiguation from a data repository used for data access. It would be helpful if NIH defined the terms "archived" and "long-term preservation". We often find the researchers have a hard time distinguishing these terms from their personal data retention, the retention period of the repository they share data through, and any institutional or governmental retention policies that may apply.
- This section might make more explicit whether it is referring to just the security of preserved and archived data from completed projects or might also discuss security during the management of the data, and PII/PHI in particular, during the project. As discussed, the investigators' institutions are eager that overall security is adequately planned. NIH might consider making this topic a separate section to centralize focus on data security throughout the research project.

5. Data Preservation and Access T

- More clarity should be provided about the differences for approval for access and approval for re-use. This might be more efficiently discussed as part of the terms of use, perhaps even moving section 5 after section 6, because overall timing may be contingent on a range of policies and compliance criteria from the grant, data repository, associated governance agencies (e.g. FDA), and the investigator's institution.

Additional Considerations for the Plan

- We support the making scientific data openly available whenever possible with no cost for those accessing the data. The JHU Data Archive is open access. Non-governmental, commercial, and

academic institutional repositories may charge for depositing data; this paragraph could reiterate that such costs might be included in the budget.

- More clarity on the "obligations" satisfied by supported institutions would be helpful, since most of the policy refers to investigator obligations, not institutional obligations. Is the intention that investigators may meet their policy obligations when they choose data repositories operated by their home institution, such as the JHU Data Archive, presuming repositories more appropriate to their field are not available? NIH should make clear how they will monitor and ensure investigators carry out their data management and sharing commitments and consider ways to automate the process.

The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

- The process for implementation should be scaffolded. Subjecting researchers to immediate requirements may lead to risks and reduced delivery of appropriate needs post funding awards. One recommendation is to use three phases, gradually increasing in stringency of requirements and evaluation. The timing of the phases might correspond with grant cycles - at least one cycle before to allow participants to seek out resources and adjust as needed. Prior to each phase, NIH could release training, web content, and potentially technical infrastructure to supplement expressed needs. After each phase, it will be important to get feedback from institutions, professionals, researchers, and other stakeholders who are impacted in any way by these new requirements.
- We would suggest three phases leading to full implementation 1) a trial period that makes the submissions voluntary or as listed for the Extramural Grants "Additional Review Consideration." This would allow investigators to get feedback and for NIH reviewers to identify areas where further guidance would be helpful. Voluntary submissions might make good exemplars for training and web content. 2) Mandatory submission with a review and resubmit component specifically for those individuals new to writing the plans or to provide some courses or webinars to individuals as they prepare. 3) Full implementation.
- Before going to the full implementation, NIH should provide guidance of the available resources and issues for researchers with sharing information that includes PHI and PII. While JHU is fortunate to provide resources and consultation on the topic, formal guidance and standards are lacking on this topic. Institutions will feel nervous and skeptical about approval for researchers to share these data as will the researchers who may not have a level of support to consult. While there are abundant resources available for constructing data management and sharing plans for data that is not sensitive, there is substantially less resources at institutions that is devoted to secure storage, collaboration, and computing processes including de-identification practices - especially as it concerns large - scale and High-performance research operations. These types of resources will need to be taken into consideration to support institutions in guiding their researchers in some of these measures.

Concluding thoughts

In conclusion we also want to reiterate the importance of connection and collaboration with the research data services at the institutions that are frequent grant recipients. The professional in these positions well poised to hold trainings, provide consultations, answer questions, and foster connections to technology at the institution. We also can disseminate knowledge about external service offerings and direct researchers to appropriate information when we are part of the chain of communication with NIH. Training from NIH for data management professionals like our team at JHU would allow us to share this essential information with investigators at our institution. We can

assist in both the process of creating data management and sharing plans and implementation of proposed processes throughout the life of the research projects to ensure that compliance is as efficient and effective as possible.

Submission #92**Date:** 12/09/2018**Name:** Ana Sanchez**Name of Organization:** Duke University**Type of Organization:** University**Role:** Scientific Researcher**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Duke University has a broad array of research. The comments submitted here were drafted by me and my team as the Advancing Scientific Integrity, Services and Training program. We develop educational tools to support research integrity, include assist with data management support.

I. The definition of Scientific Data

The definition for Data Management and Sharing Plan should be modified to provide clear expectations for data management, as it is not clear whether data management is inclusive of data collection, storage, analysis, etc. as currently written.

The definition of data sharing as it relates to FAIR data principles references accessibility with others. However, the FAIR principles also relate to ensuring data are FAIR to machines, not just humans. It may be that NIH is only focusing on FAIR data principles as they relate to access by humans, but that should be clarified, if so.

The definition of scientific data should address non-data research items, which are critical to the usability of data. These include methods and workflows (both experimental and analytical methods/workflows). It would be helpful to clarify if these items are expected to be shared according to FAIR principles.

The purpose statement (Section II) focuses on “requirements for responsible management and sharing of scientific data resulting from NIH funded or supported research.” However, the document focuses on requirements for the plan and does not explicitly state what the requirements for sharing are as currently written.

II. The requirements for Data Management and Sharing Plans

We have concerns about requirements for Data Management and Sharing Plans that can be broken into the following main topics:

Requirement: It is not clear if the requirement extends to training grants. It is also not clear how the requirements do or do not dovetail with data reporting on clintrials.gov.

Funding: The document does not clearly address plans for funding preservations and sharing beyond the life of the grant. Further, it does not define the anticipated lifespan for sharing/preserving. It is not clear if NIH funding is allowed to support adequate data security and compliance with privacy protections.

Required plan details within length: The document requests many attributes to be recorded regarding each anticipated piece of scientific data, related tools, software and/or code. Within the current page-limit, this would not be feasible given the number of assays that are present within complex grants (i.e., U or M mechanisms). Since the information requested is standardized, it would be helpful if NIH provided a template to support entry of this information.

Application of policy: With respect to Data Preservation and Access, the definition of discoverable needs to be more explicit. Is the intent to be synonymous with “Findable”? Furthermore, it is not clear within this section (Section 4.1-4.7) if the stipulations in these sections are applicable in instances when data cannot be shared.

With respect to Data Preservation and Access timeline, the NIH expectation for how long data should be preserved and shared is not clear. The policy notes that data should be shared for as long as it is useful for the scientific community. This is difficult to define, especially when resources may be required to maintain data in a repository in a way that adheres to FAIR data principles.

Within Section 7 (Oversight of Data Management), it is not clear what is intended by “data distribution” and whether there is any expectation for active data distribution efforts or if making data available via repositories is sufficient.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

Adoption of these policies will require extensive resources from a variety of specialties: data scientists, technological (IT) support and repository/database expertise. Institutions will need time to ensure these resources are widely available to effectively manage data as described in the policy.

We would recommend NIH focus efforts on data sharing from clinical trials or pre-clinical research as sharing those data likely has the biggest benefit to the community. An additional

focus on ensuring public access to published data and encouraging publication of negative data could support

Submission #93**Date:** 12/09/2018**Name:** Salvatore La Rosa**Name of Organization:** Children's Tumor Foundation**Type of Organization:** Nonprofit Research Organization**Role:** Scientific Researcher**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Neurofibromatosis, NF1, NF2, Schwannomatosis, clinical, genomics, cancer, cell biology, ophthalmology, neurology

I. The definition of Scientific Data

The definition of scientific data has to be broadened to not only the data resulting from NIH-funded or –supported research, but also to all the data (including background data and/or preliminary data) used to build the research proposal. This preliminary data has to be included in the definition of data and clearly marked as an essential piece of the data that has to come out alongside the result data of the study itself. For this reason, the awardee or grantee has to seek approval (to collaborators, if necessary) to share all the data used to build the study application, clearly mark this as the preliminary data, and make sure this is included in the output of the study.

II. The requirements for Data Management and Sharing Plans

We believe that until the requirements for data sharing will stay ‘recommended’ rather than ‘required’, the policy will remain weak in essence and will not be able to achieve the full aim of incentivizing data (re)usability and transparency that is needed to move the needle in the modern era of data digitalization. If data-sharing is still not ready to be an absolute requirement than it is of pivotal importance that NIH recognizes that sharing good quality data requires additional resources and needs to be incentivized with extra dollars outside of the traditional grant scheme. Having access to extra money in addition to the grant money would prompt the investigators to start looking for or develop those skillsets required to fulfill the need of data-sharing requirement and therefore access the extra money. So, one solution is to consider access to extra money (that can be obtained by a modest reduction of allowed direct costs in the principal grant) and allow a great deal of flexibility in the use of this. Internal or external collaborations with for-profit companies that are certified data-expert or institutions that have

the skillsets to fulfill all the requirements could allow principal investigators to use the additional resources and fulfill the requirements more easily.

With this preface, we also believe that the requirement of a data sharing plan is the first essential step.

Point 4. Data Preservation and Access.

The NIH should compile a list of approved or vetted data repositories as the preferred choice for researchers as well as a list of repositories that are outside of the NIH-funded or directed initiatives and lead a sort of consortium of 'approved' repositories that can be used. Each repository should seek for NIH-approved certification if they want to host data in accordance to the NIH-guidelines or be listed as accepted repositories.

We encourage NIH to move towards international data standards and work in collaboration with the large consortia already developing such standards.

We know that only a portion of the collected data will be made available at the end of the grant, we suggest asking for what data will be left out and an explanation of why it will not be reported or deemed important.

Point 5. Data preservation and Access Timeline

There is a lack of a specific timeline for data depositing. We recommend having a specific time period of max 12 months from the award end to have all the data listed in element 1 (Data Type) available and accessible. Data producers should have prolonged but not indefinite benefit from the use of data.

Point 6. Data sharing agreements, Licensing, and Intellectual Property

6.2 Explicitly recommend the use of the creative commons licensing type for the general terms to facilitate the identification of standards.

6.3 Rewrite this paragraph to re-enforce the concept that claiming IP from research outputs does not mean that data-sharing is somehow not required or can be avoided. If there are requirements for IP filing, these should be taken into account and plan accordingly not to delay the access to data. In this case, it would be useful to explain the mechanisms of IP filing and the fact that once the IP is filed, there is no need to wait for the patent to be public, but data has to be released soon after filing.

Additional comments:

In the grant evaluation phase, reviewers or the funding IC need to check on compliance with past or current awards and their respective data sharing plans when future funding or support decisions are sought. In this regard, a statement by the applicants that all requirements on

existing grants are in good standing order must be obtained with a minimum level of details on specific accomplished tasks.

In addition to the data sharing plan for the specific grant, a section of the data sharing plan should specifically ask for examples of data already shared by the applicant. Listing of DOIs for data or other links is important. An idea would be to add a new section to the NIH bio sketch for published datasets.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

We believe that the implementation of a requirement for data-sharing in all the grants is the only way to push data sharing and achieve results. If this is not possible, then incentives need to be rolled out first, like the possibility to access extra money for data sharing in order to achieve and make the data sharing plan realistic and achievable.

Similar to the thousands of option for scientific publication, NIH should invest or incentivize the creation of a similar ecosystem for data sharing with multiple options and the creation of multiple alternatives outside of the NIH field of operation. Also, NIH should seek compliance of existing repositories to certain standards of quality, privacy and reliability to be used as options by the investigators. This is a critical step in the adoption of the new requirements.

If the enforcement of a requirement for data-sharing is yet not possible, then the policy should be rolled out as a recommendation for sharing with a required detailed data sharing plan. Even if not required, it should be stated in the grant guideline that reference to already shared data for past or existing grants by the investigator will be considered in the overall scoring of the application.

Submission #94**Date:** 12/09/2018**Name:** Loic LE MARCHAND**Name of Organization:** University of Hawaii Cancer Center**Type of Organization:** University**Role:** Scientific Researcher**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Epidemiology

Cancer Research

Disparity Science

Minority Health

II. The requirements for Data Management and Sharing Plans

Our Experience with data sharing:

I write on behalf on the investigators of the Multiethnic Cohort Study, a prospective cohort study funded by NCI since 1992 that includes 215,000 individuals of five main ethnic/racial groups (Native Hawaiians, Japanese Americans, whites, African Americans and Latinos) (<http://www.uhcancercenter.org/mec-overview>). Our group has been very supportive of the principles of data sharing and we sincerely believe in its value to maximize the knowledge that can be gained from scientific data. We have embraced the opportunity to share the epidemiologic resource that we have developed over several decades with outside investigators (<http://www.uhcancercenter.org/mec-researchers/other-institutions-using-the-mec-resource>) and have, we believe, an excellent track record in this area. This view is shared by our External Advisory Committee and is reflected in the summary statements of our grants. We have committed considerable time and resources to developing a data sharing system (<http://www.uhcancercenter.org/mec-researchers/mec-data-sharing>) which does not differentiate between internal and external investigators. This on-line system allows us to manage and track applications, reviews, outcomes, IRB approvals, Data Use Agreements and data transfers while preserving the confidentiality and integrity of the data, and following all regulatory requirements. The experience that we have gained from a decade of data sharing has led to several observations:

1) Much effort is required to explain the design of the study and characteristics of the collected data. This is one reason that we assign a MEC investigator as a contact person to each project to help applicants determine whether our data/samples are appropriate to answer the research question, to assist them in their data request, and to ensure that users remain cognizant of the limitations of the data and study design when interpreting the findings.

2) The true value of an epidemiologic cohort is in annotating clinical and omics data with rich exposure data over part of the lifecourse. These data are complex and dynamic. A simple data dictionary is not sufficient for investigators to successfully use the data. Exposures are often measured by different variables, each having different strengths and limitations. Questions, questionnaires, and geospatial data capture exposures for different time periods and biospecimens were collected in different years than the exposure information. The data are not static; additions and corrections are made regularly to the database through verification and editing, and as a result of active and passive follow up. The validity of the results and their interpretation very much depends on the user's depth of knowledge of the data. New users always need considerable assistance from experienced users. Utilization of a reduced, "boiled down", form of the data would not make optimal use of the resource and may lead to the wrong results.

3) The cohort was assembled by sending a letter explaining the goal of the study and a questionnaire to residents of Hawaii and Los Angeles County. Return of a filled questionnaire was taken as consent to enter the study by UH and USC IRBs. Participants who provided a biospecimen completed a consent form at time of collection that stipulates the use of the samples, in broad terms, for understanding the environmental and genetic causes of chronic diseases, including in future research. However, neither of these consent processes mentioned broad data sharing in which ourselves and our IRBs would not have a role. Also, we share data according to the Standards for Privacy of Individually Identifiable Health Information (Privacy Rule) minimum necessary standard, where we provide a minimal set of variables required to address the research question. However, by playing an active role in the data sharing process and helping the applicants through the learning process, we have been able to successively share the data with many outside investigators. So far, our IRBs have only allowed limited sharing of genetic data in a public database (dbGAP).

Our cohort includes disadvantaged minorities, including an indigenous population, which have been harmed in the past by negligent researchers/institutions. We feel a strong responsibility about ensuring that the research using MEC data is ethical, scientifically valid, and socially acceptable in order to fulfill our commitment to the participants, their relatives and our communities.

Our concerns with the proposed changes:

We are concerned by the proposed changes in the NIH data sharing rules that would require the posting and archiving of a broad range of cohort data in a centralized public repository with controlled access.

1. We believe that misuse of the broad data, or even using a simplified data set, may lead to bad science and could harm some of the populations included in our study.
2. A distinction should be made between new cohort studies, that will be able to obtain written informed consent specifically for data to be widely shared on publicly available databases, and existing cohorts that began 10 to 40 years ago without this kind of consent. Requiring existing cohort studies to re-consent participants is not a feasible option. This would likely result in only a fraction of the cohort giving written consent, while some participants may withdraw altogether out of concern that their privacy will be jeopardized. To preserve continued active follow-up in the existing cohorts, it would be best to permit the cohorts to continue sharing through current methods, especially if they have not been shown to be inadequate.
3. Some of the studies using our data rely in part on data from Medicare, California Neighborhoods Data System, and other highly regulated data resources. These organizations and existing data use agreements do not allow the study to share the data with others except under very specified conditions.
4. We already have a proven process and systems in place for outside investigators to access the data. We strongly believe that outside investigators need to receive help from the study investigators in analyzing and interpreting the data, something that would not be possible under a system in which all data are stored on an NIH controlled access database.

Instead of creating one-size-fits-all rules, we suggest that NIH makes use of peer-review to determine whether a cohort adequately and sufficiently meets the requirement for data sharing with outside investigators. As far as we know, NCI or NIH has not provided evidence that the current data sharing systems do not allow the desired access to external investigators or that there are systemic or localized problems that should be fixed.

Attachment:

Comment on Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research

Our Experience with data sharing:

I write on behalf on the investigators of the Multiethnic Cohort Study, a prospective cohort study funded by NCI since 1992 that includes 215,000 individuals of five main ethnic/racial groups (Native Hawaiians, Japanese Americans, whites, African Americans and Latinos) (<http://www.uhcancercenter.org/mec-overview>). Our group has been very supportive of the principles of data sharing and we sincerely believe in its value to maximize the knowledge that can be gained from scientific data. We have embraced the opportunity to share the epidemiologic resource that we have developed over several decades with outside investigators (<http://www.uhcancercenter.org/mec-researchers/other-institutions-using-the-mec-resource>) and have, we believe, an excellent track record in this area. This view is shared by our External Advisory Committee and is reflected in the summary statements of our grants. We have committed considerable time and resources to developing a data sharing system (<http://www.uhcancercenter.org/mec-researchers/mec-data-sharing>) which does not differentiate between internal and external investigators. This on-line system allows us to manage and track applications, reviews, outcomes, IRB approvals, Data Use Agreements and data transfers while preserving the confidentiality and integrity of the data, and following all regulatory requirements. The experience that we have gained from a decade of data sharing has led to several observations:

- 1) Much effort is required to explain the design of the study and characteristics of the collected data. This is one reason that we assign a MEC investigator as a contact person to each project to help applicants determine whether our data/samples are appropriate to answer the research question, to assist them in their data request, and to ensure that users remain cognizant of the limitations of the data and study design when interpreting the findings.
- 2) The true value of an epidemiologic cohort is in annotating clinical and omics data with rich exposure data over part of the lifecourse. These data are complex and dynamic. A simple data dictionary is not sufficient for investigators to successfully use the data. Exposures are often measured by different variables, each having different strengths and limitations. Questions, questionnaires, and geospatial data capture exposures for different time periods and biospecimens were collected in different years than the exposure information. The data are not static; additions and corrections are made regularly to the database through verification and editing, and as a result of active and passive follow up. The validity of the results and their interpretation very much depends on the user's depth of knowledge of the data. New users always need considerable assistance from experienced users. Utilization of a reduced, "boiled down", form of the data would not make optimal use of the resource and may lead to the wrong results.
- 3) The cohort was assembled by sending a letter explaining the goal of the study and a questionnaire to residents of Hawaii and Los Angeles County. Return of a filled questionnaire was taken as consent to enter the study by UH and USC IRBs. Participants who provided a biospecimen completed a consent form at time of collection that stipulates the use of the samples, in broad terms, for understanding the

environmental and genetic causes of chronic diseases, including in future research. However, neither of these consent processes mentioned broad data sharing in which ourselves and our IRBs would not have a role. Also, we share data according to the Standards for Privacy of Individually Identifiable Health Information (Privacy Rule) minimum necessary standard, where we provide a minimal set of variables required to address the research question. However, by playing an active role in the data sharing process and helping the applicants through the learning process, we have been able to successively share the data with many outside investigators. So far, our IRBs have only allowed limited sharing of genetic data in a public database (dbGAP).

Our cohort includes disadvantaged minorities, including an indigenous population, which have been harmed in the past by negligent researchers/institutions. We feel a strong responsibility about ensuring that the research using MEC data is ethical, scientifically valid, and socially acceptable in order to fulfill our commitment to the participants, their relatives and our communities.

Our concerns with the proposed changes:

We are concerned by the proposed changes in the NIH data sharing rules that would require the posting and archiving of a broad range of cohort data in a centralized public repository with controlled access.

1. We believe that misuse of the broad data, or even using a simplified data set, may lead to bad science and could harm some of the populations included in our study.
2. A distinction should be made between new cohort studies, that will be able to obtain written informed consent specifically for data to be widely shared on publicly available databases, and existing cohorts that began 10 to 40 years ago without this kind of consent. Requiring existing cohort studies to re-consent participants is not a feasible option. This would likely result in only a fraction of the cohort giving written consent, while some participants may withdraw altogether out of concern that their privacy will be jeopardized. To preserve continued active follow-up in the existing cohorts, it would be best to permit the cohorts to continue sharing through current methods, especially if they have not been shown to be inadequate.
3. Some of the studies using our data rely in part on data from Medicare, California Neighborhoods Data System, and other highly regulated data resources. These organizations and existing data use agreements do not allow the study to share the data with others except under very specified conditions.
4. We already have a proven process and systems in place for outside investigators to access the data. We strongly believe that outside investigators need to receive help from the study investigators in analyzing and interpreting the data, something that would not be possible under a system in which all data are stored on an NIH controlled access database.

Instead of creating one-size-fits-all rules, we suggest that NIH makes use of peer-review to determine whether a cohort adequately and sufficiently meets the requirement for data sharing with outside

investigators. As far as we know, NCI or NIH has not provided evidence that the current data sharing systems do not allow the desired access to external investigators or that there are systemic or localized problems that should be fixed.

Sincerely,

Loïc Le Marchand, MD, PhD
Professor, Epidemiology
Associate Professor for Ethnic Diversity
University of Hawaii Cancer Center

Submission #95**Date:** 12/10/2018**Name:** Sandra Orchard (ISB Chair)**Name of Organization:** International SOciety for Biocuration**Type of Organization:** Other**Other Type of Organization:** Professional Body**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

The ISB is a non profit organization for biocurators, software developers, and researchers with an interest in biocuration. Biocuration involves the translation and integration of information relevant to biology and the biomedical sciences into a database or resource that enables integration of the scientific literature as well as large data sets. Accurate and comprehensive representation of biological knowledge, as well as easy access to this data for working scientists and a basis for computational analysis, are primary goals of biocuration.

II. The requirements for Data Management and Sharing Plans

Data management and sharing plans should include details of public domain repositories through which the data will be made available, the data standards which will be adopted when preparing the data for database deposition and a clear statement of accompanying information (meta-data) that will be included to ensure that the data is reusable by the scientific community. Interim/final reports should track the progress of such depositions and include citation of accession numbers when a deposition has been accepted by a public domain repository.

The NIH should publish lists of trusted core data repositories that are appropriate for each data type and are stably funded, to ensure long-term maintenance of this data. This should include the provision of biocurators who will ensure data meets the required standard. The Provisions should include a requirement that data be licensed under one of the recognized open data licenses at <https://opendefinition.org/licenses/>, or an explanation of why such a license cannot be adopted. All data and metadata necessary to reproduce analyses and results in publications should be released at the time of publication, for the purposes of both reproducibility and secondary analysis.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible

phasing could relate to needed improvements in data infrastructure, resources, and standards

Identifying, and actively supported, selected core resources will be critical for this initiative to have long-term sustainability. The NIH should publish a continually-updated list of compliant data repositories for each data type. NIH should also develop and publish specific criteria by which repositories can qualify as "NIH-compliant". Those criteria can address issues like data preservation plans, issuance of stable DOIs, API access, etc. NIH should be part of a global mechanism to financially support repositories that are not created/maintained by NIH itself.

For data types with no established data repository, grantees should be strongly encouraged or required to utilize a general purpose data repository (like Harvard Dataverse, Figshare, Data Dryad, GigaDB, etc.).

Submission #96**Date:** 12/10/2018**Name:** Alex Bateman**Name of Organization:** The UniProt Consortium**Type of Organization:** Other**Other Type of Organization:** Other**Role:** Scientific Researcher**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Data resource provision

I. The definition of Scientific Data

The current definitions of scientific data are not clear about whether the knowledge created by research teams and published within the scientific literature are considered as data. While we note that the current definitions are focussed on capturing key data types for submission to data depositories, we realise that the final goal is to capture knowledge. Therefore, the NIH data sharing and management plans should take steps to include how researchers will enable the capture or researcher submission of knowledge into knowledgebases.

II. The requirements for Data Management and Sharing Plans

In the future, we suggest that data management plans explicitly consider how knowledge will be transferred to the relevant knowledgebases in the field. This would be part of data preservation and access as described in section 4.2 of the proposed Provisions for a Draft NIH Data Management and Sharing Policy. We envision two main routes for this (i) through active submission or (ii) through steps to enhance the capture of knowledge from the literature.

Active submission

Most knowledgebases have routes to submit feedback on specific entries so that new knowledge can be incorporated into these entries. Some resources even allow direct editing of the knowledge by researchers following validation (e.g. WikiPathways).

Enhancing capture of knowledge

Many journals have guidelines for how to include accession identifiers in relevant resources when they are mentioned in the text of papers. The advantage of adding these identifiers is based on the fact that data archives, such as PDB and GenBank, and knowledgebases, such as

UniProt and RefSeq can then automatically identify relevant papers and thus speed up the integration of knowledge into widely used knowledgebases. One example where identifier inclusion is beneficial is when researchers may not be explicit about the origin species of the protein they are discussing in the paper. This vagueness can significantly weaken the usefulness of research knowledge and be easily rectified by the inclusion of relevant identifiers. Unfortunately, the adoption of accession identifiers into the scientific literature is rather low; perhaps less than 10% of relevant papers do this. By strengthening the adoption of the use of compact identifier standards (Wimalaratne et al. 2018. PMID:PMC5944906) and research resource identifiers (Bandrowski et al. 2015. PMID:PMC4648211) the NIH can improve the ability of knowledge to be captured. Thus, we recommend that in the data preservation and access section of the data management and sharing policy, researchers are explicitly instructed to include relevant database and knowledgebase identifiers when publishing their research work.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

No comments

Submission #97

Date: 12/10/2018

Name: Nicole Henwood

Name of Organization: NF2 BioSolutions

Type of Organization: Nonprofit Research Organization

Role: Patient Advocate

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

Gene therapy for NF2

II. The requirements for Data Management and Sharing Plans

As a physician and the mother of a child with a rare disease, I can not stress the importance of requiring data sharing. My foundation is trying to accelerate gene therapy for NF2 and the costs of bio distribution and toxicology studies is prohibitive at this point in time for even though there have been several other bio distribution studies using the same vector. Being able to build upon prior research in the biologics arena in particular will be invaluable to those of in the rare disease community.

Submission #98

Date: 12/10/2018

Name: Claire Zhu

Name of Organization: NCI

Type of Organization: Government Agency

Role: Government Official

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

Cancer genomics

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

I feel that the only effective way for implementing these rather complex requirements is to incorporate these requirements in the funding application form itself, rather than having the PI providing a free-form text as an attachment. A structured data collection (i.e. embedded in the application form) would allow NIH more control on the use of common data elements, and to more easily track compliance and produce reports.

Submission #99

Date: 12/10/2018

Name: Mary Jo Hoeksema

Name of Organization: Population Association of America/Association of Population Centers

Type of Organization: Professional Org/Association

Role: Member of the Public

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

demography or population research

I. The definition of Scientific Data

see attached letter

II. The requirements for Data Management and Sharing Plans

see attached letter

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

see attached letter

Attachment:

**Population Association of America
President**

Dr. Wendy Manning
Bowling Green State University

Vice President

Dr. John Iceland
Pennsylvania State University

President-elect

Dr. John Casterline
Ohio State University

Vice President-elect

Dr. Noreen Goldman
Princeton University

Secretary-Treasurer

Dr. Bridget Gorman
Rice University

Past President

Dr. Amy Tsui
Johns Hopkins University

Dr. Lisa Berkman

Harvard University

Dr. Kathleen Cagney

University of Chicago

Dr. Jason Fields

US Bureau of Census

Dr. Emily Hannum

University of Pennsylvania

Dr. Jeffrey Morenoff

University of Michigan

Dr. Jenna Nobles

University of Wisconsin, Madison

Dr. Mary Beth Ofstedal

University of Michigan

Dr. Krista Ferreira

University of North Carolina

Dr. Zhenchao Qian

Brown University

Dr. James Raymo

University of Wisconsin, Madison

Dr. Leah Van Wey

Brown University

Dr. Kathryn M. Yount

Emory University

**Association of Population Centers
President**

Dr. Steve Ruggles

University of Minnesota

Vice President

Dr. Jennifer Van Hook

Pennsylvania State University

Treasurer

Dr. Andrew Foster

Brown University

Secretary

Dr. Sara Curran

University of Washington

December 7, 2018

National Institutes of Health
Office of Science Policy
6705 Rockledge Drive, Suite 750
Bethesda, MD 20892

To whom it may concern:

On behalf of the over 3,000 scientists who are members of the Population Association of America (PAA) (www.populationassociation.org) and the over 40 federally supported population research centers at U.S. based research institutions comprising the Association of Population Centers (APC), we are pleased to respond to the “Request for Information (RFI) on Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research (NOT-OD-19-014).”

Population scientists include demographers, economists, and sociologists who conduct research on population trends and the individual, societal, and environmental implications of population change. They rely on discretionary grant support from the National Institutes of Health (NIH) and accurate and timely data from the federal statistical agencies to produce research findings and conduct research training activities. Population scientists also have unique expertise in data collection, dissemination, and archiving strategies. Thus, the draft NIH data management and sharing policy is especially central to our organizations.

In sum, our organizations support the data management and sharing principles expressed by the NIH. We feel strongly that “data should be made as widely and freely available while also safeguarding the privacy of participants and protecting confidential and proprietary data.” To this end, we believe data sharing should be consistent with the FAIR (Findable, Accessible, Interoperable, and Re-usable) data principles. Further, we believe that data collected as part of any NIH award should be shared regardless of the award’s size. We also support the use of centralized archives for long-term dissemination and support, as well as the development of archives to handle analysis and dissemination of restricted data, such as “Data Sharing for Demographic Research” (DSDR) program, which is funded by the National Institute of Child Health and Human Development (NICHD).

To inform further refinements to the policy, we offer several recommendations.

Recommendation #1: Develop policy for sharing and archiving data extracts

We encourage NIH to articulate how data extracts, analysis files, constructed variables, etc... that are derived from primary sources of data (surveys and other sources), and are used in specific analyses and publications, will be shared. There should be a policy and opportunities for these extracts to be archived in appropriate repositories to facilitate future research, including replication studies. In addition, archiving extracts should be consistent with policies of reuse established by primary data collectors.

Recommendation #2: Address management of paradata

NIH should encourage the systematic collection, documentation, and dissemination of paradata—i.e., data about the data collection process. These data can help users better understand and interpret primary data and support survey methods research that benefits future and ongoing data collection activities.

Recommendation #3: Reward data collection and sharing

The RFI focuses largely on compliance and enforcement. We recommend stipulating enhanced incentives to ensure greater compliance. For example, citations benefit data collectors, offering them recognition and reward, but requires establishing new norms about citing data files. The final policy should provide clear citation guidance, including recommendations for how to cite secondary data that are created and shared with the research community. Both primary and secondary data that are eligible for citation should receive an NIH data catalog record analogous to a PMID or PMCID (in addition to be cataloged using DOIs or other persistent identifiers). Data collectors should also be required to provide clear guidance to users who cite their data--especially for complex, multi-part, and long-running surveys.

Recommendation #4: Address costs of data sharing

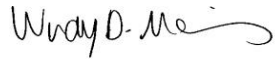
Sharing data properly and widely incurs costs usually towards the end of a project after the data have been collected and processed, and funds are exhausted. As part of its data management and sharing policy, we encourage NIH to consider options that could help offset costs associated with data sharing. These strategies could include: holding a fraction of funds in “escrow” for release at end of project for data sharing; award separate supplements to cover data sharing costs; and/or separate data sharing and archiving grants similar to an R03 program ([PAR-16-149](#)) that the NICHD has successfully implemented.

Recommendation #5: Timing

Ideally, a timeline for data sharing should be identified, and should ideally occur before the end of the grant—though this is not always possible. We encourage NIH to implement its new data management and sharing policy quickly and efficiently rather than slowly phasing in the policy.

Thank you for considering our recommendations as you develop a data management and sharing policy for NIH funded or supported research. We are pleased to offer our organizations as resources as the agency develops its final policy.

Sincerely,

Handwritten signature of Wendy Manning in black ink.

Wendy Manning, Ph.D., President
Population Association of America

Handwritten signature of Steve Ruggles in black ink.

Steve Ruggles, Ph.D., President
Association of Population Centers

Submission #100

Date:12/10/2018

Name: Marcin Cieslik

Name of Organization: University of Michigan

Type of Organization: University

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

Genomics, Precision Oncology

II. The requirements for Data Management and Sharing Plans

In my opinion the policy should be prescriptive "this is what you have to do" rather than descriptive "please tell us what you are planning to do" - a descriptive policy just makes life difficult - as researchers will try to guess what they are supposed to do/write in the 2pg statement and reviewers will be trying to figure out whether this plan enough of not enough. Just set and enforce reasonable rules that apply to everyone.

Submission #101**Date:** 12/10/2018**Name:** Michael Litzsinger**Name of Organization:** Project Data Sphere, LLC**Type of Organization:** Nonprofit Research Organization**Role:** Scientific Researcher**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Oncology clinical trial data, images and genomics. Project Data Sphere, LLC's focus is on enabling researchers to investigate cancer data that spans organizations and institutions with the goal of developing new insights that lead to improved outcomes for cancer patients.

I. The definition of Scientific Data

Scientific Data represents a broad range of information. This includes patient-level data collected during the execution of a clinical trial, real-world data collected during patient care, genomic data (especially somatic genomics data), metadata regarding clinical trials and research proposals.

II. The requirements for Data Management and Sharing Plans

Project Data Sphere, LLC, has been aggregating clinical trial data from industry and institutional organizations for over 4 years. To date, the Project Data Sphere platform contains patient-level data representing more than 140,000 patients spanning more than 180 datasets. As our mission evolves, we are beginning to aggregate imaging and genomics data in addition to the core patient-level data that has been ongoing.

One of the critical challenges that we have identified is that data elements for a common research program may exist in multiple locations. Traditional trial data may be stored with the investigator, radiologic images may be stored in a contracted repository, and genomic data may be stored elsewhere (and perhaps categorized at a latter point in time). To enable comprehensive data to be successfully shared, the natural key that links trial data, to images, to genomics data exists at the point of care. However, the nature of data sharing typically requires patient data to be de-identified or anonymized (and please note that although these terms are often used interchangeably, they carry very specific definitions with which many scientists and informaticians may not be familiar). This could mean, for example, that trial data is shared at one point in time, but when the images or genomics data are additionally shared, the natural key that exists at the point of care is frequently not available. That is, a patient whose original

identifier is #12345 may have their identifier anonymized to #ABCDE when their trial data is shared. At a latter point in time, the imaging data might be anonymized to #ZYXWV, and so forth, rendering it impossible to reliably re-integrate a patient's comprehensive data.

Although the simplest approach could be to simply anonymize all of a patient's data at the same time in order to preserve the linkages between their data, this approach may not be practical in many cases, or may prohibitively delay the act of data sharing itself. There are, however, different strategies that can be applied to preserve an identifying key that can be re-applied to different data sharing activities for a patient, and implementing these strategies is critical for preserving a comprehensive data portrait for a patient, and doing so in an efficient manner.

It is, additionally, important to avoid considering data sharing activities in isolation. It may be possible, for example, that NIH develops and implements a robust data sharing policy that limits sharing to an NIH-only data sharing platform. This would mean, for example, that researchers wishing to investigate NIH data together with industry data (or data from other sources) would need a mechanism for bringing this data together. This could potentially be built in to any NIH data sharing infrastructure, or external data sharing platforms that are already in use could be utilized. Similarly, the community of researchers would potentially benefit by being able to develop enhancements to the shared data, and then sharing these back to the community to foster greater collaboration and transparency, while at the same time reducing data re-work across researchers.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

Developing and then executing an idealized process could take years, if not decades. In the meantime, new, potentially life-saving insights would remain undiscovered while comprehensive data sharing solutions are developed. Many patients cannot wait. There are proven examples of successful data sharing platforms that enable novel clinical research. These platforms should be investigated and used, as is, as a starting point for NIH data sharing activities. Over time, through real world learnings, they can then be expanded and improved to support emerging scientific research needs.

Think big. Start small. Do it now.

Submission #102

Date: 12/10/2018

Name: Arman Yashar Khojandi

Name of Organization: National Institute of Mental Health

Type of Organization: Government Agency

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

Neuroscience, computational data science

Submission #103**Date:** 12/10/2018**Name:** Margaret Levenstein**Name of Organization:** Inter-university Consortium for Political and Social Research**Type of Organization:** University**Role:** Scientific Researcher**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Social and behavioral science

An international consortium of more than 780 academic institutions and research organizations, the Inter-university Consortium for Political and Social Research (ICPSR) provides leadership and training in data access, curation, and methods of analysis for the social science research community. ICPSR archives over 10,000 data collections comprising 250,000 files of data and documentation, which are downloaded millions of times each year. ICPSR data form the foundation for tens of thousands of research articles, reports, and books that advance science. ICPSR was also one of the founding members of the Data Documentation Initiative (DDI), which has become an international standard for metadata in the social sciences, and we provide the home office for the DDI Alliance. The ICPSR Summer Program, established in 1963 as a complement to ICPSR's data services, is internationally recognized as the leader for training in research methodologies and technologies used across the social, behavioral, and health sciences.

I. The definition of Scientific Data

ICPSR supports NIH's proposed definition of Scientific Data, including the importance of metadata to accompany data. While we agree with the statement "scientific data may include certain individual level and summary or aggregate data," we believe scientific data *must* (not just "may") include metadata. Data without metadata are generally meaningless and impossible to re-use.

One item missing from the definition of Scientific Data is original software created in the course of research. This may have been implied in the term "recorded factual material," but we believe it is important to emphasize. The UK's Wellcome funding body, for instance, specifies original software directly alongside data within the scope of its outputs management plan requirements.

II. The requirements for Data Management and Sharing Plans

Plan Review and Evaluation:

We agree that plans should not be factored into the overall impact score of an extramural grant. There is currently too much variability among peer reviewers when evaluating data management plans to make this a useful addition to the overall impact score. Instead, we recommend that data management plan reviews be conducted by trained NIH staff, who can provide consistent, objective oversight and be able to work with potential awardees to ensure that concerns are addressed as a contingency of NIH funding.

We also recommend that plan guidelines across NIH ICs include common definitions, expressed as clearly and unambiguously as possible.

Plan Elements:

We agree with many of the proposed plan elements, which align closely with the elements we currently recommend be included in a DMP. However, two elements related to data quality appear to be missing, and we encourage these be added. The first is data organization: How will the data be managed during the project, with information about version control, naming conventions, etc.? The second is quality assurance: What are the procedures for ensuring data quality during the project? Scientific Data need to be well-managed, complete, and self-explanatory to facilitate FAIRness (Findable, Accessible, Interoperable, and Re-usable). Messy or dirty data, with missing or inconsistent metadata, can be unnecessarily challenging and time consuming to re-use.

While metadata is indirectly included under data type standards, we recommend explicitly asking about metadata, including a discussion of the metadata standards used. We cannot overemphasize the importance of good metadata to data management.

We are happy to see a plan element asking “where scientific data will be archived to ensure its long-term preservation.” Many disciplines have good repository options available and NIH should emphasize and encourage use of these existing domain repositories. Additionally, we recommend that NIH encourage use of certified data archives, such as through the CoreTrustSeal data repository certification. Certified repositories undergo external audits and help insure data are FAIR and long-lived in a trustworthy repository. Certified repositories can then provide letters of support indicating willingness to archive researchers’ data.

DMP elements should be clearly defined, with a clear reason provided for including each element. From the University of Michigan Library: “In addition to providing a clear explanation of the requirements (including definitions of key terms and concepts), agencies should provide a rationale behind the requirements in their guidance. For example, it may not be clear to the researcher why they would need to include the file formats of their data set in a DMP. Instructions provided by the agency could include a statement on the need to encourage the use of community supported, open formats and to help repositories better plan and prepare to receive the data.”

Generally, we encourage NIH to align plan elements with other funders, as well as with other ICs. Several other major funding agencies now require data management plans, including the National Science Foundation. While a one-plan-fits-all approach will not work for the heterogeneous range of data covered across disciplines, aligning core plan elements across funders would make life easier for researchers since they would know what to expect regardless the funder. Alignment of data management plan elements will also encourage standardization of metadata, lowering the cost to researchers to produce metadata and increasing the ease with which it can be automated.

We encourage NIH to make plans machine-readable and machine-actionable for interoperability and data mining. The Research Data Alliance is working on 'Active' DMPs, which will "define a common information model and specify access mechanisms that make DMPs machine-actionable; this will help to make systems interoperable and will allow for automatic exchange, integration, and validation of information provided in DMPs. These initiatives will increase the extent to which DMPs are integrated in the research lifecycle and the management of research information, bringing benefits to research teams, institutions and funders."

Compliance and Enforcement:

We encourage NIH ICs to monitor for compliance using machine-actionable DMPs and make DMPs an explicit component of reporting to the funding agency. Non-compliant DMPs would be enforced at the end of the grant through withholding a portion of the remaining funds.

We also encourage NIH to share the information collected from data management plans with the public. We agree with this recommendation from the University of Michigan Library: "Funding agencies should consider sharing the information collected from data management plans in aggregate to better inform the repository, publishing, curation and other communities who provide support for data management, sharing and preservation. Having access to the content of DMPs would help these communities better understand the nature and types of data being generated and to better anticipate and respond to researcher needs. If a specific repository is named in the DMP, consider sharing the DMP with the repository."

Other

We agree with the following statement from the Scope and Requirements section: "Reasonable costs associated with data management and sharing could be requested under the budget for the proposed project." Data management and sharing are real expenses. Proposals should accurately describe these costs and how they will be paid. Data repositories can help estimate these costs.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and

Several elements of this updated plan should be adopted 3 months after publication, including the definition of Scientific Data and the core plan elements. This provides some buffer for funding applicants to adjust, but since the changes are helpful in providing clarification and guidance, it would be advantageous to implement soon.

For the elements of the plan involving reporting, compliance, and enforcement, a longer delay, such as 12 months, could give funding applicants more time to prepare, as well as NIH ICs more time to develop systems and train staff to monitor reporting and compliance.

Attachment:

Response to RFI: [Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research](#)

From: Inter-university Consortium for Political and Social Research (ICPSR), Institute for Social Research, University of Michigan

Contact: Margaret C. Levenstein, Director

Date: December 10, 2018

About ICPSR

An international consortium of more than 780 academic institutions and research organizations, the Inter-university Consortium for Political and Social Research (ICPSR) provides leadership and training in data access, curation, and methods of analysis for the social science research community. ICPSR archives over 10,000 data collections comprising 250,000 files of data and documentation, which are downloaded millions of times each year. ICPSR data form the foundation for tens of thousands of research articles, reports, and books that advance science. ICPSR was also one of the founding members of the Data Documentation Initiative (DDI), which has become an international standard for metadata in the social sciences, and we provide the home office for the DDI Alliance. The ICPSR Summer Program, established in 1963 as a complement to ICPSR's data services, is internationally recognized as the leader for training in research methodologies and technologies used across the social, behavioral, and health sciences.

I. [The definition of Scientific Data](#)

ICPSR supports NIH's proposed definition of Scientific Data, including the importance of metadata to accompany data. While we agree with the statement "scientific data may include certain individual level and summary or aggregate data," we believe scientific data *must* (not just "may") include metadata. Data without metadata are generally meaningless and impossible to re-use.

One item missing from the definition of Scientific Data is original software created in the course of research. This may have been implied in the term "recorded factual material," but we believe it is important to emphasize. The UK's Wellcome funding body, for instance,

specifies original software directly alongside data within the scope of its outputs management plan requirements.¹

II. [The requirements for Data Management and Sharing Plans](#)

Plan Review and Evaluation:

We agree that plans should not be factored into the overall impact score of an extramural grant. There is currently too much variability among peer reviewers when evaluating data management plans to make this a useful addition to the overall impact score. Instead, we recommend that data management plan reviews be conducted by trained NIH staff, who can provide consistent, objective oversight and be able to work with potential awardees to ensure that concerns are addressed as a contingency of NIH funding.

We also recommend that plan guidelines across NIH ICs include common definitions, expressed as clearly and unambiguously as possible.

Plan Elements:

We agree with many of the proposed plan elements, which align closely with the elements we currently recommend be included in a DMP.² However, two elements related to data quality appear to be missing, and we encourage these be added. The first is data organization: How will the data be managed during the project, with information about version control, naming conventions, etc.? The second is quality assurance: What are the procedures for ensuring data quality during the project? Scientific Data need to be well-managed, complete, and self-explanatory to facilitate FAIRness (Findable, Accessible, Interoperable, and Re-usable). Messy or dirty data, with missing or inconsistent metadata, can be unnecessarily challenging and time consuming to re-use.

While metadata is indirectly included under data type standards, we recommend explicitly asking about metadata, including a discussion of the metadata standards used. We cannot overemphasize the importance of good metadata to data management.

¹ Wellcome Website, Developing an outputs management plan:
<https://wellcome.ac.uk/funding/guidance/developing-outputs-management-plan>
(accessed November 30, 2018)

² ICPSR Web site, Elements of a Data Management Plan:
<https://www.icpsr.umich.edu/icpsrweb/content/datamanagement/dmp/elements.html>
(accessed November 28, 2018)

We are happy to see a plan element asking “where scientific data will be archived to ensure its long-term preservation.” Many disciplines have good repository options available and NIH should emphasize and encourage use of these existing domain repositories. Additionally, we recommend that NIH encourage use of certified data archives, such as through the CoreTrustSeal data repository certification.³ Certified repositories undergo external audits and help insure data are FAIR and long-lived in a trustworthy repository. Certified repositories can then provide letters of support indicating willingness to archive researchers’ data.

DMP elements should be clearly defined, with a clear reason provided for including each element. From the University of Michigan Library: “In addition to providing a clear explanation of the requirements (including definitions of key terms and concepts), agencies should provide a rationale behind the requirements in their guidance. For example, it may not be clear to the researcher why they would need to include the file formats of their data set in a DMP. Instructions provided by the agency could include a statement on the need to encourage the use of community supported, open formats and to help repositories better plan and prepare to receive the data.”⁴

Generally, we encourage NIH to align plan elements with other funders, as well as with other ICs. Several other major funding agencies now require data management plans, including the National Science Foundation. While a one-plan-fits-all approach will not work for the heterogeneous range of data covered across disciplines, aligning core plan elements across funders would make life easier for researchers since they would know what to expect regardless the funder. Alignment of data management plan elements will also encourage standardization of metadata, lowering the cost to researchers to produce metadata and increasing the ease with which it can be automated.

We encourage NIH to make plans machine-readable and machine-actionable for interoperability and data mining. The Research Data Alliance is working on ‘Active’ DMPs,⁵ which will “define a common information model and specify access mechanisms that make DMPs machine-actionable; this will help to make systems interoperable and will allow for automatic exchange, integration, and validation of information provided in DMPs. These initiatives will increase the extent to which DMPs are integrated in the research lifecycle

³ CoreTrustSeal Web site: <https://www.coretrustseal.org/> (accessed November 28, 2018)

⁴ Jake Carlson. An Analysis of Data Management Plans from the University of Michigan. <http://hdl.handle.net/2027.42/136230>

⁵ See the Active DMPs Interest Group: <https://www.rd-alliance.org/groups/active-data-management-plans.html>, DMP common standards WG: <https://www.rd-alliance.org/groups/dmp-common-standards-wg>, and Exposing DMPs WG <https://www.rd-alliance.org/groups/exposing-data-management-plans-wg>

and the management of research information, bringing benefits to research teams, institutions and funders.”⁶

Compliance and Enforcement:

We encourage NIH ICs to monitor for compliance using machine-actionable DMPs and make DMPs an explicit component of reporting to the funding agency. Non-compliant DMPs would be enforced at the end of the grant through withholding a portion of the remaining funds.

We also encourage NIH to share the information collected from data management plans with the public. We agree with this recommendation from the University of Michigan Library: “Funding agencies should consider sharing the information collected from data management plans in aggregate to better inform the repository, publishing, curation and other communities who provide support for data management, sharing and preservation. Having access to the content of DMPs would help these communities better understand the nature and types of data being generated and to better anticipate and respond to researcher needs. If a specific repository is named in the DMP, consider sharing the DMP with the repository.”⁷

Other

We agree with the following statement from the Scope and Requirements section: “Reasonable costs associated with data management and sharing could be requested under the budget for the proposed project.” Data management and sharing are real expenses. Proposals should accurately describe these costs and how they will be paid. Data repositories can help estimate these costs.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards.

Several elements of this updated plan should be adopted 3 months after publication, including the definition of Scientific Data and the core plan elements. This provides some

⁶ Turning FAIR into reality: Final Report and Action Plan from the European Commission Expert Group on FAIR Data. <https://doi.org/10.2777/1524>

⁷ Jake Carlson. An Analysis of Data Management Plans from the University of Michigan. <http://hdl.handle.net/2027.42/136230>

buffer for funding applicants to adjust, but since the changes are helpful in providing clarification and guidance, it would be advantageous to implement soon.

For the elements of the plan involving reporting, compliance, and enforcement, a longer delay, such as 12 months, could give funding applicants more time to prepare, as well as NIH ICs more time to develop systems and train staff to monitor reporting and compliance.

Submission #104

Date: 12/10/2018

Name: Iain Hrynaszkiewicz

Name of Organization: Springer Nature

Type of Organization: Other

Other Type of Organization: Scholarly publisher

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

Springer Nature publishes research across all disciplines including all areas of health research, which are all equally important.

I. The definition of Scientific Data

Please see the attached document for our comments

II. The requirements for Data Management and Sharing Plans

Please see the attached document for our comments

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

Please see the attached document for our comments

Attachment:

Springer Nature Limited
The Campus
4 Crinan Street
London N1 9XW
United Kingdom

T +44 20 7833 4000
www.springernature.com

Iain Hrynaszkiewicz
Head of Data Publishing
iain.hrynaszkiewicz@springernature.com

NIH Office of Science Policy
NIH Office of the Director
SciencePolicy@od.nih.gov

10th December 2018

Springer Nature response to Request for Information on Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research

Prepared by Iain Hrynaszkiewicz, Head of Data Publishing and Rebecca Grant, Research Data Manager

To Whom it May Concern,

Introductory comments

Springer Nature is committed to open research and offers researchers, institutions and their funders open access (OA) options for journals, books and sharing research data. We also offer expertise and services that support research data management and sharing.

Springer Nature has an established interest in policies on data sharing and data management and welcomes this opportunity to provide comments to the NIH on the development of its policies. To our knowledge, more than 80 funding agencies globally now have data sharing policies with around half of those with a policy requiring data management plans (DMPs). However, how policies are implemented, and their compliance monitored, and how these activities are resourced are vitally important^{1,2}.

Many Springer Nature journals publishing in the life sciences, including Nature and the BMC journals, have operated data sharing policies for many years. More recently, in 2016, Springer Nature launched an initiative to standardise, harmonise, and implement consistent data sharing policies to all its journals. Several other large publishers, including Elsevier, Wiley, Taylor & Francis, BMJ journals, have since introduced similar initiatives and globally publishers are collaborating via the Research Data Alliance (RDA) to standardise and implement journal policies.

Specific comments requested by NIH

Definition of Scientific Data

Our first comment is regarding the term “scientific data”. Given research is increasingly interdisciplinary and international, and to ensure an inclusive policy to all NIH grantees, we suggest using “research data” instead of “scientific data”. Springer Nature - which works with researchers from the humanities through to the life and physical sciences - consciously uses “research data” in the context of data policy development and implementation across our organisation, and product and service portfolios (<https://www.springernature.com/gp/authors/research-data>).

In our experience of research data policy and service development and implementation, defining the scope of the policy is critical. As a research publisher, this generally means focusing on research data that specifically support scholarly publications and we welcome this explicit definition (“including, but not limited to, data used to support scholarly publications”) in the policy draft. In addition, it may be helpful to reference or support established community norms for data sharing (such as DNA/RNA sequence deposition), in cases where they are not covered by the first definition. We additionally recommend NIH define whether or not their definition of “scientific data” includes:

- All data generated by a grant
- All data analysed (including data from third parties) by grant holders in their research

Material from (electronic) laboratory notebooks, and preliminary analyses, can be used to support scholarly publications (for example, in³) so we suggest these not be excluded from the policy and definition. Also, we recommend being explicit about whether research software and code are included in the policy.

In addition, we would caution referencing “Data sharing...in a manner that is consistent with the FAIR principles”. While the FAIR principles are important framework for data sharing and increasingly used by policy makers, institutions and those already active in the research data management (see Jisc report for more information⁴), a recent survey by Digital Science found that more than 60% of researchers were not familiar with FAIR⁵. This might suggest that, where researchers are the audience, a more general or accessible definition for data sharing may be helpful.

Requirements for Data Management Plans

As part of our Nature Research Academies training on research data sharing, we also provide training to researchers in preparing data management plans and interpreting the guidance provided by their funder or institution. We have found that it is beneficial to focus on the practical elements of data management planning, to ensure that a realistic and actionable plan is developed. For this reason, it might be beneficial to include additional resources for researchers to assist them in preparing their plans, e.g. examples of well-written plans and a list of sources where they can find information on good practice in data storage, metadata creation, and repository selection.

More detailed guidance could be provided on the following:

- Required retention periods for the data, including a defined minimum, for example 10 years from the date the data were last accessed.
- Data types which do not have a discipline-specific repository available.
- Issues of copyright and data ownership.

It seems somewhat arbitrary to have a two-page limit on these plans - a one- to two-page minimum might work better in order to encourage detailed planning.

In relation to compliance and enforcement, consideration should be given to the likelihood of a plan changing from the time that a grant proposal is drafted, to when the data collection begins, through data analysis and archiving. It would be more beneficial for the researcher to consider the plan as a living document which should continuously be referred to and continuously be updated. For this reason, annual reviews could aim to ensure that the plan is being updated on a regular basis, rather than being complied with rigidly.

To provide additional opportunities for researchers to consider their plan in a practical way, it might also be helpful to add the following topics:

- Short term storage (while the research is being conducted).
- Resourcing (both technical infrastructure required and the skills and/or training needs of the researchers to carry out the plan).

Timing, implementation, resources

The focus of the policy as worded appears to be on the requirements for data sharing and data management plans, with fewer details on compliance and enforcement, and on how researchers, support staff (including programme officers) will be supported to implement and monitor compliance with the policy. An analysis of research data management policy implementation from several case studies by Cameron Neylon (2017), highlighted, amongst other relevant findings, that “for DMP requirements to be supportive of culture change they need to be well supported with expertise, systems and guidance in place,” and that “Requirements imposed on data sharing...must be auditable and audited”².

Where and how researchers can best receive support - such as through formal training, through supervision or dedicated data management staff and help desks - will vary. Support for creating and implementing DMPs can be available from institutions, funding agencies and third parties such as the Digital Curation Centre and scholarly publishers’ researcher services (<https://www.springernature.com/gp/open-research/institutions/research-data-services>). Leveraging a wider “market place” for resources to support open science has become a key part of the European Commission’s Open Science Cloud (<https://www.eosc-portal.eu/for-providers>). We would welcome more detail on how NIH will support researchers in preparing and implementing their DMPs. As well as support in terms of tools and training, we recommend explicit clarity about use of NIH-funding for costs associated with data management and curation, storage and publication fees of articles that describe datasets (such as data papers, data notes, data descriptors).

We recommend more consideration be given to the mechanisms by which compliance with the policy can be demonstrated by researchers. Monitoring of compliance with data sharing policies and data management plans is, understandably, challenging because the practices, expectations and available infrastructure (including repositories) for sharing research data vary by discipline. Other funding agencies, and research publishers have implemented requirements for transparency in reporting of information about data availability, as a prerequisite to monitoring compliance. This is achieved by requiring researchers provide, in their publications, a ‘Data availability statement’ (sometimes called ‘Data accessibility statement’ or ‘Data sharing statement’).

Provision of Data availability statements is a requirement of hundreds - if not thousands - of journals and publishers’ research data policies. These statements are a common feature of publishers’ data sharing policies. Data availability statements in researchers’ publications are also part of the policies a number of funding agencies including the seven UK Research Councils (<https://www.ukri.org/files/legacy/documents/rcukcommonprinciplesondatapolicy-pdf/>). There is, also, evidence that consistent statements of data availability, combined with a mandatory deposition policy, is the most effective approach for journals to ensure data availability supporting publications long term⁶.

Journals and publishers that require these statements in publications include all the Nature and BMC (BioMed Central) journals, as well as PLOS, BMJ, and others. More recently, publishers and the Belmont Forum (which includes the US National Science Foundation) group of agencies are collaborating to introduce consistent requirements for Data availability statements for the member agencies, their grantees, and publishers⁷. This more consistent reporting of data availability also

supports the utility of literature search and evaluation tools and services designed to determine availability of data supporting scholarly publications (for example, <http://blog.europepmc.org/2018/11/mapping-out-path-to-data.html>).

For the above reasons, we recommend NIH consider including requirements for the provision of data availability (accessibility) statements in its Data Sharing and Data Management policy requirements.

Other common features of research data policies, to publishers and funding agencies, include support for formal citation (referencing) of research data in reference lists (bibliographies), provision of researcher support (helpdesk) functions, and collaboration with trusted data repositories for implementation.

In addition to reviewing DMPs for compliance, consideration should be given to how researchers can be credited and rewarded for good practice. This can be facilitated, for example, by promoting the sharing of datasets in repositories that assign persistent identifiers (such as DOIs) to datasets so that datasets can be formally cited, and tracked, in scholarly publications. Data citation was a topic of a previous NIH RFI, to which Springer Nature responded, and more information on our proposals for data citation can be found in our response⁸. We note that the National Science Foundation encourages researchers to list their “research products”, including datasets and software, on their bibliographic sketches⁹. It is also possible to share DMPs publicly, in repositories and furthermore some journals, such as *BMC Research Notes*, will consider publishing them as peer-reviewed articles.

Other remarks

Regarding the statement: “NIH encourages the sharing of data for as long as it is useful to the scientific community.” It would be helpful to clarify if this refers to the NIH’s overall position on data sharing, and if so if this represents any change in stringency of the policy compared to the 2003 policy. That is, does the NIH *encourage* data sharing by all grant holders or *require* it?

Regarding the length of time that data archiving, the UK Research and Innovation (UKRI)’s guidelines reference a minimum of 10 years from date of last access. Similarly, when assessing data repositories for Springer Nature’s recommended repository list¹⁰, we typically look for sustainability of infrastructure for a minimum of 10 years (<https://www.nature.com/sdata/policies/data-policies#repo-suggest>).

Acknowledgements

For comments on the first draft of this proposal thanks to: Sowmya Swaminathan, Head of Editorial Policy, Nature Research Group, Springer Nature, and Grace Baynes, VP Research Data and New Product Development, Springer Nature.

Springer Nature is happy to be contacted for more information about our response (researchdata@springernature.com or iain.hrynaszkiewicz@nature.com) and encourages NIH's further participation in international fora such as the RDA data policy standardisation and implementation Interest Group (<https://www.rd-alliance.org/groups/data-policy-standardisation-and-implementation-ig>).

Yours faithfully,



Iain Hrynaszkiewicz

Bibliography

1. Grant, R. & Hrynaszkiewicz, I. The impact on authors and editors of introducing Data Availability Statements at Nature journals. *BioRxiv* (2018). doi:10.1101/264929
2. Neylon, C. Building a culture of data sharing: policy design and implementation for research data management in development research. *RIO* **3**, e21773 (2017).
3. Panetta, J. L. *et al.* Reptile-associated *Borrelia* species in the goanna tick (*Bothriocroton undatum*) from Sydney, Australia. *Parasit. Vectors* **10**, 616 (2017).
4. Allen, R. & Hartland, D. FAIR in practice - Jisc report on the Findable Accessible Interoperable and Reuseable Data Principles. (2018).
5. Science, D.. The State of Open Data Report 2018. (2018). doi:10.6084/m9.figshare.7195058.v2
6. Vines, T. H. *et al.* Mandated data archiving greatly improves access to research data. *FASEB Journal* fj.12-218164- (2013).
7. Murphy, F. & Samors, R. J. Belmont Forum Data Accessibility Statement Policy and Template - Endorsed 18 October 2018. (2018).

8. Springer Nature response to NIH RFI on Strategies for NIH Data Management, Sharing, and Citation. (2017). doi:10.6084/m9.figshare.4616293.v1.
9. Piwowar, H. Altmetrics: Value all research products. *Nature* **493**, 159 (2013).
10. Data, S. Scientific Data recommended repositories. (2018). doi:10.6084/m9.figshare.1434640.v12

Submission #105

Date: 12/10/2018

Name: Anne Klibanski, M.D.

Name of Organization: Partners HealthCare

Type of Organization: Healthcare Delivery Organization

Role: Institutional Official

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

Neuroendocrine

Attachment:



FOUNDED BY BRIGHAM AND WOMEN'S HOSPITAL
AND MASSACHUSETTS GENERAL HOSPITAL

December 7, 2018

Francis S. Collins, MD, PhD
National Institutes of Health
Bethesda, MD

Submitted electronically: <https://osp.od.nih.gov/provisions-data-managment-sharing/>

Re: RFI on Proposed Provisions for Draft Data Management and Sharing Policy

Dear Dr. Collins:

Thank you very much for providing the research community with an opportunity to comment on the NIH proposed policy. I am writing on behalf of Partners HealthCare System (Partners) which is not a recipient of federal research funds but which provides financial and administrative oversight of research grants, cooperative agreements, and contracts awarded to its member hospitals: Brigham and Women's Hospital (BWH), Massachusetts General Hospital (MGH), McLean Hospital (McLean), Spaulding Rehabilitation Hospital (SRH), and Massachusetts Eye and Ear Infirmary (MEE). Partners is one of the nation's leading non-profit biomedical research organizations; its hospitals are the principal teaching affiliates of Harvard Medical School. In FY 18, Partners hospitals received slightly less than \$800 million in NIH/HHS support. Thus, any proposed change in NIH data management and sharing requirements is of vital interest to us.

Let me begin by stating my colleagues and I conceptually support data sharing as a means of enabling researchers to test the validity of scientific findings, explore new scientific pathways, and shorten the time for ideas to move from the bench to the bedside. Yet, the devil is in the details for data sharing to be successful. The proposed policy is so broad and all-encompassing, we believe if implemented it would be extremely difficult for the NIH to achieve its objective of enhancing science, let alone for Principal Investigators (PI) and institutions to meet their compliance requirements.

Some of our investigators have suggested that the proposed policy appears to be an extension of data sharing requirements for genetic data to scientific data more generally. Genetic data sharing through dbGaP and similar repositories works because genetic data can be supported with standard file formats for data submission. We find it difficult to envision how the many possible experimental designs for laboratory-based experiments would be submitted and archived in a way that could be interpreted by an outside user.

We strongly recommend that the NIH revise the proposed policy to scale back its requirements, add clarity to definitions, and provide meaningful examples for investigators. We also recommend that the NIH consider convening a group of NIH-funded investigators to work with NIH research and administrative leadership to develop a policy that is more realistic and achievable from an investigator's perspective.

Please see below for our comments on specific sections of the proposed policy.

1. Section I

- a. Definitions: The definition of Scientific Data is extremely broad and confusing. We recommend considering the definition of Research Data in OMB Circular A-110 as a substitute. This definition would already be familiar to most of the research community.
- b. Lab notebooks: Throughout the policy there is confusion about lab notebooks and whether they should be shared. Their role/purpose in a “data sharing policy” should be clarified. We maintain that lab notebooks, while critical to the scientific process, are not Scientific Data; they are a means for recording experiments and the Scientific Data generated.
- c. Reasonable effort to digitize scientific data: While institutionally we are requiring our investigators to transition to digital recordkeeping, we do not recommend including a statement about digitizing scientific data within the current policy. Not all Scientific Data can be digitized; this makes the data no less valuable to research.

2. Section II. Purpose: Making Scientific Data accessible in a “timely manner:” Researchers generate data daily. We recommend clarifying this section by adding timelines for posting/sharing published and unpublished data. We recommend adding a section to the Progress Report where the PI can inform the NIH of data accessibility. The policy should be flexible. Not all data will be ready for sharing or posting in a repository at the same time. Investigators may want to refrain from posting/sharing unpublished data until it has been published. These situations should be taken into consideration in this section.

3. Section III. Scope and Requirements:

- a. We are concerned that requiring a data management/sharing plan for each application/proposal submission, when the overall funding success rate hovers at 20% or less, creates a significant administrative burden for PIs submitting applications. We recommend the NIH consider requiring the plan as part of the first progress report. These plans will not have the benefit of peer review, but is peer review necessary if the strength or weaknesses of the plan will not be considered in the impact score? Continuation funding for year 2 could be delayed until a plan acceptable to the Program Officer is submitted.
- b. In the general statement that data management/sharing plans will be required regardless of mechanism, we recommend that the NIH review the different funding mechanisms for appropriateness. For example, a data sharing plan would not be appropriate for a shared instrumentation grant; nor would it be appropriate for a conference grant. We also recommend that the NIH consider eliminating the requirement for institutional training grant applications. We recognize that Scientific Data are generated under training grants, but the management and sharing of the data will vary across the training grant based on requirements of each trainee’s mentor who often come from different departments/research labs with different data management/sharing requirements.
- c. The policy states, “Reasonable costs associated with data management and sharing could be requested under the budget for the proposed project.” Will supplemental funds be available for these costs? If not, the data management/sharing requirement will only reduce the amount of funding available for the actual research project. We can envision situations where institutions with limited resources will have to provide their investigators with institutional funds or create local data repositories because the NIH funds were simply not enough to complete the project and pay for the costs associated

with external data repositories essentially creating yet another unfunded mandate for grantee institutions.

4. Section IV. Requirements for Data Management and Sharing Plans

- a. General comment: We do not believe PIs will be able to provide all the information the NIH is requiring within a two-page limit.
- b. Scoring/Peer Review Process: If the NIH continues to require plans as part of the grant application/contract proposal, we agree whether a plan is acceptable or unacceptable to reviewers should not be included in the overall impact score.
- c. Plan Elements: We recommend that the NIH create a form with drop down boxes for the PI to identify the plan elements relevant for his/her research. The elements should be minimal and allow for PI flexibility.
- d. Describe type and amount of scientific data to be collected and used in the project: This may be difficult for some types of projects. The example provided is for a specific type of project in which the number of cases/patients/individuals may be known at submission. In many lab-based projects, investigators may improvise and adjust the work making use of techniques that may not have been envisaged initially. We are concerned that PIs may feel providing this type of information will restrict their ability to modify the research as they move forward.
- e. Related Tools, Software and/or Code: Please clarify what the NIH is expecting. For example, would the PI have to justify use of a specific image analysis software product?
- f. 4.1 Indicate where Scientific Data will be archived to ensure long-term preservation: We recommend that the NIH create data repositories to meet this new mandate. As we indicated above, many institutions do not have the resources to develop and maintain repositories for their NIH-funded investigators. Grantee institutions cannot continue to absorb unfunded mandates. Moreover, we are concerned at the possible development of numerous and heterogeneous and possibly rogue repositories.
- g. 4.4 Describe alternative plans for maintaining, preserving and providing access to scientific data should the original plan not be achieved: If the NIH is truly interested in this information, we recommend not requiring submission of a “Plan B” as part of the data management/sharing plan in their application/contract proposal. We recommend adding a section to the data management/sharing reporting section of the annual progress report to describe any changes necessary because the original plan could not be achieved.
- h. 5. Data Preservation and Access Timeline: We question the usefulness of requiring this information in the data management/sharing plans. It may be impossible at the beginning of the project to estimate timelines. This may lead PIs to develop meaningless timelines which become a compliance requirement if the application is funded. We recommend removing this requirement.
- i. 6. Data Sharing Agreements, Licensing and Intellectual Property:
 - i. “NIH encourages terms that provide for the broadest use of data resulting from NIH-funded or -supported research.” Please confirm/clarify that this statement applies to data generated as part of the study, i.e., the data would not exist if not for the study; and does NOT include any additional, pre-existing clinical data, e.g., annotated, longitudinal data pulled from a patient’s medical record.
 - ii. 6.1 “Describe any relevant data sharing agreements outlining...how scientific data can and cannot be used.” Please confirm/clarify that this applies only to Scientific Data generated as part of the study. In a situation where the project is supported by NIH and industry or a foundation, the non-NIH sponsors may limit data sharing. Would an SBIR grant be relevant here?

6.3 "[I]ndicate how intellectual property... will be managed in a way to maximize sharing of scientific data." While Scientific Data do not constitute IP, any plan to maximize sharing should not infringe upon the nature of the IP and should preserve ownership rights.

5. Compliance and Enforcement
- a. Community-based Standards: The NIH should specify these standards within the policy or at a minimum provide examples. When we consulted our investigators to develop our response, they were unsure what the standards were and where they might find them.
 - b. I/C Monitoring Plans: The policy should include information on how I/Cs will monitor plans, reporting requirements, how to modify plans during the lifetime of the grant. If an I/C determined non-compliance, what would be the enforcement mechanism?
 - c. We are very concerned about compliance/enforcement requirements extending beyond the end of the grant's performance period. If this requirement continues in the policy, the NIH should identify the authority that allows the requirement to continue in perpetuity. Comments made during the NIH webinar on the RFI seemed to suggest that the NIH does not consider data sharing requirement as continuing beyond the project end date. The proposed policy contradicts this point and should be clarified. How will the NIH monitor? How will a grantee know if a former award is out of compliance?

Once again, thank you for providing an opportunity for the research community to submit comments. Please do not hesitate to contact me for any additional information.

Yours sincerely,



Anne Kibanski, M.D.
Chief Academic Officer
Partners HealthCare

Laurie Carrol Guthart Professor of Medicine
Dean for Academic Programs at Partners HealthCare
Harvard Medical School

Submission #106**Date:** 12/10/2018**Name:** Chuck Cook**Name of Organization:** EMBL-European Bioinformatics Institute**Type of Organization:** Other**Other Type of Organization:** Intergovernmental treaty organization**Role:** Institutional Official**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

EMBL-EBI runs over 40 major life sciences data resources and undertakes research in computational biology.

I. The definition of Scientific Data

This response is on behalf of EMBL-European Bioinformatics Institute (EMBL-EBI). EMBL-EBI manages over 40 data resources that include: archives of experimental data submitted directly by researchers; knowledgebases that add value to primary experimental data through curation and analysis; ontology resources that provide descriptor frameworks; and literature resources which support scientific publication.

The scientific, economic, and societal value of scientific research is maximized if the outputs of that research are preserved and made available for reuse by other researchers. All data generated by public funding should, in principle, be preserved and made publicly available and open access for future use. If data reuse and scientific reproducibility are the drivers behind open data, then archiving the data that support this is key. For large data generation types like biological images, this may just be supporting the archiving of 'reference' data, as agreed by suitable community stakeholders, rather than all the primary data generated. Exceptions to complete open access may also be required in some circumstances, such as individual human genetic data, but the requirement for preservation remains even when researcher access is managed.

Scientific data includes:

- 1) Primary data produced by laboratory research, such as nucleotide sequences, other 'omics, and biological images.

- 2) Metadata associated with the primary data: date, species, biosample, methodology, environmental conditions, instrument settings, etc. These metadata are required to put primary data in context for further analyses.
- 3) Added value data generated from primary data. At EMBL-EBI added-value knowledgebases combine different types of primary data and enhance them through curation, creation of analytical algorithms, and development of biology-driven portals for display of new knowledge gained from this work.
- 4) The scientific literature. The scientific literature is itself a data source, and downstream curation and data mining from the literature are a major component of adding value described above.

II. The requirements for Data Management and Sharing Plans

To ensure adequate preservation, data management plans should include detailed descriptions of what types of data will be generated by the research, including metadata, and where those data will be preserved. If there is a recognized public access database for those data they should be deposited in that database.

Recognition of data resources for deposition is at present somewhat informal, with input from research communities, journals, and some resource providers, but no formal mechanism for recommending resources for particular data types. In Europe the ELIXIR infrastructure has begun formalizing the data deposition process by recommending specific data resources as deposition databases. EMBL-EBI strongly supports this effort to formalize recommended deposition resources and encourages NIH to do so as well.

If no public data resource is available for deposition other provision should be made. At EMBL-EBI we host the BioStudies resource that hosts data for which there is no established public repository. NIH should encourage researchers to deposit as much research data as possible in public repositories rather than creating bespoke local solutions within their institutions or laboratories: deposition in public repositories vastly increases the likelihood of data being findable, accessible, interoperable and reusable (Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3: 160018).

As a funder NIH should aid researchers in identifying the appropriate repository for their data, either through formal designation or through referral to appropriate standards-bodies and/or journals. Journals also have a stake in this: NIH might, for example, require NIH-funded researchers to publish only in journals that mandate appropriate data deposition.

By extension NIH also has an obligation to ensure that the life sciences data resource's infrastructure is funded for the long term to ensure that deposition data resources can continue to operate to preserve research data. NIH itself hosts many data resources, particularly within NCBI, and other NIH institutes have long history of supporting data resources. However, this

support has, to date, always been time-limited, with no assurance of continuation following grant-end. Given the importance of these data resources to the worldwide scientific effort, NIH should continue working with other funders and the global scientific community to identify fair funding criteria and to sustain funding for those data resources, which are essential for the continued success of the life sciences research effort.

For all data resources the two most costly budget categories are staff (salary and benefits) and computational infrastructure (storage, compute, networking, and data center capacity), where staff costs always substantially higher than computational infrastructure costs. With researchers generating exponentially more data for submission to public archives, the running costs for public archives will continue to increase. In general, costs for infrastructure scale well: a ten-fold increase in data submissions requires a ten-fold increase in storage capacity but not a ten-fold increase in data center staff and the decreasing costs of technology help to support the system. However, that same ten-fold increase in submissions may require very substantial increases in staff for processing the data, particularly for resources that require substantial curation or other non-scalable data analyses. If NIH intends to support public archiving of data generated by NIH research funding it should make provision to support the increasing costs of the public data resources that it supports, both within NIH and through extramural funding.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

We will not suggest here specific dates for implementation of any NIH data management policies. We do, of course, support implementation as soon as possible, but also urge careful development of any plans, with input from other funders, data resource managers, and researchers, to ensure coordination of NIH efforts with those of other funders worldwide.

Submission #107**Date:** 12/10/2018**Name:** Emily Haozous**Name of Organization:** Pacific Institute of Research and Evaluation**Type of Organization:** Nonprofit Research Organization**Role:** Scientific Researcher**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Social Science research, epidemiology and population health, and genomics.

I. The definition of Scientific Data

This definition does not honor non-western epistemologies and is focused only on a positivist understanding of knowledge. While the overall Data Sharing policy is clearly focused on bench science and the accessibility of clinical data, the narrow definition established in this document also excludes the wide body of work completed by qualitative researchers, many of whom are funded by NIH and generate large amounts of interview data.

II. The requirements for Data Management and Sharing Plans

The broadened scope that is contained within the revised data management and sharing plans policy is alarming in its clear omission of the special government-to-government relationship that the federal government has with federally-recognized American Indian and Alaska Native (AIAN) tribal nations. Although the revision does state that there are special circumstances in which data is not required to be shared, the lack of clarity regarding tribes' sovereign status and related data ownership leaves readers to draw their own conclusions as to which datasets can or cannot be withheld from a national repository. AIANs have an ongoing struggle to protect tribal data. This revision is an opportunity to include language that insists that funded researchers consult with tribes regarding the status of research on tribal lands and subsequent use of their data, setting the standard that AIAN researchers have been struggling to establish for decades. Not only is this an opportunity, it also provides researchers working with AIAN data with the necessary ethical and legal foundation they need for working with tribal data.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible

phasing could relate to needed improvements in data infrastructure, resources, and standards

Language regarding AIAN data usage must be integrated into all data sharing and management plan documents and websites immediately. I propose documents also are required to include the following questions:

- 1) Does this research include data from American Indians and/or Alaska Natives?
- 2) If yes, describe the plan for managing these data with consideration of sovereign status of federally-recognized tribes (example: include tribal resolution with plan for tribal review prior to publication, tribal notification and approval of all datasets prior to upload to public repository)

Submission #108

Date: 12/10/2018

Name: Carole Mitnick on behalf of

Name of Organization: Harvard Medical School, Department of Global Health & Social Medicine

Type of Organization: University

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

global health, epidemiology, anthropology

This response is submitted on behalf of other faculty in the department: Drs. Mercedes Becerra, Anne Becker, Mary Kay Smith Fawzi, Molly Franke, Bethany Hedt, Ann Miller, Megan Murray, Vikram Patel, Eugene Richardson, Gustavo Velasquez, Norma Ware.

I. The definition of Scientific Data

see attached

II. The requirements for Data Management and Sharing Plans

see attached

Attachment:

Please find a joint comment on proposed NIH requirements for data sharing:

https://osp.od.nih.gov/wp-content/uploads/Data_Sharing_Policy_Proposed_Provisions.pdf#scientific-data

First, some overarching observations about the proposed requirements:

- 1) Any data-sharing requirements and guidance for evaluation by reviewers should promote explicit efforts to reduce inequalities in access to data (in a global health context, this means that plans should promote/guarantee involvement of local researchers; in all contexts, plans should promote access by affected populations).
- 2) The expectations for, and nature of, data sharing should be different when the objective of sharing is reproducibility/validation (as implied by the definition of “scientific data”) compared to when the objective of sharing is for secondary/new research. The former should be nearly universal, relatively unfettered, and should occur at the time of publication. The latter may be more restricted and may occur at a later date, i.e., after all planned analyses have been conducted by investigators of the NIH-funded effort.
- 3) It is difficult to comment on the guidance without seeing evaluation criteria. Our comments should inform guidance as well as development of evaluation criteria.
- 4) Costs (and expertise required) to permit sharing of data are substantial; long-term, public (curated) access entail significant expenditure both during and after the end of the funding period. A mechanism should be considered that supports these efforts above normal grant/contract ceilings. Otherwise there is a risk that those who embrace data sharing completely will effectively have their research penalized financially relative to those investigators who take a more minimalist approach to data sharing.

Specific points:

I. Definitions:

1. Scientific data:

In the definition “The recorded factual material commonly accepted in the scientific community as necessary to validate and replicate research findings including, but not limited to, data used to support scholarly publications,” “replicate” should be changed to “reproduce”. And, this definition does not speak to sharing for purposes of secondary analysis or new research. The remainder of the proposed guidance suggests these broader uses of shared data. As noted above, expectations should be different for these different data-sharing objectives.

How does this definition, and the proposed guidance, apply to qualitative research? Qualitative research studies, by design, are not intended to be replicated, making the sharing of data for this purpose not very meaningful. Also, sharing qualitative data without risking breach of confidentiality is a concern. Removing identifying information from raw qualitative data, e.g. interview transcripts, is not in itself sufficient to eliminate the risk of a confidentiality breach. If the guidance applies to qualitative data, it should clearly specify the form in which qualitative data should be shared to maintain confidentiality.

2. Metadata: This should be more explicitly considered part of scientific data if it’s essential to the reproducibility of published results and interpretation of scientific data. Metadata that makes primary data more understandable/usable should be required as part of data sharing. NIH should consider a definition of metadata similar to that used for replication datasets in the Dataverse project (<https://dataverse.org/best-practices/replication-dataset>), which allows for discoverability of replication datasets as well as a comprehensive “list of code, scripts, documents, and data files that are needed in order to make replication possible.”

II. The requirements for Data Management and Sharing Plans

1. Plans should include efforts to minimize ‘parasitic or parachute research’ in global health (as described in [https://www.thelancet.com/pdfs/journals/langlo/PIIS2214-109X\(18\)30239-0.pdf](https://www.thelancet.com/pdfs/journals/langlo/PIIS2214-109X(18)30239-0.pdf)) and to maximize equity in access to data for local researchers in global health and to affected populations in all types of population research. Plans should include and be evaluated for:

- 1) specific attention to discoverability for these groups;
- 2) commitment to data sharing with local researchers (global health); with other groups whose access is constrained by structural inequalities; and with affected communities.
- 3) plans for training of these groups to enhance ability to use shared data.

2. Timing & scope of access:

The document should provide guidance on the expectations around timeline for sharing data. It would be reasonable to expect investigators to share data used for specific publication promptly upon publication to fulfill validation/reproducibility goals. For broader secondary analyses/new research, the expectation should be less restrictive: investigators should be permitted a period of exclusive use of data that extends beyond the end of the grant period. At some point after the end of the grant period, sharing could be expected to commence. Mechanisms to evaluate requests (for compliance with principles that discourage parasitic/parachute research) should be permitted. These costs should not be part of the allowed budget for grants/contract.

We recommend a difference in expectations for sharing when data are *collected with* NIH funds vs. when data are *analyzed with* NIH funds. For example, if a pre-existing dataset was created without the use of NIH funds, but NIH funds support secondary analyses of these data, the data-sharing expectations should be limited to those necessary to reproduce or validate analyses supported by NIH funds. Secondary analysis of data that is not publicly shareable (for purposes other than validation/reproduction) should still be eligible for NIH funding.

Compliance: Consideration should be given to consequences if results aren’t reproduced or validated. Some form of adjudication should be permitted.

Submission #109**Date:** 12/10/2018**Name:** Chris Bourg**Name of Organization:** MIT Libraries**Type of Organization:** University**Role:** Institutional Official**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

All areas of research under NIH are important to MIT Libraries.

I. The definition of Scientific Data

We support the definition of scientific data proposed. It follows closely on the definition given in the Uniform Guidance (<https://www.federalregister.gov/d/2013-30465/p-834>). We would further suggest expanding this definition to include examples of expected data types, similar to that presented in the DOE Policy for Digital Research Data Management Glossary (<https://www.energy.gov/datamanagement/doe-policy-digital-research-data-management-glossary>), “The term digital data encompasses a wide variety of information stored in digital form including: experimental, observational, and simulation data; codes, software and algorithms; text; numeric information; images; video; audio; and associated metadata. It also encompasses information in a variety of different forms including raw, processed, and analyzed data, published and archived data.” This expansion would have the effect of prompting an expanded view of the types of data necessary for validation and replication, which may have been limited by the reference to “data used to support scholarly publications.” The NIH may also consider eliminating this phrase to enhance a broadened understanding of the importance of sharing negative results that still face challenges to becoming part of the scholarly record through the publication process.

II. The requirements for Data Management and Sharing Plans

We support updating the previous 2003 data sharing plan to include a data management plan. NIH should take every opportunity to align the elements of the plan and the implementation with the goals of the NIH Strategic Plan for Data Science.

We support data management plans being evaluated as an Additional Review Consideration (Section IV, Plan Review and Evaluation – Extramural Grants). We further recommend including

data management plans in the overall grant proposal impact score. This provides NIH an opportunity to make a strong data management statement. This would also embed compliance at the onset. This does, however, place a new burden of responsibility on reviewers to evaluate content that may exceed their familiarity with best data management practices and infrastructure. Including the data management plan in the impact score as we suggest may require substantial training and guidance for reviewers, which ultimately may align with the workforce development goal of the NIH Strategic Plan for Data Science. Strong consideration of the development of this reviewer support is needed.

We support the proposed plan's overall expanded sections and are especially pleased that it addresses timelines, oversight responsibility, and code and software management. That said, we note a few areas where further clarification and infrastructure development is required.

Section IV, Plan Elements – 1. Data Types. The phrasing provided: “and how raw or processed the data will be” is awkward and invites confusion that may result in equally unclear researcher statements on their produced data. While it is necessary to generalize terms in consideration of the many possibilities of research output, the NIH should consider what level of information regarding the extent to which data is processed is helpful or necessary and provide a revised statement that adequately directs researchers to address this need.

Section IV, Plan Elements – 3. Standards. While we support the use of Common Data Elements (CDEs) to promote interoperability across disciplines, the existing resources for researchers to connect with and utilize existing standards and CDEs can be insufficient and confusing, limiting their overall uptake. For example, the CDE Resource Portal as the user interface for the NIH-supported CDEs can be challenging for researchers to navigate, identify, and adopt appropriate CDEs. The NIH should consider a process to connect newly granted research proposals with data management plans mentioning specific CDE's with the authoring IC.

Section IV, Plan Elements – 4. Data Preservation and Access. We would encourage the NIH to consider framing this section from the perspectives of data citation and reuse. Reframing in this way may help researchers to consider the activities involved in data storage and access in a more cyclical and future-motivated way rather than simply fulfilling a requirement. This reframing may require more direct references to practices such as data citation and providing adequate documentation that describes the data for effective reuse.

Section 4.1 references archiving, preservation, and storage in its first two sentences. It is unclear which of these distinct activities researchers might be intended to address, so we would encourage the NIH to consider which of these activities are necessary and practical and provide the clarity that will allow researchers to address these activities responsibly. As most of this document is concerned with issues of data storage and access it may be useful to define the term “preservation” when it is used purposefully.

Section 4.4 asks researchers to provide contingency plans for any number of failure modes. This may be both time and space prohibitive as the recommended limit for the proposed DMP is

two pages. An alternative to this may be that the NIH consider data management and sharing plans to be living, dynamic documents, rather than static. This may enable more meaningful engagement with these plans throughout the research process and support more effective compliance and enforcement as plans may be more reflective of actual practices and less fraught with self-protecting vagaries. Incorporating review of and necessary revisions to DMPs as part of the annual Research Performance Progress Report (RPPRs) process (as stated in Section V) will help move towards making the DMP a continued and active aspect of research.

Section IV, Plan Elements – 5. Data Preservation and Access Timeline. The content of this element addresses issues of storage and access rather than the complexities normally associated with preservation processes. The NIH should consider retitling this element to accurately reflect its content.

Additionally, separating out the timeline for sharing data (5.1) from when data can be shared to be used by others (5.2) is an odd distinction.

Section IV, Plan Elements – 7. Oversight of Data Management. In addition to the active project data management roles outlined, the NIH should consider including post-project completion roles and responsibilities regarding ongoing access to shared data. Data management responsibilities do not end at project completion.

Lastly, we strongly support including a compliance step in the annual Research Performance Progress Report (RPPRs) process (Section V. Compliance and Enforcement). A way to approach easing some of the review burden may be to consider deploying machine-readable document structures for data management and sharing plans. Ultimately, integration with this existing, annual requirement will not only help to ensure timely compliance but will allow researchers to flag and address barriers or necessary changes to their data management strategies early and often.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

Prior to sharing and implementing this updated plan, a plan for providing adequate guidance and training for reviewers and researchers should be developed. Incorporation of the DMP as an Additional Review Consideration provides a compliance and feasibility checkpoint that could be eroded if reviewers aren't given the tools to adequately evaluate this important aspect of a proposal. Additionally, compliance workflows should be established or existing documents (RPPRs) should be updated.

Phasing should also take into account necessary improvements to data infrastructure, resources, and standards. The prevalence of research data generated via NIH grants that require responsible sharing mechanisms and controlled access highlights a particular weakness

in the research data ecosystem. There are few systems that support these procedures adequately, resulting in the research community potentially building habits of not sharing these data even when the infrastructure is eventually adequately developed. Addressing these and other infrastructure shortcomings may require reconsideration of funding provisions for data centers and data storage, accessibility, curation, preservation and archiving resources.

Submission #110**Date:** 12/10/2018**Name:** Mary M. Langman**Name of Organization:** Medical Library Association**Type of Organization:** Professional Org/Association**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

biomedical research and clinical data

I. The definition of Scientific Data

Metadata:

There are many categories of metadata (e.g., descriptive, preservation, technical, structural, administrative). For this reason, MLA recommends that the policy notes the different categories, and provides recommendations about which to leverage in the case of research documentation efforts.

If electronic lab notebooks are called out as not being an example of scientific data, they could be provided as an example of research documentation or metadata that are required to make the data usable.

Scientific Data:

MLA recommends clarifying at what level the data should be shared/deposited: raw data, processed data, and/or analyzed data. If the data is going to be usable, guidelines should also clarify what else must be made available (e.g., software to process the data).

In order to replicate research findings, researchers need the raw data and the processing code or a software citation. Having examples of what this scientific data includes, along with examples of what scientific data does not include, would be helpful for researchers.

Scientific data should NOT be defined just as those datasets that underlie publications, but also include data outputs from research not necessarily published to avoid bias.

II. The requirements for Data Management and Sharing Plans

Sharing Scientific Data:

If there are perceived barriers to sharing scientific data; guidelines should outline what would be acceptable reasons for not sharing.

Plan Review and Evaluation:

NIH should provide guidelines that include details for how the plan review and evaluation will be accomplished. Guidelines could include training for reviewers and/or the development of a rubric for evaluation. Perhaps a more consistent way would be to have plans reviewed by a set group of NIH personnel, possibly from the National Library of Medicine (NLM) or the NIH Library, with the necessary skill sets for evaluating Data Management Plans (DMP) with a score that is integrated into the evaluation of the entire application.

Plan review and evaluation should affect the overall score of an application and affect the overall success of an application. Otherwise, these will not gain traction or importance and remain afterthoughts to the process for some researchers.

Plan Elements:

A two-page limit will restrict much of the required elements from being described in depth (i.e., data types, related tools and software, data standards, data preservation, access (including timelines) and discoverability, terms for re-use and redistribution, limitations on access, and oversight of data management). We recommend that this limit not be put into place.

Budget:

We recommend requiring researchers to provide reasonable costs associated with data management and sharing under the budget for the proposed project to ensure that researchers account for this cost at the outset. It would be helpful for NIH to provide examples of potential costs associated with each component of the Research Data Management (RDM) plan; examples should include links to the costs repositories might charge for storage.

Support Documentation and Guidance:

NIH should make available a portfolio of successful sample applications and include the DMP. The National Institute of Allergies and Infectious Diseases (NIAID) makes sample applications available (<https://www.niaid.nih.gov/grants-contracts/sample-applications#r15>) but their samples do not include data management or sharing plan information.

NIH should acknowledge that DMP are living documents that are likely to change throughout the research process, and provide guidelines about how to update and version plans as they change.

Data Types:

A requirement to list the types and estimated amount of scientific data resulting from NIH-funded or supported research should include an estimate of size as many repositories have size

limits on both file size and total size of data uploaded. This should be spelled out in the policy as researchers may not equate “amount” with “size” and not include this information.

For scientific data derived from human participants or specimens, applicants should briefly describe the process of de-identification or aggregation they plan to use.

Related Tools/Software and/or Code:

The guidelines should require that all created code and scripts be shared alongside the data in order to make it replicable.

While open source software is preferred, all research software should be well documented, and version and packages used need to be included, especially for reproducibility.

Standards:

MLA recommends including guidance for metadata when identifying standards and ontologies that apply to the scientific data to be collected; e.g. suggested data formats, data identifiers, definitions, and other data documentation.

Data Preservation and Access:

In funding announcements, MLA recommends that institutional data repositories are recognized in funding announcements as potentially acceptable solutions for data deposit, and that appropriate discipline-related repositories are detailed for researchers consideration, such as directing them to use the Registry of Research Data Repositories (<https://www.re3data.org/>).

Repository selection should be guided by criteria (<https://www.datasealofapproval.org/en/information/requirements/>) that increase discoverability, guarantee the integrity and authenticity of the data, and have a continuity plan that ensures ongoing access to and preservation of its holdings, rather than the cost of the repository. For this reason, MLA suggests that the following sentence be removed from the provisions: “Investigators would be encouraged to consider using repositories that make scientific data available at no cost for extended periods of use”.

We recommend including a statement in the DMP outlining why a certain repository will be selected. Alternatives/suggestions should be provided by reviewers when appropriate if those listed in the DMP are inadequate/not standard for a discipline.

As proposed in the National Library of Medicine’s Strategic Plan (https://www.nlm.nih.gov/pubs/plan/lrp17/NLM_StrategicReport_Synopsis_FINAL.pdf) linkages between publications and datasets must occur.

If researchers are not able to deposit datasets into a repository for any reason, we recommend creating a metadata record which describes the data. Many health sciences libraries curate research datasets, regardless of where they are stored (varying from a personal server to repositories), and make them discoverable via a data catalog; e.g. several are funded through the National Network of Libraries of Medicine (<https://www.datacatalogcollaborationproject.org/partners/>). Research institutions and NIH should consider this model for insuring the discovery of funded research data especially if the data was not deposited into a repository.

Oversight of Data Management:

We strongly recommend prompting researchers to seek assistance from specialist librarians and information professionals who have expertise in managing data. Librarians are pivotal in facilitating the storage, description, maintenance, preservation, and access to scientific data, and providing expertise, consultations, and outreach to the campus community and others around scholarly resource assessment and metrics (https://www.coar-repositories.org/files/Competencies-for-RDM_June-2016.pdf)

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

Compliance and Enforcement:

MLA supports the provision for developing data management and sharing guidelines that support compliance and enforcement of the policy.

Libraries and librarians should play a leading role in coordinating an institution-wide initiative to educate and support the efforts of researchers and others in compliance and enforcement of data management and sharing policies and procedures.

MLA recommends that NIH, in addition to those definitions already provided, direct researchers to existing resource(s) that define additional terms used in this document related to data management and sharing. An example of an existing resource is the National Network of Libraries of Medicine Data Thesaurus (<https://nnlm.gov/data/thesaurus>).

Submission #111

Date: 12/10/2018

Name: Jason Williams

Name of Organization: Cold Spring Harbor Laboratory

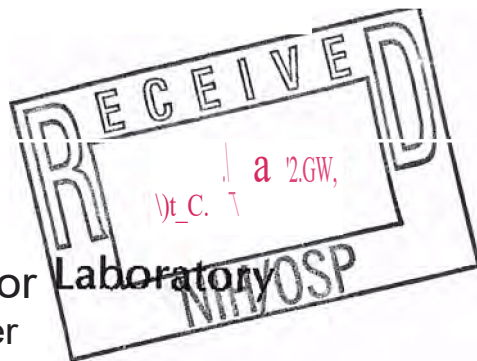
Type of Organization: Nonprofit Research Organization

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

cancer, neuroscience, plant biology, genomics, bioinformatics

Attachment:



Cold Spring Harbor Laboratory
DNA Learning Center

1 Bungfown Road
Cold Sprl11g Horbor, NY 11724
Phone: (516) 367-5170
Fax: (516) 367 182
Internet: www.dnalc.org
Email: dnalc@ahli.edu

Re: Request for Information (RFI) on Proposed Provisions for a Draft Data Management and Sharing Polley for NIH Funded or Supported Research*

December 10, 2018

Dear Colleagues,

The NIH mission is to "seek fundamental knowledge about the nature and behavior of living systems and the application of that knowledge to enhance health, lengthen life, and reduce illness and disability." It is highly appropriate that NIH develop data management and sharing policy cognizant of the challenges (and opportunities) brought about unprecedented rapid development of **data**. Before addressing my specific responses to the Proposed Provisions document I wanted to raise a key issue I have not seen addressed - one which limits NIH's access to the best possible advice,

NIH mechanisms for collecting feedback In the areas of data practice have generated skepticism about the organization's willingness or ability to act upon community recommendations. Since the release of the Nit-I Strategic Plan for Data Science in June 2018, my own personal experience is that investigators specifically from communities of bloinformatics, computational biology, and data science have expressed disappointment that comments made did not result in meaningful improvement In the final plan. Clearly, NIH action should not be based on my personal anecdotes. However, while NIH has been diligent about soliciting comments (and crucially, posting them publicly) there Is no sense from investigators I've talked to, or from a reading of NIH announcements, Just how final decisions will be made. In the absence of a more transparent process, the most cynical reading of this exercise is that NIH can ignore community recommendations and proceed to follow a pre-determined agenda. If investigators doubt NIH's commitment, many of the comments NIH needs to hear most will never be offered. I speak of NIH as a whole because even though actions of individual institutes may be their own, they all contribute to the community's judgment of faith and good will In these processes.

My *primary recommendation* is that NIH go beyond disclosing the comments made to this and other RFIs by also making available to the community a public and complete explanation of how recommendations are acted on and for what causes. For several reasons elaborated ir, my previous response⁰ NIH is at a 'structural' disadvantage In developing data-related policies. Ideally NIH would suggest a reasonable implementation plan ahead of an RFI describing how recommendations would be vetted and implemented and follow through on their commitment to be responsive to community recommendations.

Unfortunately, large-scale initiatives to support the data and computational underpinnings of NIH-sponsored science have not had a track record at success (e.g. caBIG, BD2K, etc.). I still have high-hopes for the Data Commons (if it can progress towards being community-driven for the benefit of NIH, not NIH-entangled for the benefit of no one). Without doing more to be responsive and open, NIH risks becoming irrelevant as a policy shaper and will fade into a reactionary position as data generated domestically and globally dwarf projects directly funded by NIH. I hope NIH decides to think about how think outside the box (or the NIH oampus) to build community consensus through a more transparent and accountable process. I think I see thing happening and hopeful for further assurance.

SPECIFIC COMMENTS

The definition of Scientific Data

- Software is a concept missing from the definition of scientific data. In one sense, software is a form of metadata. For example, in a wet-lab experiment, a description of a cell line or antibody lot might be a crucial descriptor of a dataset. Often, scientific data is uniquely wedded to the software used to produce that data - from base calling software which may be involved in the production of "raw" sequence data, to the software used to produce any of several downstream analysis products. The definition of scientific data could benefit from explicit acknowledgment of this unique relationship.
- Related to software, provenance is also a concept missing from the proposed plan. Scientific data is characterized by its life cycle (see one elaboration at <https://www.dataone.org/data-life-cycle>). In particular, the need for constant versioning and updating merits reflecting on this in the definition of "scientific" data as a dynamic concept.

The requirements for Data Management and Sharing Plans

- *Regarding "Related Tools Software and/or Code":*
 - For any software/code used (but not developed) in an analysis, it should be a requirement that a full description of the software be provided including version numbers, links to source code/binaries, etc.
 - As indicated, the NIH should strongly encourage the use of open-source software that is freely available. As funders and institutes are migrating to all open-access publication, it may be worthwhile to consider how NIH can progress to implementing requirements for the usage of open software and data formats.
 - I agree with the recommendation that where proprietary software must be used, there is explanation provided.
 - It should be a further recommendation that software used in analysis be fully-documented for example by making available version-controlled scripts, makefiles or workflow language descriptions, etc. Where possible, investigators should make use of modern reproducibility approaches such as containers (e.g. Docker, Singularity), virtual machine images, etc. Documentation should follow recommendations that increase the reproducibility of analysis such as minimum information standards (e.g. minimum reporting guidelines for biological and biomedical investigations; <http://www.nature.com/nbt/journal/v26/n8/full/nbt.1411.html>, or more updated recommendations being produced by GA4GH, Research Data Alliance, etc.).
 - Where any scripts or other software is developed as part of an investigation, this must be accompanied by an appropriate open-source license and available in a public repository upon submission of any pre-print and/or by the time of submission for peer-review. Any code/software should be available by the end of funding regardless of publication status. The same recommendation on version controls and containerization apply here as well.
- *Regarding "Data Preservation and Access":*
 - There is no comment on how investigators should address data that does not need to be kept. There should be a clear description of how investigators determine what intermediate or derivative data products do not require preservation. There may also need to be specific

recommendations for documenting how sensitive data will be discarded - but these may be sufficiently addressed by other legal and policy requirements patient data.

- a Although unique identifiers are addressed, there should be more specific guidance on where identifiers should be required, and what defines appropriate long-term storage solutions.
- o There should be a description of how sharing will be achieved for large (> 1GB?) datasets. For very large data sets, sharing becomes increasingly difficult. While there may not be an obligation for the researcher to make every dataset equally available, it should be possible to characterize (and perhaps NIH could implement a scoring system) that allows classification of data sharing. FAIR metrics projects being developed within the NIH Data Commons and elsewhere are already working on these objectives.

- *Regarding "Data Preservation and Access Timeline":*

- o That data funded by NIH (with the exception of protected records) must be made available should be made explicit.


The optimal timing, including possible phased adoption, for NIH to consider In Implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards.

- I agree with the document's suggestion that for extramural grants, the DMP could be evaluated as acceptable/unacceptable and as part of an Additional Review Consideration.
- The NIH needs to invest in training and/or learning materials on how to effectively critique a data management plan. In general, it is likely that a substantial number of reviewers in a study section or other panel (or program officers) may not have training in the practices of computational sciences or data management. As such, they may not be the most effective adjudicators of a data management plan. NIH could increase the quality of its review by offering training. Groups such as Data Carpentry, DataONE, etc. have training material/curriculum that could be applicable here.
- The NIH should invest in training, including the development of learning resources in order to assist investigators with development and execution of data management plans.

Additional Recommendations

- The data management/policy landscape is very dynamic. NIH should implement an annual review of its guidance.
- Guidance on the data management policy should directly solicit advice from recognized organizations including the Research Data Alliance, GA4GH, ELIXIR, Force11 and others. A formal advisory mechanism here may be appropriate.

Sincerely,



Jason Williams

Old Spring Harbor Laboratory

- This response represents my only my own personal opinions

- See my comments on the NIH Data Plan and that of other community members here: https://github.com/JasonJWilliamsNY/2018_nih_datascience_rfi

Submission #112**Date:** 12/10/2018**Name:** Taneisha Wilson**Name of Organization:** Society of Academic Emergency Medicine and Academy of Emergency Physicians**Type of Organization:** Professional Org/Association**Role:** Scientific Researcher**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Clinical Research, epidemiology

I. The definition of Scientific Data

- Access to the data should be contingent on evidence of a plan for responsible use by the party who wants it, including a written protocol that has been approved or exempted by a valid IRB.
- o This should help mitigate any conflicts that exist for commercial/policy interests who might massage data to obtain biased results.
- For researchers accessing human subjects research it would be prudent to have confirmation of specific training regarding the ethics of using human subjects' data.
- Similarly, we suggest that you provide guidance for language in informed consents pertaining to how the data will be stored, archived, and accessed by those other than the researcher.
- Patient consent relating to specific disease topics should include disclosure regarding data sharing.
- o For example, if genomic data is shared for asthma, using that data for other disease processes may outside of the initial consent if not specified.
- Original researchers should have the ability to screen and deny requests if it seems out of the realm of the original consent.

II. The requirements for Data Management and Sharing Plans

Decreasing Researcher Burden

- Consider excluding the Plan from the initial application and requiring the Plan as a Just-In-Time addition or a post funding addition.
- More expansively, the IC should work with the applicant once funded to develop the Plan for implementation.
 - o These changes will limit the burden of unnecessary paperwork for applicants as well as provide flexibility and customization of the Plan to the specific IC.
- The proposal mentions “free” services for archiving data. However, this seems unrealistic for many projects funded by the NIH.
 - o Any archiving of data must consider the possibility of data breaches and should therefore preferentially utilize vetted storage solutions using deidentified data, ideally through a federally managed archiving system.
 - o The ICs should include a supplement for implementing the Plan if costs are associated.
- Overall, a streamlined form for how and when data sharing will be enacted would guide researchers and increase compliance. We strongly recommend a document which functions as a tool kit for researchers.
- Providing clean data, a clear data dictionary, and reviewing requests, etc. is a large undertaking.
 - o Data files with large amounts of data can have a very steep learning curve and can be easily botched by a new, external investigator – despite their best intentions.
 - o There are substantial costs for sharing data but also for trouble-shooting the problems that emerge when data is misinterpreted or misused.
 - o Please outline specifically who will provide the COST of sharing data (the original researcher, or the requesting researcher.)
 - o Statistical support from specific ICs may be required for large cumbersome datasets.
- Recipients of federal grant should have to upload a deidentified .CSV database on a no-cost archiving system after publication of the major findings and filing of intellectual property protection.
- While the data should be available to the public, a mechanism should be in place to allow investigators to know who accessed their data and why.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and

- The deadline for this sharing should be project specific based on the data collected since project timelines and analyses are project specific.
- o Within 3 years from data lock might be reasonable for many projects, though this will necessarily vary based on the study.
- o The proposed time period should be included in the initial grant application and included as a scoring criterion. Mechanisms to request extensions should be made available.

Attachment:

NIH RFI: Data Sharing and Management, Emergency Medicine Researchers Respond

Responsible data sharing, and management is an integral part of scientific inquiry. As emergency medicine researchers with NIH funding, we commend the NIH for thoughtfully compiling the current proposal. We acknowledge the initial effort to decrease researcher burden. Other positive aspects of the current document include the suggestion to shift some of the burden of data storage and management to an IC task instead of solely on the researcher. These highlights are only a few noteworthy aspects of the proposal. However, if this proposal is to move forward, we have outlined some amendments and additions for your consideration.

Decreasing Researcher Burden

- Consider excluding the Plan from the initial application and requiring the Plan as a Just-In-Time addition or a post funding addition.
- More expansively, the IC should work with the applicant once funded to develop the Plan for implementation.
 - These changes will limit the burden of unnecessary paperwork for applicants as well as provide flexibility and customization of the Plan to the specific IC.
- The proposal mentions “free” services for archiving data. However, this seems unrealistic for many projects funded by the NIH.
 - Any archiving of data must consider the possibility of data breaches and should therefore preferentially utilize vetted storage solutions using deidentified data, ideally through a federally managed archiving system.
 - The ICs should include a supplement for implementing the Plan if costs are associated.
- Overall, a streamlined form for how and when data sharing will be enacted would guide researchers and increase compliance. We strongly recommend a document which functions as a tool kit for researchers.

Data Sharing

- Providing clean data, a clear data dictionary, and reviewing requests, etc. is a large undertaking.

- Data files with large amounts of data can have a very steep learning curve and can be easily botched by a new, external investigator – despite their best intentions.
- There are substantial costs for sharing data but also for trouble-shooting the problems that emerge when data is misinterpreted or misused.
- Please outline specifically who will provide the COST of sharing data (the original researcher, or the requesting researcher.)
- Statistical support from specific ICs may be required for large cumbersome datasets.
- Recipients of federal grant should have to upload a deidentified .CSV database on a no-cost archiving system after publication of the major findings and filing of intellectual property protection.
- The deadline for this sharing should be project specific based on the data collected since project timelines and analyses are project specific.
 - Within 3 years from data lock might be reasonable for many projects, though this will necessarily vary based on the study.
 - The proposed time period should be included in the initial grant application and included as a scoring criterion. Mechanisms to request extensions should be made available.
- While the data should be available to the public, a mechanism should be in place to allow investigators to know who accessed their data and why.

Ethical Data Use

- Access to the data should be contingent on evidence of a plan for responsible use by the party who wants it, including a written protocol that has been approved or exempted by a valid IRB.
 - This should help mitigate any conflicts that exist for commercial/policy interests who might massage data to obtain biased results.
- For researchers accessing human subjects research it would be prudent to have confirmation of specific training regarding the ethics of using human subjects' data.

- Similarly, we suggest that you provide guidance for language in informed consents pertaining to how the data will be stored, archived, and accessed by those other than the researcher.
- Patient consent relating to specific disease topics should include disclosure regarding data sharing.
 - For example, if genomic data is shared for asthma, using that data for other disease processes may be outside of the initial consent if not specified.
- Original researchers should have the ability to screen and deny requests if it seems out of the realm of the original consent.

Submission #113**Date:** 12/10/2018**Name:** Audie Atienza, PhD (representing ICF generally)**Name of Organization:** ICF**Type of Organization:** Other**Other Type of Organization:** Global Consulting Organization**Role:** Scientific Researcher**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

ICF provides the federal government with support for data collection, data curation, data analysis, and data dissemination. ICF has expertise in various health research domains including: population health survey, local area health surveys, clinical trials registry data, cancer registry data, public health/epidemiology data, HIV/AIDS information, Genetic and Rare Disease information, electronic health record data, environmental/behavioral aspects of health, and common data elements repositories.

I. The definition of Scientific Data

NIH states:

“Scientific Data: The recorded factual material commonly accepted in the scientific community as necessary to validate and replicate research findings including, but not limited to, data used to support scholarly publications. Scientific data do not include laboratory notebooks, preliminary analyses, completed case report forms, drafts of scientific papers, plans for future research, peer reviews, communications with colleagues, or physical objects, such as laboratory specimens. For the purposes of a possible Policy, scientific data may include certain individual level and summary or aggregate data, as well as metadata. NIH expects that reasonable efforts should be made to digitize all scientific data.” (https://osp.od.nih.gov/wp-content/uploads/Data_Sharing_Policy_Proposed_Provisions.pdf)

ICF’s Comment: GSA defines “data” as “recorded information, regardless of form or the media on which it may be recorded. The term includes technical data and computer software.” (see: https://www.acquisition.gov/far/current/html/Subpart%2027_4.html) As the proposed NIH data management and sharing plans list “software” as an element in the plan, NIH may wish to consider including “software” in their definition of data as well.

II. The requirements for Data Management and Sharing Plans

A) The Data Sharing Policy Proposed Provisions states:

“Plan Elements: Plans could have a two-page limit and address the following research elements: (i) data types, (ii) related tools and software, (iii) data standards, (iv) data preservation, access (including timelines) and discoverability, (v) terms for re-use and redistribution, (vi) limitations on access, and (vii) oversight of data management. Examples of guidance about how these Plans could be implemented are included below:

2. Related Tools, Software and/or Code: Indicate what software/computer code will be used to process or analyze the scientific data (the inclusion of scripts may be helpful), why the software/code was chosen, and whether it is free and open source. If software/code that is not free and open source is needed to access or further analyze the scientific data, briefly describe why this particular software/code is needed. Describe whether there is an alternative free and open source software/code that may be used to further analyze the scientific data.”

ICF’s Comment: It would be useful for future researchers who wish to re-analyze existing data or replicate findings to know if certain software/computer code and/or hardware is required to process or analyze the data, not just why it was chosen or free.

B) With respect to contracts, the Data Sharing Policy Proposed Provisions states:

“IV. Requirements for Data Management and Sharing Plans

Plan Review and Evaluation: The funding or supporting NIH IC would consider the evaluation and determine the acceptability of the Plan, which could be implemented in a variety of ways, including:

Contracts: Plans could be included as part of the technical evaluation performed by NIH staff and incorporated in the subsequent terms of the contract.”

AND

“V. Compliance and Enforcement

During the Funding or Support Period

Compliance with the Plan would be determined by the funding or supporting NIH IC and reviewed at a minimum, annually [e.g., at the time of annual Research Performance Progress Reports (RPPRs)].

Contracts: The Plan would become a Term and Condition of the Award, and compliance with and enforcement of the Plan would be consistent with the award and the Federal Acquisition Regulations (FAR),¹⁸ as applicable.”

ICF's Comment: NIH should reference FAR Subpart 27.4—Rights in Data and Copyrights (https://www.acquisition.gov/far/current/html/Subpart%2027_4.html). This regulation provides the reader with the major data right types relevant to contracts, which may be relevant or applicable to the proposed data sharing policy.

C. Data Preservation and Access - Data Disposition

ICF's Comment: If scientific data posted to a federal government repository(ies) [see the section: V. Compliance and Enforcement] represents a federal record as defined by the Federal Records Act [see: <https://www.archives.gov/about/laws/fed-agencies.html>], the preservation of that data would be governed by federal regulation.

Thus, NIH may wish to note that “Data preservation must be consistent with the NIH Strategic Plan for Data Science” along with other federal regulations, as appropriate.

NIH may also wish to have researchers prepare as part of the Data Preservation and Access plan “data/biospecimens disposition” statements, like the NCI NCTC Banks [see: <https://nctnbanks.cancer.gov/biospecimens-access/disposition.html>].

Submission #114**Date:** 12/10/2018**Name:** Carl McKinley**Name of Organization:** Regenstrief Institute**Type of Organization:** Nonprofit Research Organization**Role:** Institutional Official**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Clinical, genomics, social determinants, retrospective and prospective health research, public health, policy

II. The requirements for Data Management and Sharing Plans

Plan review and evaluation:

Plans should not be factored into the overall impact score through the peer review process.

- Data sources, especially clinical data sources, often have unique contractual restrictions with regards to sharing individual level data with third parties. A funding or supporting NIH IC which requires individual data sharing may put access to clinical data sources in jeopardy for future research proposals.
- Recognize that an IRB may not approve waivers of authorization for individual level data to be stored or shared in perpetuity.

Plan Elements:

Data Preservation and Access:

- Similar to our comment above, clinical data sources may restrict access to their data if individual level data is required to be shared. Clinical data sources may amend access agreements to restrict NIH funded or supported research projects from clinical data sources.
- If the Plan requires sharing protected health information, Current PHI cybersecurity best practice is to ensure the return or destruction of health information once the approved use is complete. NIH should consider allowing time limits to be placed on Plan (for example, the study data containing PHI shall be placed in a repository for restricted or enclave access for 10 years and then destroyed).

Submission #115**Date:** 12/09/2018**Name:** Matthew Trunnell**Name of Organization:** Fred Hutchinson Cancer Research Center**Type of Organization:** Nonprofit Research Organization**Role:** Institutional Official**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Our focus areas are treatment and prevention of cancer and infectious diseases.

I. The definition of Scientific Data

As written the definition covers only data generated in support of a research question, not data generated for the purposes of enabling new research questions. Examples of the latter would include the Cancer Genome Atlas, the 1000 Genomes project and other efforts to create data sets for reference and discovery. While these large-scale efforts have made use data coordinating centers with explicit mandates around sharing, the general data sharing policy should anticipate smaller-scale efforts focused on data generation with goal of creating a data set and not just the "data exhaust" of conventional research activities.

II. The requirements for Data Management and Sharing Plans

I applaud the inclusion of explicit review and evaluation of data management and sharing plans.

While the overall document makes reference to FAIR standards, these are not represented explicitly in the requirements for the data management and sharing plan. Significant progress has been made by the community in generating quantitative metrics for FAIRness (see, for example, <https://fairshake.cloud/project/>). While there are not yet community standards for metrics relevant to all data types, it is reasonable (and desirable) to ask of data management and sharing plans that they assert the FAIR metrics against which they propose to have the plan assessed. The document does make reference to "community-based standards" in regards compliance _to_ the plan. The requirements _of_ the plan should also consider community-based standards, which might be manifest in rubrics for assessing FAIRness.

The suggestion in section 2 that submitters define the software tools that will be used in advance of any data generation seems poorly conceived. The software landscape in bioinformatics evolves rapidly and much of that is driven by the availability of data rather than

the anticipation of data. Better to suggest that the plan include descriptions of how information about software and tools will be shared, suggesting, for instance, the use of computational "notebooks" (e.g., Jupyter), and/or use of community standards for describing workflows (e.g., CWL, WDL) and repositories for software that will ensure that others have access.

Section 4.2 introduces the concept of using a persistent unique identifier or other standard indexing tools to make data discoverable. This could be tied more explicitly to the FAIR framework by casting this in the context of making data "findable". In general the plan requirements should refer to FAIR concepts to underscore the importance to NIH asserted in the introduction to the document. Indeed, the body of the document largely ignores the FAIR framework.

The suggestion in section 7 that a data manager/data steward be assigned is a good one. A data steward should be named in advance, and if that responsibility shifts during a project, NIH should be notified. As we move into a time of more intentional data sharing, the community of project-level or institutional data stewards will become more important and NIH may have occasion to interact with this community directly.

The use of the phrase "scientific data archiving" is confusing given the context which suggests that this would make data "available ... for extended periods of use." In the IT context, "archiving" refers to cold storage of data, from which data would have to be retrieved to be used. It may be more appropriate to discuss "long-term hosting" of scientific data.

NIH might consider placing emphasis on making data citable as part of the data management and sharing plan.

Submission #116**Date:** 12/10/2018**Name:** James M. Musser, MD, PhD**Name of Organization:** Federation of American Societies for Experimental Biology (FASEB)**Type of Organization:** Professional Org/Association**Role:** Scientific Researcher**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Biological and Biomedical Research

I. The definition of Scientific Data

Definition of Scientific Data: FASEB appreciates that “Scientific Data” as defined in the proposed provisions explicitly excludes items such as laboratory notebooks, preliminary data, case report forms, draft manuscripts, and physical specimens and emphasizes the need for access to the data underlying publications. However, we encourage slight expansion of this definition to recognize the impact of negative results that may be excluded from publications. As noted in FASEB’s 2016 recommendations to enhance research reproducibility, transparency regarding experiments not yielding positive results is also critical to scientific knowledge. Thus defining scientific data as all findings, both positive and negative, contributing to a line of research inquiry ensures transparency of the underlying data, thus contributing to the rigor and reproducibility of final published work.

II. The requirements for Data Management and Sharing Plans

Breadth of Proposed Requirements for Data Management and Sharing Plans: One lesson that can be gleaned from the implementation of data management and sharing plans at other federal agencies is that an open-ended requirement for data management and sharing plans will not yield the desired result of information exchange and data re-use. Therefore FASEB recommends that NIH work with the stakeholder community to develop a framework for data management and sharing plans that is flexible and adaptable to the breadth of research activities supported by NIH. To facilitate development of appropriate plans, we encourage NIH to develop supplemental resources and guidance for the information sought in the data management and sharing plans, including a form that balances free text with check box responses, and example forms that demonstrate plans that meet NIH’s expectations for reporting versus those that would not fulfill agency requirements. We also recommend that

NIH conduct a pilot implementation of the policy for a random sample of grants prior to final rollout to ensure templates and guidance documents are clear and lead to the development of appropriate data management and sharing plans.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

Phased Implementation of Requirement: FASEB recognizes that the requirement for data management and sharing plans will aid NIH in its broader efforts to demonstrate proper stewardship of federal funds. We also recognize that there will be unforeseen challenges as NIH proceeds with implementation of a new data management and sharing policy. Therefore, we encourage consideration of a tiered approach for implementing any final policy to both the extramural and intramural research communities to ensure preparedness for fulfilling requirements, making course corrections, and fostering community compliance.

FASEB understand that this RFI represents the first step in a longer journey to increase access to scientific data resulting from NIH funding, and we appreciate NIH's willingness to engage the scientific community in the development of its data management and sharing policy. We encourage continuation of this active engagement, such as through RFIs, public meetings, or even a designated working group or task force, to ensure feasibility of and community support for the final plan.

Attachment:



Proposed FASEB comments in response to NOT-OD-19-014, “Request for Information (RFI) on Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research”

Comments submitted electronically via [RFI website](#) on December 10, 2018

The Federation of American Societies for Experimental Biology (FASEB) appreciates the opportunity to provide comments in response to [NOT-OD-19-014](#), RFI on Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research. FASEB is comprised of 30 scientific societies, collectively representing over 130,000 biological and biomedical researchers who produce and use a wide variety of data, core data resources, and analytic tools.

In reviewing the [proposed provisions](#), we found that the cross-cutting recommendations made in our [comments](#) on NIH’s draft Strategic Plan for Data Science and the guiding principles highlighted in our 2016 [Statement on Data Management and Access](#) are also applicable to this RFI. FASEB also recognizes that this is the first of several steps in the implementation of an NIH-wide data management and sharing policy. However, one overarching concern that arose throughout our deliberations was variability in terms of individual investigators’ expectations, experience, and resource needs to ensure key data from NIH funded projects are consistent with the FAIR (Findable, Accessible, Interoperable, and Re-usable) data principles. In addition, it will be necessary for the final policy to strike a fine balance between data accessibility and administrative effort and not serve as a deterrent to those seeking to share or reuse datasets. Below we highlight specific concerns that need to be addressed before finalizing a data management and sharing policy for NIH funded research.

Definition of Scientific Data: FASEB appreciates that “Scientific Data” as defined in the proposed provisions explicitly excludes items such as laboratory notebooks, preliminary data, case report forms, draft manuscripts, and physical specimens and emphasizes the need for access to the data underlying publications. However, we encourage slight expansion of this definition to recognize the impact of negative results that may be excluded from publications. As noted in FASEB’s 2016 [recommendations](#) to enhance research reproducibility, transparency regarding experiments not yielding positive results is also critical to scientific knowledge. Thus defining scientific data as all findings, both positive and negative, contributing to a line of research inquiry ensures transparency of the underlying data, thus contributing to the rigor and reproducibility of final published work.

Breadth of Proposed Requirements for Data Management and Sharing Plans: One lesson that can be gleaned from the implementation of data management and sharing plans at other federal agencies is that an open-ended requirement for data management and sharing plans will not yield the desired result of information exchange and data re-use. Therefore FASEB recommends that NIH work with the

stakeholder community to develop a framework for data management and sharing plans that is flexible and adaptable to the breadth of research activities supported by NIH. To facilitate development of appropriate plans, we encourage NIH to develop supplemental resources and guidance for the information sought in the data management and sharing plans, including a form that balances free text with check box responses, and example forms that demonstrate plans that meet NIH's expectations for reporting versus those that would not fulfill agency requirements. We also recommend that NIH conduct a pilot implementation of the policy for a random sample of grants prior to final rollout to ensure templates and guidance documents are clear and lead to the development of appropriate data management and sharing plans.

Phased Implementation of Requirement: FASEB recognizes that the requirement for data management and sharing plans will aid NIH in its broader efforts to demonstrate proper stewardship of federal funds. We also recognize that there will be unforeseen challenges as NIH proceeds with implementation of a new data management and sharing policy. Therefore, we encourage consideration of a tiered approach for implementing any final policy to both the extramural and intramural research communities to ensure preparedness for fulfilling requirements, making course corrections, and fostering community compliance.

FASEB understand that this RFI represents the first step in a longer journey to increase access to scientific data resulting from NIH funding, and we appreciate NIH's willingness to engage the scientific community in the development of its data management and sharing policy. We encourage continuation of this active engagement, such as through RFIs, public meetings, or even a designated working group or task force, to ensure feasibility of and community support for the final plan.

Submission #117**Date:** 12/10/2018**Name:** American College of Radiology**Name of Organization:** American College of Radiology**Type of Organization:** Other**Other Type of Organization:** Non-profit physician professional association**Role:** Medical Provider**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Medical imaging

I. The definition of Scientific Data

In general, the American College of Radiology (ACR) finds the definition of scientific data employed by NIH in this proposed plan to be acceptable. There is reference however to one specific term that is unclear. The proposed plan specifically names “completed case report forms” as an example of work that is not scientific data. ACR generally uses electronic data capture (EDC) systems to collect data from sources and therefore the completed case report forms are inherent in the database populated through EDC. In order to share case-level data it would seem data collected through EDC would need to be shared as completed case report forms. However, we would agree that completed case report forms which are source documents used to support data entry into a structured database should not be included as scientific data.

II. The requirements for Data Management and Sharing Plans

There need to be clear boundaries on what is included in the Data Management and Data Sharing Plan in order to minimize redundancy with other research documents and requirements. In this proposed policy, data management should refer strictly to the handling of data for the purpose of data sharing and should not include specific measures taken to ensure quality of data, processing of data for scientific endpoints, or the analysis of such data. These aspects are typically outlined in the study protocol (if applicable) or other study guidance documents. More specifically, Plan Element #2, as proposed, would require the inclusion of a description of the use of tools that would be employed to analyze medical images collected as part of the research project and would also seem to include statistical analysis tools to be used

by the researchers; these methods and tools are not relevant to the stated intention of this proposal (i.e.; ability to promulgate data for sharing across the broader scientific community) and are already addressed in other research study documents which can be provided as part of the data sharing process.

The ACR has collected clinical images for research trials for two decades and in the process the ACR has developed tools to assist in the collection, dissemination, analysis and archival of images. As a result of our experience we have gained significant insight into some of the complexities associated with these processes, to include:

- Lack of a consistent anonymization/de-identification method across the research community and the challenges associated with AI which may result in the ability to identify subjects previously thought to be anonymized/de-identified
- Methods required to anonymize/de-identify all sources of images given variability of modalities and vendor methods
- Appropriate indexing of clinical images with other clinical metadata, especially when data may be sourced from different institutions at different points in time
- Creating searchable archives which enable researchers to identify the most appropriate cases for specific research objectives, both within an existing data set and across multiple datasets
- Growth in the physical size of the datasets both in terms of quantity of images and the size of the imaging files. This has resulted in an increase in the cost of data transfer and the time required to transfer data and has encouraged researchers to develop methods to introduce analytic tools into the data hosting environment. Data sharing plans and NIH funding to support such plans should include flexibility to determine cost-efficient methods for sharing large clinical imaging archives and should promote the use of federated archive models.

To achieve the vision of this proposal, NIH must therefore be prepared to fund the cost associated with the creation of image data archives which are scalable and extensible and which reflect sufficient flexibility to promote efficient use of the image archives. Solutions such as the NCI Imaging Data Commons will contribute to such a solution but funding to support the transfer of such data and/or to create federated archive models will also be essential.

While the scope of this proposal focuses on the roles and responsibilities of the researchers, NIH can further enhance the quality and scope of shared datasets and minimize the cost to share data by helping to address the need for increased standardization across the community related to patient consent to share data, IRB considerations, de-identification procedures, and methods to create limited data sets.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible

phasing could relate to needed improvements in data infrastructure, resources, and standards

ACR has no comments on this topic at this time

Submission #118**Date:** 12/10/2018**Name:** Anonymous**Name of Organization:****Type of Organization:** University**Role:** Scientific Researcher**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

basic biomedical science (non-clinical i.e. no contact with patient population)

II. The requirements for Data Management and Sharing Plans

The living document that describe the FAIR principles for data imposes a very high bar; the data sets should be human and machine readable and this standard should be met at many levels (i.e., many languages) and on all computer architectures. These goals are laudable and eventually achievable, but for much of scientific data we are not there yet.

Plans should be mandatory for data collected with established methodology (e.g., methods accepted within the community for > 5 years). For example, if an investigator makes use of an established method and proposes to use this method to investigate a question then the plan should be drafted before new data is collected and should include: (i) the collection, (ii) analysis, (iii) archiving and curation and (iv) sharing data workflows. Most investigators can describe (i) and (ii) albeit they may be less familiar with the machines and formats that are used to store, visualize and ensure the robustness of their data. However, most scientists have minimal or no training on items (iii) and (iv) and will need assistance in this respect if their data sets are to start to approach FAIR guidelines. This assistance requires new tools that curate, aggregate and archive data as it is being collected. In addition to the consensus (by investigators) on what needs to be reported and shared. I suggest the plans should be reported for established methods, not new technologies as the data garnered for new technologies is still being researched. This also allows the inventor time to capitalize on their new techniques. I make one further suggestion that when experimental data is obtained through the direct sacrifice of mammals then a comprehensive plan that tackles these work-flows should be given highest priority than data that did not need the sacrifice of mammals. I state this because of the ethics associated with the use of animals and the need to maximize the knowledge gained with such inputs. This promotes good stewardship of data, and of the animals used to create the data.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

See some discussion above. There is no optimal timing here the scientific community for the most part is not prepared for DMPs, resources will be needed to do this. These resources (beta versions) will need to be tested by peers. Once data management plans for each method are drafted and tested, the plans should be reported (published) for others to use and build upon. Here one can envisage a new journal (part of journal) that could publish these plans. In this way they can be reused by the community, and the process (in principle) should require less innovation and become routine.

Submission #119

Date: 12/10/2018

Name: Catherine Luria

Name of Organization: Laboratory of Systems Pharmacology, Harvard Medical School

Type of Organization: University

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

systems pharmacology

Attachment:

**Response to “Proposed Provisions for a Draft NIH Data Management and Sharing Policy”
on behalf of the Laboratory of Systems Pharmacology, Harvard Medical School**

Peter Sorger

Director and P.I. of the Laboratory of Systems Pharmacology

Head of the Harvard Program in Therapeutic Science

Otto Kraye Professor of Systems Pharmacology, Department of Systems Biology

Laura Maliszewski

Executive Director of the Laboratory of Systems Pharmacology

Catherine Luria*

Scientific Program Manager, Laboratory of Systems Pharmacology

Alyce Chen

Scientific Program Manager, Laboratory of Systems Pharmacology

Jeremy Muhlich

Director of Software Engineering, Laboratory of Systems Pharmacology

Caroline Shamu

Scientific Director for HMS Research Cores and Technology

Director ICCB-Longwood Screening Facility

Assistant Professor, Department of Radiology (MGH)

Assistant Professor, Department of Biological Chemistry & Molecular Pharmacology

*For further information, please contact Catherine Luria (catherine_luria@hms.harvard.edu).

Comments on Section II: The requirements for Data Management and Sharing Plans

Over the past five years, we have been involved in multiple large NIH/NCI grants that involve data sharing and management activities, including the NIH Library of Integrated Network-Based Cellular Signatures (LINCS) and Illuminating the Druggable Genome (IDG) programs and the new NCI Human Tumor Atlas Network (HTAN) program. We therefore have considerable experience in implementing such activities from the perspective of practicing scientists and NIH grantees. We have commented on this topic in a perspective in *Science Translational Medicine*¹ and written multiple papers attempting to improve the reproducibility of one important type of data: preclinical assays of drug response²⁻⁴.

Overall, we are highly supportive of the development of a more robust set of policies and associated infrastructure for data management and sharing that conforms with FAIR principles. However, we believe that this can only be accomplished as a part of a multi-part, multi-year strategy that also includes: (i) education – including education of graduate students and fellows, (ii) much more substantial investment in innovation and in computational tools and

approaches, (iii) development of infrastructure for validating, storing and disseminating diverse types of data and (iv) incentives for timely and useful data deposition as opposed to simple mandates and penalties associated with specific data types. It is essential to recognize that meeting FAIR standards will not be cheap, and that the annotation and re-use of heterogeneous data arising from perturbational studies (the vast majority of the mechanism-oriented research in the NIH portfolio) is fundamentally different from storing and disseminating a single type of data on a steady-state sample (e.g. a genome sequence).

In many cases however, formats and reporting standards have not yet been developed. This is true of most types of microscopy data, the many variants of mass spectrometry, multiplex immuno-assays on cell and tissues lysates or on components of the microenvironment and emerging data types such as spatial transcriptomics or multiplex imaging. True innovation will be required to adequately annotate experiments in which these types of data are collected over time following genetic or drug-mediated perturbation of a system.

None of this should discourage us from moving towards better data standards, but extensive work will be required for data types (and even file formats) for which no standards currently exist. Existing standards must be more actively supported. For example, the OMERO image management standard that we helped to develop over a decade at MIT (and is now in widespread use) has never received any NIH support despite multiple attempts. The entire development team was moved from the US to the UK, where it is now headed by Jason Swedlow with EU/UK funding. We might support these sorts of activities in the future. NIH-led development of such infrastructure as well as community education before new requirements are implemented will ensure that PIs provide realistic data plans that can be implemented successfully once a grant is funded.

1. Data type:

- Controlled vocabulary for data types should be provided for those who are writing and submitting NIH grants that will be subject to this requirement. Furthermore, definitions are needed for describing how the data will be shared (e.g. individual vs. aggregated vs. summarized). Standardized definitions for data-related terms will make these data plans more consistent, easier for grant applicants to prepare and easier for reviewers and program officers to evaluate and enforce.
- Guidance is required on minimum standards for the types of materials that should be shared along with scientific data. Standardized formats for study protocols and information about data collection instrumentation are essential for data Interoperability.
- Efforts to standardize data types and repositories should be international. Particularly in the area of pre-clinical, basic research data, EMBL/EBI is well ahead of anything in the US.

2. Related Tools, Software and/or Code:

- Standards for specifying information about software and code should be provided. In particular, more information is needed to guide investigators on grants that are data

intensive or that include significant software and computational tool development, for whom this requirement is substantial.

- The suggestion to include scripts may be unreasonable at the grant proposal stage as these are often developed in the course of the study and investigators should not be penalized if scripts are not included or are later updated. Also, the two-page limit is a prohibitive format for including code within the Data Management Plan. Moreover, code is typically accessed via a repository (e.g. GitHub), not as plain text.
- More specific guidance should be provided for investigators developing new software tools during the course of a grant. How should tools that are not yet developed be described at the outset? What minimum information is required?
- As mentioned above, substantially more funding for software development and hardening is required. The apps we all enjoy using (e.g. Google Maps) have involved a much higher level of refinement than any of the code we are forced to use for storage and annotation of scientific data. cBioPortal and Cytoscape are two examples of well-developed code – both require large teams and many years of investment.

3. Standards

- Before requiring investigators to meet data standards for their shared data, standards need to be developed for the vast majority of non-genomic data types. It is not sufficient to establish an ontology: tools must be developed to annotate data according to these standards, to impose uniform vocabularies and to validate annotations.
- The Common Data Element Resource Portal suggested by NIH in the Proposed Provisions is largely limited to disease-specific clinical studies and is not necessarily relevant to many basic research projects. For example, searching for “image” or even “imag” retrieved no results. This suggests there are currently no NIH standards for sharing image data, e.g. microscopy results. A more broadly applicable and more easily searchable resource than the Common Data Element Resource Portal should be in place to provide information about existing data standards for NIH investigators. For data types for which no NIH data repository exists, a list of accepted non-NIH repositories will be required; these will need persistent unique identifiers for deposited data. For example, the PRIDE database maintained by EMBL/EBI is becoming the standard for deposition of proteomics data. NIH should support and mirror this and similar types of repositories, by analogy with mirrored genomics databases.

Other considerations:

- Data archiving: If data must be archived and shared long-term or even in perpetuity, NIH needs to address how data storage will be funded if data storage fees continue to be charged to grantees after the grants which supported data production have ended. The possibility of retiring some types of data (e.g. RAW files or image-based screening data) after a predetermined period must be considered.
- Standards development: We hope that NIH will catalyze community groups to develop community-based standards for data types for which no standards exist already. The task of developing these standards is too large for individual grantees and standards

developed by small groups are unlikely to result in FAIR data. In some cases, innovative machine-human collaboration (e.g. AI) are likely to be required.

- Standards dissemination: Before the proposed data management and data sharing requirements are implemented, existing data standards should be more actively supported and disseminated. Many existing standards are currently difficult to locate and sometimes poorly documented, meaning that research groups struggle to find and correctly implement existing standards. If existing standards are not easy to find, some groups may “reinvent the wheel” and develop new, redundant standards, which again, reduce data FAIRness.
- As noted above, a standard is useless without the software infrastructure needed to implement and validate it. In our experience, this often requires some ability in scripting – we teach all of our trainees basic Python coding skills. It is for this reason that data annotation and education and closely interrelated.

References

1. AlQuraishi M, Sorger PK. Reproducibility will only come with data liberation. *Sci Transl Med*. 2016 May 18;8(339):339ed7. PMID: PMC5084089
2. Hafner M, Niepel M, Chung M, Sorger PK. Growth rate inhibition metrics correct for confounders in measuring sensitivity to cancer drugs. *Nat Methods*. 2016 Jun;13(6):521–7. PMID: PMC4887336.
3. Niepel M, Hafner M, Williams EH, Chung M, Barrette AM, Stern AD, Hu B, Gray JW, Birtwistle MR, Heiser LM, Sorger PK. A multi-center study on factors influencing the reproducibility of in vitro drug-response studies. *bioRxiv* [Internet]. 2017 Jan 1; Available from: <http://biorxiv.org/content/early/2017/11/03/213553.abstract>
4. Hafner M, Niepel M, Sorger PK. Alternative drug sensitivity metrics improve preclinical cancer pharmacogenomics. *Nat Biotechnol*. 2017 Jun 7;35(6):500–502. PMID: PMC5668135

Submission #120**Date:** 12/10/2018**Name:** Christine Zardecki**Name of Organization:** RCSB Protein Data Bank**Type of Organization:** University**Role:** Scientific Researcher**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Data archiving, structural biology

I. The definition of Scientific Data

RCSB Protein Data Bank welcomes the formulation of the NIH Data Management and Sharing Policy.

Of particular significance is the clearly articulated NIH commitment to following the FAIR Principles for data management, which the Protein Data Bank has embraced since its inception in 1971.

In our view, the plan could be improved by expanding the definition of scientific data to include all digital artifacts that support published findings. These could include images, spectra, other experimental measurements, software, and workflows.

The key to preserving all Scientific Data to enable reproducibility would be the establishment of repositories for responsible archiving. Within the US, there are a handful of Data Deposition Repositories in the biological sciences that archive enormous volumes of experimental data and metadata for widespread use by global research and educational communities (e.g., GenBank, Protein Data Bank). Support for repositories is important for managing and sharing scientific data.

While the plan encourages the use of standards where these exist, a common scenario in leading edge research is the absence of a community data standard. Additional effort and coordination will be required to rapidly extend existing data standards. Failure to address these cases may significantly reduce the value and impact of data extracted from these rapidly evolving and emerging research areas. Archival deposition repositories have traditionally taken on the maintenance and development of community data standards, and these important activities should continue to be supported by NIH data management activities.

The discussion of data standards in the current data management planning document focuses mainly on data representation. The application or development of community standards for data quality is also important. The interpretation and reuse of data may be strongly influenced by data quality considerations, and application of data quality measures could be more explicitly included as a requirement of the data management plan.

II. The requirements for Data Management and Sharing Plans

To adhere to the FAIR Principles, data should follow controlled vocabularies that would make them computer-searchable. Controlled vocabularies can be used to provide data that is well described, expertly-curated, standardized, and richly annotated. These vocabularies could be developed and maintained by data repositories, and used to apply standard validation protocols, ensure data integrity, and provide interoperability with other data resources.

Providing Data Deposition Repositories with the resources necessary to sustain and evolve their capabilities to deliver high quality data should be a key objective of the NIH Strategic Plan.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

There is an urgent need for the adoption of a data management and sharing policy, particularly one that involves controlled vocabularies for data and support for repositories.

Submission #121**Date:** 12/10/2018**Name:** Michaela Seiber, MPH**Name of Organization:** Collaborative Research Center for American Indian Health**Type of Organization:** Nonprofit Research Organization**Role:** Bioethicist/Social Science Researcher**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Social science, community based participatory research

I. The definition of Scientific Data

The Collaborative Research Center for American Indian Health has reviewed the proposed changes along with tribal nation collaborators and researcher. The consensus in terms of the definition of scientific data was that it was lacking in addressing cultural knowledge, Indigenous ways of knowing, and plans to address protected populations (i.e. pregnant women, fetus). Tribal nations already see researchers coming in and disregarding cultural knowledge, and we feel that the NIH definition of scientific data MUST address this aspect. A statement about tribal sovereignty and a tribe's ability to define scientific data how they see applicable is also suggested. It should also be clear that any data collected with a tribe belongs to the tribe and they have the right to dictate how it is used.

II. The requirements for Data Management and Sharing Plans

A statement about tribal sovereignty and a tribe's ability to require more in terms of data management and sharing plans should be included. Tribal research data is unique and needs to have additional protections and allowances for tribes to exercise data sovereignty. An additional unique aspect of tribal data is in regard to identifiable data. Tribal nations must be allowed to approve/disapprove of having their tribe identified via data in (not limited to) publications, presentations, reports, etc. When it comes to multi-center projects, there must be a way for tribes to exclude any identifiable data (if it is collected - such as tribal nation, district, village, city, reservation) from reports, presentations, publications, etc. Any restrictions to data sharing should be made up front and be known to both the researcher and the tribal nation and if a researcher plans to share any data in any way, that must be disclosed up front or when the researcher becomes aware of intent to share. No data should be shared without prior tribal approvals. Regarding large tribal datasets - the tribe/tribal IRB may make the decision to house the data, and if this is the case, it should be the first option for location to house the data; to

Submission #122**Date:** 12/10/2018**Name:** Lisa Simpson**Name of Organization:** Academy Health**Type of Organization:** Professional Org/Association**Role:****Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Health Services Research, Dissemination and Implementation Science

I. The definition of Scientific Data

The definition of scientific data offered in the proposed provisions is most helpful when considered through the lens of data sharing, to test the validity of research findings. Seen in the context of the Science Data Lifecycle Model (<https://www.usgs.gov/products/data-and-tools/data-management/data-lifecycle>), the definition appears to be focused on data at the Analyze stage - since preliminary analyses, which may be used for data processing and preparation, are not included. However, for the goals of testing validity, combining data sets, and exploring new frontiers, the definition may need to be expanded to include data in the Process stage. As more data are available in different forms for research and analysis, data for analysis is increasingly dependent on processing activities. This is important because some findings should be tested beyond the final analysis and more with data prior to being selected and filtered. In addition, data sets from different studies may be difficult to combine effectively without applying consistent processes for preparation.

An illustrative example of the value of pre-processed data for sharing is when phenotypes are computed from other data elements, such as with observational studies using data extracted from electronic health records. This is an increasingly common type of research performed and used by AcademyHealth members. Sharing these data sets is less valuable for validation, combination and exploration if it only includes the final computed value. Another example when scientific data are used with machine learning. Common approaches of applied machine learning include a process of feature selection or engineering, where characteristics of multiple data variables are evaluated together to select or derive a smaller set of variables that become the focus of the machine learning computation. Under the current definition of scientific data used, the feature selection or engineering stage could be interpreted as preliminary analyses and excluded from the policy, yet these data would be critical to sharing goals.

Another type of data that is increasingly important for inclusion in the definition of scientific data is synthetic data. Data resources that are created through random generation processes or disruption of data elements in known data sets are often used in simulation and modeling research. These are important in the scientific process. Testing the validity and accuracy of such research will require that these data are shared with other researchers.

We recognize that expanding the current definition to include pre-processed data for sharing may complicate other provisions in this policy. However, this may be necessary to better enable the policy to achieve its defined goals for sharing. Further, practices of data management may be identified that can allow an expanded scope of data sharing while limiting the challenges for data management and governance that would be introduced.

The definition of scientific data also needs consideration of and clarity around requirements regarding data that are generated through primary data collection vs. existing data sets. Some existing data sets used in health services research are licensed or have data use agreements that might restrict broader sharing. Provision #6, “Data Sharing Agreements, Licensing, and Intellectual Property” under “Requirements for Data Management and Sharing Plans” shows awareness of licensing and data use agreements, but could be easily interpreted as only applying to those that might be defined by the supported research. It would also be preferable to explicitly include these data in the definition of scientific data, and the provisions should provide guidance on how they should be considered in this policy. Inappropriate consideration of these existing data sets under this policy could lead to either misuse or reduced use of these data. Such data should not be simply excluded from the policy requirements. Methods for data processing or analysis could be shared, thereby providing pathways for validation and promoting exploration.

The concept of data management, which is used in definitions supporting the definition of scientific data, itself needs better definition in the provisions. Above we have applied existing models of data lifecycles for data management to help interpret the definition of scientific data. If the policy is more explicit about data management and the stages of the data lifecycle to which the policy pertains, it can better define scope and guide compliance.

Finally, AcademyHealth supports the emphasis on digitizing scientific data that is made in the definition, as digital data are generally more easily shared. We recommend expansion to include use of standards where possible as well. Data shared in digital formats may be more easily stored and distributed, but their actual use will be more limited when the data are represented in non-standard formats. Data for sharing should be stored and represented in a way that they can be both distributable and interoperable.

II. The requirements for Data Management and Sharing Plans

AcademyHealth believes the expanded requirements for data management and sharing plans can advance data sharing by requiring more detailed consideration and commitment by scientists to share data. Our members have observed that while some consideration of data

sharing has been required in prior proposals for NIH-supported research, it has been difficult for researchers to develop plans that support the goals of data sharing. Often, plans are specified that do not promote open data for science, but rather provide a minimal commitment to make some degree of data sharing possible. We also believe that the support of and incentives for investigators to develop and implement acceptable data management and sharing plans will be critical to the success of the policy.

We recognize that these new requirements will require expanded efforts by investigators and their institutions for adherence, especially during early stages before best practices are established and systems for open research data are optimized. NIH will need to closely evaluate the research funding investigators and institutions need to facilitate open research data, and provide adequate support for these activities. NIH has applied open sharing to various projects, typically those with large multi-institutional programs where ground rules are equally applied to all participants. Experience with these projects will be useful in considering what additional support may be necessary. NIH should also evaluate costs and methods of engagement for additional data preparation by investigators providing scientific data that may be needed beyond an award period. NIH should ensure incentives are appropriate for investigators who agree to share data and should support and fund storage and sharing as part of the grant process.

Beyond these initial projects where open sharing has been applied, we believe that many investigators and institutions are limited in their experience and ability for developing appropriate data sharing plans. This can make the initial implementation of the policy difficult and may lead to inconsistent implementation while practices in effective data sharing evolve. Therefore, NIH should prioritize and recommend timelines for measures that reflect data sharing performance and institutional compliance, from plan development to actual data sharing. Either the provisions should specifically define the prioritization and timelines, or the provisions should reference guidelines that may be adapted over time. Of these options, the use of guidelines may be more flexible and better adapt to changing knowledge of best practices.

NIH should also fund additional research on research support operations and technical platforms that could improve the efficiency of developing and implementing an appropriate data management and sharing plan. NIH should also support research and training on how to best manage and share data under this new policy. Such support will be important in accelerating the discovery best practices.

Portions of the draft policy provisions pertaining to the acceptance of appropriate data management and sharing plans prior to awarding funding are reasonable, as long as the requirements for the plans are specified in advance in funding guidelines, and appropriate support is provided to investigators in developing plans. Investigator adherence to and advance of best practices in data sharing may be best achieved if the appropriateness of plans is considered more directly in scoring for funding, rather than just as a condition for funding. This

could be implemented as a specific question during peer review or may be more effective as a component of the other dimensions for scoring (e.g., Approach).

Specific expectations for data management and sharing plans can improve the overall data management of proposals, which is an ancillary benefit to the provisions. Clarity of expectations and best practices is also needed regarding informed consent. Participants in research studies have a right to know how data will be shared and should be informed during the consent process. The provisions also need to consider how protected health information should be considered in sharing plans. Appropriately managing privacy requires better guidance, as asking for broad consent for subsequent data use places the burden of protecting privacy on research participants who must decide up front whether or not to consent to any potential uses. This is an important risk, because if data sharing requires protected health information sharing, consent may be more difficult to obtain from participants, particularly those who may have reasonable concerns about use of their data too broadly. Elements of good data stewardship and fair information practice principles should be included as well.

The current provisions suggest that data management and sharing plans could have a two-page limit. We appreciate the intent of the suggestion to make the plans concise; however, such a limitation may be difficult practice without precise guidance in how different research elements should be addressed. For example, typical data use agreements are well beyond the recommended page limit, even when including only those sections most relevant to the proposed policy. Without clear expectations, the recommended page limit may not be useful.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

As mentioned above, flexibility is needed in implementing these provisions, as they can represent a significant burden to investigators and institutions until standard practices for data management and sharing emerge. In addition, specific guidance is needed for investigators who have little experience and infrastructure to support open science data. For this reason, we believe guidelines should be referenced, established and frequently updated as standard practices evolve. NIH should also support research to improve discovery of effective practices, and should support the effective dissemination of these best practices to investigators, institutions, and institutional review boards (IRBs). Special consideration for dissemination should be given for IRBs, who would have a greater burden with implementing step-wise recommendations.

The most important consideration for timing is that these requirements should apply only to new funding awards, and to awards where the requirements for data sharing can be specified in funding guidelines. First, data sharing considerations should be understood before the consent process is defined, as participants have a right to know in how their data may be

shared. Second, while it is reasonable for the sharing plans to be acceptable as a condition for funding, the negotiation of what constitutes an effective plan and terms should not be done individually as a condition of an award.

In terms of phased adoption for different types of research or funding mechanisms, we recommend the following considerations. First, NIH has for various projects applied open data sharing, and these have typically been large multi-institutional programs. Extending implementation first from these groups seems reasonable. Second, clinical and translational research domains may be appropriate for early implementation of the provisions, due to the proximal impact the findings may have on direct application to health. Third, training programs and small grants programs may be more appropriately considered for later implementation, with larger programs prioritized to mitigate additional reporting and management burdens. Fourth (and perhaps most important), the policy should be implemented as quickly as is reasonable, as there are clear benefits to open research data.

Finally, NIH should advance the infrastructure for storage of shared data. Current recommendations are for considering repositories available at no cost for extended periods of time. Without NIH support, such repositories may become either compromising or unsustainable. If investigators and institutions agree to share data, the actual storage of that data to enable sharing should not be a primary concern of the data contributors.

Attachment:

AcademyHealth Response

NIH Proposed Provisions for a Future Draft Data Management and Sharing Policy

December 10, 2018

As a member-based organization that serves the research community, AcademyHealth is deeply interested in the production and use of evidence to improve health and the performance of the health system. Our work, and that of our more than 4,000 individual and organizational members, is directly impacted by policies regarding the collection, use, governance and sharing of data to facilitate research. As such, AcademyHealth applauds the National Institutes of Health (NIH) for adopting a plan to increase access to scientific publications and data in 2015, and we are encouraged by the current effort to consider a new data management and sharing policy in support of that plan. We commend the NIH for offering these proposed provisions, which provide a helpful foundation for considering these issues, and suggest that more definition is needed both to ensure that researchers understand NIH expectations and that the overall vision to encourage responsible data management and sharing is realized. In this spirit, that we are grateful for the opportunity to provide comments on the three primary topics of interest for the NIH.

The Definition of Scientific Data

The definition of scientific data offered in the proposed provisions is most helpful when considered through the lens of data sharing, to test the validity of research findings. Seen in the context of the Science Data Lifecycle Model (<https://www.usgs.gov/products/data-and-tools/data-management/data-lifecycle>), the definition appears to be focused on data at the *Analyze* stage - since preliminary analyses, which may be used for data processing and preparation, are not included. However, for the goals of testing validity, combining data sets, and exploring new frontiers, the definition may need to be expanded to include data in the *Process* stage. As more data are available in different forms for research and analysis, data for analysis is increasingly dependent on processing activities. This is important because some findings should be tested beyond the final analysis and more with data prior to being selected and filtered. In addition, data sets from different studies may be difficult to combine effectively without applying consistent processes for preparation.

An illustrative example of the value of pre-processed data for sharing is when phenotypes are computed from other data elements, such as with observational studies using data extracted from electronic health records. This is an increasingly common type of research performed and used by AcademyHealth members. Sharing these data sets is less valuable for validation, combination and exploration if it only includes the final computed value. Another example when scientific data are used with machine learning. Common approaches of applied machine

learning include a process of feature selection or engineering, where characteristics of multiple data variables are evaluated together to select or derive a smaller set of variables that become the focus of the machine learning computation. Under the current definition of scientific data used, the feature selection or engineering stage could be interpreted as preliminary analyses and excluded from the policy, yet these data would be critical to sharing goals.

Another type of data that is increasingly important for inclusion in the definition of scientific data is synthetic data. Data resources that are created through random generation processes or disruption of data elements in known data sets are often used in simulation and modeling research. These are important in the scientific process. Testing the validity and accuracy of such research will require that these data are shared with other researchers.

We recognize that expanding the current definition to include pre-processed data for sharing may complicate other provisions in this policy. However, this may be necessary to better enable the policy to achieve its defined goals for sharing. Further, practices of data management may be identified that can allow an expanded scope of data sharing while limiting the challenges for data management and governance that would be introduced.

The definition of scientific data also needs consideration of and clarity around requirements regarding data that are generated through primary data collection vs. existing data sets. Some existing data sets used in health services research are licensed or have data use agreements that might restrict broader sharing. Provision #6, “Data Sharing Agreements, Licensing, and Intellectual Property” under “Requirements for Data Management and Sharing Plans” shows awareness of licensing and data use agreements, but could be easily interpreted as only applying to those that might be defined by the supported research. It would also be preferable to explicitly include these data in the definition of scientific data, and the provisions should provide guidance on how they should be considered in this policy. Inappropriate consideration of these existing data sets under this policy could lead to either misuse or reduced use of these data. Such data should not be simply excluded from the policy requirements. Methods for data processing or analysis could be shared, thereby providing pathways for validation and promoting exploration.

The concept of data management, which is used in definitions supporting the definition of scientific data, itself needs better definition in the provisions. Above we have applied existing models of data lifecycles for data management to help interpret the definition of scientific data. If the policy is more explicit about data management and the stages of the data lifecycle to which the policy pertains, it can better define scope and guide compliance.

Finally, AcademyHealth supports the emphasis on digitizing scientific data that is made in the definition, as digital data are generally more easily shared. We recommend expansion to include use of standards where possible as well. Data shared in digital formats may be more easily stored and distributed, but their actual use will be more limited when the data are represented in non-standard formats. Data for sharing should be stored and represented in a way that they can be both distributable and interoperable.

The Requirements for Data Management and Sharing Plans

AcademyHealth believes the expanded requirements for data management and sharing plans can advance data sharing by requiring more detailed consideration and commitment by scientists to share data. Our members have observed that while some consideration of data sharing has been required in prior proposals for NIH-supported research, it has been difficult for researchers to develop plans that support the goals of data sharing. Often, plans are specified that do not promote open data for science, but rather provide a minimal commitment to make some degree of data sharing possible. We also believe that the support of and incentives for investigators to develop and implement acceptable data management and sharing plans will be critical to the success of the policy.

We recognize that these new requirements will require expanded efforts by investigators and their institutions for adherence, especially during early stages before best practices are established and systems for open research data are optimized. NIH will need to closely evaluate the research funding investigators and institutions need to facilitate open research data, and provide adequate support for these activities. NIH has applied open sharing to various projects, typically those with large multi-institutional programs where ground rules are equally applied to all participants. Experience with these projects will be useful in considering what additional support may be necessary. NIH should also evaluate costs and methods of engagement for additional data preparation by investigators providing scientific data that may be needed beyond an award period. NIH should ensure incentives are appropriate for investigators who agree to share data and should support and fund storage and sharing as part of the grant process.

Beyond these initial projects where open sharing has been applied, we believe that many investigators and institutions are limited in their experience and ability for developing appropriate data sharing plans. This can make the initial implementation of the policy difficult and may lead to inconsistent implementation while practices in effective data sharing evolve. Therefore, NIH should prioritize and recommend timelines for measures that reflect data sharing performance and institutional compliance, from plan development to actual data sharing. Either the provisions should specifically define the prioritization and timelines, or the provisions should reference guidelines that may be adapted over time. Of these options, the use of guidelines may be more flexible and better adapt to changing knowledge of best practices.

NIH should also fund additional research on research support operations and technical platforms that could improve the efficiency of developing and implementing an appropriate data management and sharing plan. NIH should also support research and training on how to best manage and share data under this new policy. Such support will be important in accelerating the discovery best practices.

Portions of the draft policy provisions pertaining to the acceptance of appropriate data management and sharing plans prior to awarding funding are reasonable, as long as the requirements for the plans are specified in advance in funding guidelines, and appropriate support is provided to investigators in developing plans. Investigator adherence to and advance of best practices in data sharing may be best achieved if the appropriateness of plans is considered more directly in scoring for funding, rather than just as a condition for funding. This could be implemented as a specific question during peer review or may be more effective as a component of the other dimensions for scoring (e.g., Approach).

Specific expectations for data management and sharing plans can improve the overall data management of proposals, which is an ancillary benefit to the provisions. Clarity of expectations and best practices is also needed regarding informed consent. Participants in research studies have a right to know how data will be shared and should be informed during the consent process. The provisions also need to consider how protected health information should be considered in sharing plans. Appropriately managing privacy requires better guidance, as asking for broad consent for subsequent data use places the burden of protecting privacy on research participants who must decide up front whether or not to consent to any potential uses. This is an important risk, because if data sharing requires protected health information sharing, consent may be more difficult to obtain from participants, particularly those who may have reasonable concerns about use of their data too broadly. Elements of good data stewardship and fair information practice principles should be included as well.

The current provisions suggest that data management and sharing plans could have a two-page limit. We appreciate the intent of the suggestion to make the plans concise; however, such a limitation may be difficult practice without precise guidance in how different research elements should be addressed. For example, typical data use agreements are well beyond the recommended page limit, even when including only those sections most relevant to the proposed policy. Without clear expectations, the recommended page limit may not be useful.

The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

As mentioned above, flexibility is needed in implementing these provisions, as they can represent a significant burden to investigators and institutions until standard practices for data management and sharing emerge. In addition, specific guidance is needed for investigators who have little experience and infrastructure to support open science data. For this reason, we believe guidelines should be referenced, established and frequently updated as standard practices evolve. NIH should also support research to improve discovery of effective practices, and should support the effective dissemination of these best practices to investigators, institutions, and institutional review boards (IRBs). Special consideration for dissemination should be given for IRBs, who would have a greater burden with implementing step-wise recommendations.

The most important consideration for timing is that these requirements should apply only to new funding awards, and to awards where the requirements for data sharing can be specified in funding guidelines. First, data sharing considerations should be understood before the consent process is defined, as participants have a right to know in how their data may be shared. Second, while it is reasonable for the sharing plans to be acceptable as a condition for funding, the negotiation of what constitutes an effective plan and terms should not be done individually as a condition of an award.

In terms of phased adoption for different types of research or funding mechanisms, we recommend the following considerations. First, NIH has for various projects applied open data sharing, and these have typically been large multi-institutional programs. Extending implementation first from these groups seems reasonable. Second, clinical and translational research domains may be appropriate for early implementation of the provisions, due to the proximal impact the findings may have on direct application to health. Third, training programs and small grants programs may be more appropriately considered for later implementation, with larger programs prioritized to mitigate additional reporting and management burdens. Fourth (and perhaps most important), the policy should be implemented as quickly as is reasonable, as there are clear benefits to open research data.

Finally, NIH should advance the infrastructure for storage of shared data. Current recommendations are for considering repositories available at no cost for extended periods of time. Without NIH support, such repositories may become either compromising or unsustainable. If investigators and institutions agree to share data, the actual storage of that data to enable sharing should not be a primary concern of the data contributors.

Additional considerations

These are additional considerations for the provisions that were noted by our members.

More clarity is needed regarding “Compliance and Enforcement.” The concept of making data available to the scientific community “as long as it is useful” lacks definition, is highly subjective, and is problematic. Investigators cannot require unlimited consent from research subjects, and this as written is similar to requesting unrestricted access to data. Terms of the agreement for sharing should be clear and unambiguous.

We support the goals of open research data and these efforts to advance them by the NIH. We also recognize that there will be situations where due to licensing restrictions or privacy risks, some data may not yet be appropriate for sharing. However, we hope that exclusions will be seen as exceptional and that the burden of proof should be on the investigators or institutions to define the exclusions, rather than an equal burden to exclude or include data. Data sharing currently carries sufficient burden in implementation that it should be facilitated, and the goals of open research data for NIH-supported research should be promoted.

In Provision #6, “Data Sharing Agreements, Licensing, and Intellectual Property” under “Requirements for Data Management and Sharing Plans,” the term “intellectual property” was noted as unclear, difficult to define, and problematic. It will be difficult for researchers and the NIH to always have a working agreement about the meaning of the term, which will create issues. Instead, the NIH should define precisely what it intends in regard to the various legal rights for research data. Explicitly defining a term like “proprietary interest” may be a better approach.

As alluded to prior, it is important to increase consideration of the research participant or subject with regard to sharing data. Studies have shown that patients’ perspectives on appropriate use of data can be very different than that of investigators. These provisions are focused on the researcher perspective, which is appropriate given the context for implementation, but additional

consideration is needed of the subject perspective. How subjects and patients understand data sharing in the context of data privacy, ownership, and consent will be critical for successful implementation. Such perspectives will be important for defining governance for appropriate use of shared data, which also seemed lacking in the provisions. Researchers may be able to advocate in the data sharing plan elements for protection of proprietary interest, but the provisions are not clear in defining protections for appropriate use from patient perspectives and ensuring protection of privacy.

Conclusion

AcademyHealth appreciates the opportunity given by NIH to provide input regarding these proposed provisions for a new data management and sharing policy. Our organization has a shared interest in the goals of the policy and its successful definition and implementation. We believe our organization and members can be helpful in disseminating guidelines and best practices for data sharing plans as they are discovered and as they evolve. We believe a successful policy can achieve many benefits, but one that is not well defined or implemented can actually impede the use of various data for generating evidence that can improve health and the performance of the health system.

Submission #123

Date: 12/10/2018

Name: Anonymous

Name of Organization:

Type of Organization: University

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

neurobiology

I. The definition of Scientific Data

no comment

II. The requirements for Data Management and Sharing Plans

In response to the request for information, we would like to submit our input to the NIH on behalf of the faculty of the Department of Neurobiology at Harvard Medical School, and with the perspective of the data needs of neurobiology research.

In order to meet the requirements for Data Management and Sharing, it is essential that the NIH provides budgetary support for data storage as well as support that would enable investigators to build necessary infrastructure/interfaces to share data. In many cases, investigators must create public repositories de novo, and this requires tremendous investment of time and effort in building a website, creating the necessary parameters, and user interfaces. Needless to say, this area of software development/website building is not routinely found in individual labs, and there are no pipelines in place to facilitate these efforts. The NIH should provide budgetary support based on the project, or by indirect costs, to help investigator create useful data repositories that can be hosted on servers, and shared with the general community. Our investigators have had to pay expenses for building websites using other resources, or submit large resource grants.

The Allen Brain Institute hosts large datasets from the molecular to the connectivity level and provides easy to use interface for investigators from all over the world. As single cell sequencing data, electrophysiology data, behavioral, and connectivity datasets continue to emerge from individual laboratories, it is essential that the NIH support efforts to make these data accessible in a way that is similar to the Allen Brain Institute's services.

The Human Cell Atlas initiative has provided funding to computational biologists/software engineers to build data portals to analyze and store data created by other HCA investigators, and the NIH should consider funding similar efforts in the neurobiology space. This would encourage a standardization of metadata requirements so that the data can be easily searched and analyzed by various laboratories all over the world. The NIH must be clear in its definition of metadata and requirements to make a given dataset 'shareable'.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

no comment

Submission #124**Date:** 12/10/2018**Name:** Michael Mabe**Name of Organization:** International Association of Scientific Technical and Medical Publishers (STM)**Type of Organization:** Other**Other Type of Organization:** Publishing and Research Services; Trade Association**Role:** Institutional Official**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

All areas of research

I. The definition of Scientific Data

STM supports the intent of NIH in promoting data sharing in order to improve reproducibility and transparency, and we are pleased to see that the goal of validation and replication is included in the definition of data. In looking at what information is necessary for this goal, it is critical to distinguish between data itself and various types of presentation of data, and appropriately consider a researcher's rights to data generated in his or her research, as well as to respect intellectual property protection and copyright laws. The Data Publication Pyramid on p. 6 of the "Report on Integration of Data and Publications" (http://www.stm-assoc.org/2011_12_5_ODE_Report_On_Integration_of_Data_and_Publications.pdf), written by a coalition representing researchers, publishers, libraries and data centers, is a comprehensive overview of research data and sharing, and many of the issues involved, but, tellingly, the report does not itself provide a definition of data.

A definition must be precise enough to make a distinction between the data and various interpretations and presentations of that data, whilst at the same time being flexible enough to encompass the data practices of a wide variety of fields. It should also be consistent with other descriptions of data in federal policy and code. The definition proposed in the "Proposed Provisions for a Draft NIH Data Management and Sharing Policy" has a significant benefit in being similar to that in the 2013 OSTP memo on "Increasing Access to the Results of Federally Funded Scientific Research," upon which other federal agencies have built their data management policies. At the same time, it would be helpful if NIH would further clarify the meaning of data as primary information and not analyses or creative presentations of the information.

Our recommendations are:

(1) that “data used to support scholarly publications” be modified to “primary data that support the finding presented in scholarly publications”

(2) that the list of materials that are not considered data include all “analyses” and all “versions of scientific papers” rather than only “preliminary analyses” and “drafts of scientific papers.”

II. The requirements for Data Management and Sharing Plans

STM’s members publish in a wide variety of research areas, each of which has different practices with respect to data collection, use, and sharing. The plan requirements must be flexible enough to support the diverse nature of the research that NIH funds, while also providing guidance to all researchers to encourage and enable sharing. STM welcomes the opportunity for dialogue with NIH and all stakeholders to find ways to increase the impact of research data.

With the diversity of data practices and differences in the intensity of data usage in different fields, it may not be appropriate to limit data management plans to two pages in all cases. NIH may want to consider providing the limit as a guideline, or adjusting it in the case of multi-institutional or more complex data plans.

With respect to data preservation and access, NIH may want to provide guidance to researchers on criteria for an appropriate and trusted location for data, including plans for perpetual access and commitment to the FAIR Data principles. Several initiatives offer certification for or recommendations of trusted data repositories, including CoreTrustSeal (<https://www.coretrustseal.org/>) and Repository Finder (<https://repositoryfinder.datacite.org/about>; <https://www.re3data.org/>).

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

Whatever the timing of the implementation, STM believes that it is critically important that as the policies are implemented a robust and regular review and evaluation takes place with extensive stakeholder input. Data sharing policies are likely to have profound effects on the research enterprise which will not be fully understood until they are implemented. Regular review will enable NIH to address any unintended consequences in a timely manner, as well as help take advantage of changes in research practices or technologies.

STM, through its involvement in RDA and other initiatives, is already contributing to the development of the standards, resources, policies, and infrastructure needed to enable robust data sharing across the research community. We welcome further discussion on how NIH, STM,

and our member publishers can work together to build trust in science and promote the use of research data for the benefit of research and the public.

Attachment:

10 December 2018

Response to Request for Information (RFI) on Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research

The International Association of Scientific, Technical and Medical Publishers (STM) is the leading global trade association for academic and professional publishers. It has more than 150 members in 21 countries who each year collectively publish more than 66% of all journal articles and tens of thousands of monographs and reference works. STM members include non-profit scientific and scholarly societies, commercial publishers, and university presses who work collectively to ensure broad access to and use of the latest scientific and scholarly information. The majority of our members are small businesses and not-for-profit organizations, who represent tens of thousands of publishing employees, editors and authors, and other professionals across the United States and world who regularly contribute to the advancement of science, learning, culture and innovation throughout the nation. They comprise the bulk of a \$10 billion publishing industry that contributes significantly to the U.S. economy and enhances the U.S. balance of trade.

Publishers sit at the interface between researchers, their research and the rest of the world through our work to improve the quality and availability of information related to research. STM shares our members' commitment to supporting researchers in the sharing, discoverability, and reuse of research data. Individual publishers are developing tools and services to support researchers to make their data FAIR (Findable, Accessible, Interoperable, and Re-usable), and have actively responded to community demand for citation principles for data. STM itself has been involved in numerous projects looking at data access, citation, and preservation, the most recent example of which is support for the development of [SCHOLIX](#), an easy and universal linking mechanism between scholarly publications and research data.

In keeping with our commitment to promote sustainable open science, STM therefore supports NIH's efforts to improve data management and sharing, and welcomes this opportunity to respond to NOT-OD-19-014, "Request for Information (RFI) on Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research," as published on October 10, 2018. This submission builds on responses that STM has submitted to previous NIH RFIs on research data and digital repositories, as well as responses that STM has submitted to two RFCs on the Federal Data Strategy.

The following comments are in response to the specific areas upon which NIH has requested that respondents focus, and have been submitted to the appropriate boxes in the response form as well as being copied here.

I. The definition of Scientific Data (Provisions I, II, and III)

STM supports the intent of NIH in promoting data sharing in order to improve reproducibility and transparency, and we are pleased to see that the goal of validation and replication is included in the definition of data. In looking at what information is necessary for this goal, it is critical to distinguish between data itself and various types of presentation of data, and appropriately consider a researcher's rights to data generated in his or her research, as well as to respect intellectual property protection and copyright laws. The Data Publication Pyramid on p. 6 of the "Report on Integration of Data and Publications" (http://www.stm-assoc.org/2011_12_5_ODE_Report_On_Integration_of_Data_and_Publications.pdf), written by a coalition representing researchers, publishers, libraries and data centers, is a comprehensive overview of research data and sharing, and many of the issues involved, but, tellingly, the report does not itself provide a definition of data.

A definition must be precise enough to make a distinction between the data and various interpretations and presentations of that data, whilst at the same time being flexible enough to encompass the data practices of a wide variety of fields. It should also be consistent with other descriptions of data in federal policy and code. The definition proposed in the "Proposed Provisions for a Draft NIH Data Management and Sharing Policy" has a significant benefit in being similar to that in the 2013 OSTP memo on "Increasing Access to the Results of Federally Funded Scientific Research," upon which other federal agencies have built their data management policies. At the same time, it would be helpful if NIH would further clarify the meaning of data as primary information and not analyses or creative presentations of the information.

Our recommendations are:

- (1) that "data used to support scholarly publications" be modified to "primary data that support the finding presented in scholarly publications"
- (2) that the list of materials that are not considered data include all "analyses" and all "versions of scientific papers" rather than only "preliminary analyses" and "drafts of scientific papers."

II. The requirements for Data Management and Sharing Plans (Provision IV)

STM's members publish in a wide variety of research areas, each of which has different practices with respect to data collection, use, and sharing. The plan requirements must be flexible enough to support the diverse nature of the research that NIH funds, while also providing guidance to all researchers to encourage and enable sharing. STM welcomes the opportunity for dialogue with NIH and all stakeholders to find ways to increase the impact of research data.

With the diversity of data practices and differences in the intensity of data usage in different fields, it may not be appropriate to limit data management plans to two pages in all cases. NIH may want to consider providing the limit as a guideline, or adjusting it in the case of multi-institutional or more complex data plans.

With respect to data preservation and access, NIH may want to provide guidance to researchers on criteria for an appropriate and trusted location for data, including plans for perpetual access and commitment to the FAIR Data principles. Several initiatives offer certification for or

recommendations of trusted data repositories, including CoreTrustSeal (<https://www.coretrustseal.org/>) and Repository Finder (<https://repositoryfinder.datacite.org/about>; <https://www.re3data.org/>).

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards. (Provision V)

Whatever the timing of the implementation, STM believes that it is critically important that as the policies are implemented a robust and regular review and evaluation takes place with extensive stakeholder input. Data sharing policies are likely to have profound effects on the research enterprise which will not be fully understood until they are implemented. Regular review will enable NIH to address any unintended consequences in a timely manner, as well as help take advantage of changes in research practices or technologies.

STM, through its involvement in RDA and other initiatives, is already contributing to the development of the standards, resources, policies, and infrastructure needed to enable robust data sharing across the research community. We welcome further discussion on how NIH, STM, and our member publishers can work together to build trust in science and promote the use of research data for the benefit of research and the public.

Very truly yours,



Michael Mabe

CEO STM

Submission #125**Date:** 12/10/2018**Name:** Brett Harnet**Name of Organization:** University of Cincinnati**Type of Organization:** University**Role:** Scientific Researcher**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

All of the above

II. The requirements for Data Management and Sharing Plans

The draft data-sharing plan is quite comprehensive. It lacks one key element in my opinion that can be summed up in one word: Ebola. Not once in the 2700 word document is there mention of the words “world, global or international”. In the era of globalization, public health is defined by the jetways of the commercial airline industry. While NIH funds are focused on the United States, any research done around the globe should have at least the opportunity to – or at most the requirement to share research data. And that data needs context (meta data) that creates semantic interoperability. This includes common data models, shared ontologies (two other words not mentioned) and the use of standards. The problem with standards is [SIC] ‘everyone has their favorites’. I suggest the NIH requires all clinical/translation data be de-identified and uploaded to a single NIH cloud that has a standardized data model(s), required metadata and use of established vocabularies. This is what the document implies, but not stated beyond section V. While I am not a fan of duplicating data, in this case, we need a single, central repository. (Eliminate section 4.1.) The public - especially scientists - need a single place to go to mine this data. Compared to ten years ago, any high costs have eliminated by Amazon Web Services and Moore’s Law.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

Implementing this for funding distributed across the U.S. should be easy, but to truly achieve scale, we need the global community to participate. We are a complex and heterogeneous village but zeros and ones are still binary.

Submission #126**Date:** 12/10/2018**Name:** Alessia Daniele**Name of Organization:** Weill Cornell Medicine**Type of Organization:** Other**Other Type of Organization:**

Academic medical center with a tripartite mission of patient care, education and research

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

Weill Cornell Medicine is committed to excellence in patient care, scientific discovery and the education of future physicians and scientists in New York City and around the world. All research areas are of importance to the institution.

I. The definition of Scientific Data

Our investigators are recipients of awards from the NIH, the Agency for Healthcare Research and Quality (AHRQ), the Patient-Centered Outcomes Research Institute (PCORI), the Centers for Disease Control and Prevention (CDC), the Food and Drug Administration (FDA) and the Department of Defense's Congressional Directed Medical Research Programs (CDMRP). Frequently, they collaborate with scientists from academic research institutions across the country, as well as with our parent university Cornell University. The Association of American Universities (AAU) and the Association of Public & Land-Grant Universities (APLU) convened a Public Access Working Group and released a November 2017 report on recommendations and actions universities and federal agencies can take to, "advance public access to data in a viable and sustainable way." The report states federal agencies will need to provide funding to make data widely available and, "provide consistent and clear policies, compliance guidelines, and definitions across agencies to minimize the burden on researchers and institutions." We strongly support these claims. As you and the agency well know, research projects often leverage multiple government funding sources. This often results in research collaborations that benefit from support from two or more federal agencies, such as the National Science Foundation (NSF) and the NIH for example. We appreciate the NIH's proposed provisions acknowledging this reality through the proposed definition of Scientific Data mirroring the NSF's definition of Research Data; however, we note and caution the difference in verbiage between "Scientific" and "Research." We urge the NIH's definition be broadly applicable to avoid any potential impediments to the academic research community's ability to collaborate. We also encourage the NIH to consider the inclusion of code (e.g. SQL, R, Python) in the

definition of scientific data if such code is requisite to interact with the data. This will help address, though not resolve, reproducibility challenges as software stacks change and evolve so quickly.

II. The requirements for Data Management and Sharing Plans

Standards for data repositories should be left to the institution and investigator. What might be acceptable or sufficient for one area of study might be wholly inadequate for another and therefore a large degree of latitude should be retained. This would also allow for new systems and cloud-based repositories to be created without the impediment of outdated requirements that could take the agency an extensive amount of time to address.

Individual requests for proposals (RFPs) should delineate clear standards and minimum expectations, relevant to the specific scientific discipline, necessary to meet the data management and sharing requirement. Where possible, templates and samples should be provided to illustrate effective and compliant plans.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

The policy, as it currently reads, would essentially make institutional data repositories a requirement for receiving NIH funding and the dollar figures associated with mandating such a requirement would create significant financial strain on institutions that are not prepared, nor equipped to create or manage such costly and extensive systems. Rather than mobilizing a mass reorganization of academic and biomedical research institution's data management infrastructure, we recommend delaying the progression of any draft policy until insights and best practices can be leveraged from the Data Commons project.

Attachment:

December 10, 2018

Carrie D. Wolinetz, Ph.D.
Office of Science Policy
National Institutes of Health
6705 Rockledge Drive, Suite 750
Bethesda, MD 20892

Re: Request for Information on Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research

Dear Dr. Wolinetz and the Office of Science Policy,

Weill Cornell Medicine is committed to excellence in patient care, scientific discovery and the education of future physicians and scientists in New York City and around the world. The doctors and scientists of Weill Cornell Medicine — faculty from Weill Cornell Medical College, Weill Cornell Graduate School of Medical Sciences, and Weill Cornell Physician Organization—are engaged in world-class clinical care and cutting-edge research that connects patients to the latest treatment innovations and prevention strategies.

We are deeply grateful for the opportunity to provide comment and insight into the proposed provisions of the draft National Institutes of Health's (NIH) Data Management and Sharing Policy, and applaud the agency for taking such a transparent and inclusive approach. The current Data Management and Sharing Policy was last updated in 2003. Since then, there has and continues to be an innovational renaissance in information technology. There has also been a collective appreciation and understanding among the scientific community of the importance of data sharing for the purposes of furthering scientific advances and improving human health. The NIH has attempted to continue to encourage widespread data sharing through the creation of the NIH Data Commons Project launched last year. The project has the potential to provide a trove of insight, information and best practices that could advise this data management and sharing policy process, however it cannot currently do so due to the programs infancy. We agree there is room for improvement in the current environment; however, we strongly caution the overly broad and simplistic approach the November proposed provisions have taken. The policy, as it currently reads, would essentially make institutional data repositories a requirement for receiving NIH funding and the dollar figures associated with mandating such a requirement would create significant financial strain on institutions that are not prepared, nor equipped to create or manage such costly and extensive systems. Rather than mobilizing a mass reorganization of academic and biomedical research institution's data management infrastructure, we recommend delaying the progression of any draft policy until insights and best practices can be leveraged from the Data Commons project.

Weill Cornell Medicine is comprised of 5,866 clinical faculty, 373 basic science faculty and over 4,500 staff supporting the clinical, research and education mission of the institution. The institution is part of the New York City Consortium for the All of Us Research Program and a participant in a myriad of other NIH led programs and projects that aim to increase data sharing and usability. We received 722 NIH awards in CY2016-2017, which supported 4,818 published journal articles and citations, and \$243 million in research funding in FY2017-2018, which includes \$127.3 million in NIH funding, \$39.7 million from other government agencies and \$17.8 million for clinical trials. The breadth and depth of the basic science, translational and clinical research being conducted in our nearly 470,000 square feet of research space is extremely vast. To oversee these research

operations is our Institutional Review Board (IRB), Office of Research Compliance and tactical and specialized Information Technologies and Services teams (ITS), to name a few. Our ITS division consists of teams including Research Informatics (RI), which facilitates the conduct and administration of clinical and translational research through its Architecture for Research Computing in Health (ARCH). The Data Integration team and the Scientific Computing team create and maintain infrastructure to host computationally-intensive research systems while working closely with Weill Cornell Medicine scientists to define and develop new services for the research community.

Our investigators are recipients of awards from the NIH, the Agency for Healthcare Research and Quality (AHRQ), the Patient-Centered Outcomes Research Institute (PCORI), the Centers for Disease Control and Prevention (CDC), the Food and Drug Administration (FDA) and the Department of Defense's Congressional Directed Medical Research Programs (CDMRP). Frequently, they collaborate with scientists from academic research institutions across the country, as well as with our parent university Cornell University. The Association of American Universities (AAU) and the Association of Public & Land-Grant Universities (APLU) convened a Public Access Working Group and released a November 2017 report on recommendations and actions universities and federal agencies can take to, "advance public access to data in a viable and sustainable way." The report states federal agencies will need to provide funding to make data widely available and, "provide consistent and clear policies, compliance guidelines, and definitions across agencies to minimize the burden on researchers and institutions." We strongly support these claims. As you and the agency well know, research projects often leverage multiple government funding sources. This often results in research collaborations that benefit from support from two or more federal agencies, such as the National Science Foundation (NSF) and the NIH for example. We appreciate the NIH's proposed provisions acknowledging this reality through the proposed definition of Scientific Data mirroring the NSF's definition of Research Data; however, we note and caution the difference in verbiage between "Scientific" and "Research." We urge the NIH's definition be broadly applicable to avoid any potential impediments to the academic research community's ability to collaborate. We also encourage the NIH to consider the inclusion of code (e.g. SQL, R, Python) in the definition of scientific data if such code is requisite to interact with the data. This will help address, though not resolve, reproducibility challenges as software stacks change and evolve so quickly.

According to the proposed provisions draft, the data plan would have a two-page limit and address: (i) data types, (ii) related tools and software, (iii) data standards, (iv) data preservation, access (including timelines) and discoverability, (v) terms for re-use and redistribution, (vi) limitations on access, and (vii) oversight of data management. We value and endorse the decision for the data plan to have a relatively short two-page limit and not be factored into the overall impact score through the peer review process. We are, however, unable to assess the potential implications of the requirement that every individual NIH Institute or Center (IC) have the ability to approve a data plan. We fear this could create a reality within which what is deemed acceptable for one IC is unacceptable for another. This would only further administrative burden and complicate the already rigorous and extensive research grant application and compliance requirements that currently exist.

Additional items for which we have recommendations and/or concerns:

- We strongly encourage the NIH to include or at minimum acknowledge in any policy revision the reality that not all data is useful in perpetuity and that not all data can or even should be shared. This is especially important to consider as it pertains to confidential clinical data. It is critically important

the Science Policy Administration take these realities into consideration and build flexibility into any data management and sharing policy.

- Standards for data repositories should be left to the institution and investigator. What might be acceptable or sufficient for one area of study might be wholly inadequate for another and therefore a large degree of latitude should be retained. This would also allow for new systems and cloud-based repositories to be created without the impediment of outdated requirements that could take the agency an extensive amount of time to address.
- A plethora of additional rules, regulations and guidance would have to follow the release of this new policy. If the roll-out of the Common Rule has been any indication, the community cannot be idly waiting for government regulations as this has the potential to slow down the speed and increase the costs of true biomedical innovation.
- The data plan becoming a Term and Condition of the Notice of Award is not on its own objectionable, however, there should be some explicit acknowledgement of the indemnification of investigators and institutions if a compliance failure were to occur as a result of something beyond their control. For example, if due to the nature of their research, an investigator was utilizing a niche cloud-based data repository and the data housed in such repository were to be compromised due to no fault of the investigator, the investigator or institution should not be penalized through enforcement action, additional special terms and conditions or award termination.
- There should be a semblance of a dispute resolution structure whereby a rejected data plan that is deemed sufficient by an investigator or institution can be addressed.
- Individual requests for proposals (RFPs) should delineate clear standards and minimum expectations, relevant to the specific scientific discipline, necessary to meet the data management and sharing requirement. Where possible, templates and samples should be provided to illustrate effective and compliant plans.
- There is mention of requiring when an investigator will make their data available to secondary data users, however a data management plan should not be required for NIH funded projects that are analyzing secondary data.
- There appears to be no flexibility for an investigator or institution to change a data plan once it has been accepted. This should be reconsidered and included in any new policy.
- A portion of the policy should be dedicated to intellectual property (IP) as it is not clear how such matters would be handled as it relates to the overall data plan.
- While we favor the creation of federally sponsored repositories of specific, high value data types, we also favor the creation of a code repository. However, we have concerns about linking repositories to politics and therefore favor public/private partnerships to create and manage open source code and data repositories that can survive with “lights on” government funding and flourish with support of industry and other third parties.
- We suggest exploring the feasibility of a single institutional data management and sharing policy that is flexible, adaptable and takes into consideration the type of research and data management systems and repositories that currently exist. While some specific variation might be required for the

specific science of any given RFP, an institution-level approach could be more economically advantageous compared to requiring individual data plans for each individual award.

As it currently reads, a policy of this nature would require an institution like Weill Cornell Medicine to produce, monitor and maintain thousands of individual data plans. To accommodate this, institutional resources would have to be dramatically shifted and congressionally appropriated research funds would have to be limited to make funds available for these plans. This is especially troublesome as the recent laudable increases to the NIH budget are by no means permanent nor predictable. Allocating research funds for data management infrastructure and data repository fees would require an increase in Facility and Administrative (F&A) costs associated with NIH grants without providing an increase in programmatic research funding.

We again applaud and appreciate the opportunity to comment on this Proposed Provisions Draft and Request for Information. With data sharing at the core of scientific breakthroughs and NIH led programs such as the Precision Medicine Initiative's All of Us Research Program, Cancer Moonshot and BRAIN initiative, now is truly a crucial time to update the NIH's data management and sharing policy. However, we request the agency continue to take the transparent and pragmatic approach it appears to be taking to fully consider and weigh the concerns and recommendations detailed in this letter and the other comment letters submitted from across the academic medicine community.

If you have questions, or would like any additional information on anything aforementioned, please contact Weill Cornell Medicine's Director of Government and Community Affairs, Daniel C. Pollay Jr. at 646-962-9527 or dcp2003@med.cornell.edu.

Sincerely,



Curtis L. Cole, MD, FACP
Chief Information Officer
Associate Professor of Clinical Healthcare Policy
and Research
Associate Professor of Clinical Medicine



Hugh C. Hemmings Jr., MD, PhD, FRCA
Senior Associate Dean for Research
Chair of the Department of Anesthesiology
Joseph F. Artusio, Jr. Professor of Anesthesiology
Professor of Pharmacology

Submission #127**Date:** 12/10/2018**Name:** Amy Koshoffer**Name of Organization:** University of Cincinnati Libraries**Type of Organization:** University**Role:** Institutional Official**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

We are an R1 institution and have a very active research community across many disciplines.

I. The definition of Scientific Data

In general we agree with the definition offered in the Public Access Plan <https://grants.nih.gov/grants/NIH-Public-Access-Plan.pdf> . Simply put, scientific data is the information to be analyzed to address the research questions. However, scientific data is not inherently self explanatory. As part of the new policy, we would want to see a strong emphasis on defining best practices for data collection and documentation, data standards that are consistent across funding and government agencies (NIH, NSF, DOE, OSTP) and supporting documentation to help ensure that data collected is useful to current and future researchers within disciplines and in support of trans disciplinary research. Researchers should strive to create data that is FAIR as stated in the Force 11 data principles - <https://www.go-fair.org/fair-principles/>

II. The requirements for Data Management and Sharing Plans

In considering requirements surrounding data management and sharing plans, emphasis should be placed on several factors which include but not limited to: (i) creating quality data that can be reused by current and future researchers; (ii) focus on data that is also machine readable; and (iii) ensure that language used in the plans contain consistent definitions that have meaning across multiple organizations. The instructions in the RFP should also contain guidance language that encourage researchers to be very specific in their plans to manage their data. Since it is clear that all data cannot be shared, specific examples which highlight compliance should be put into place for organizations to use as a template in developing their own data management and sharing plans. Standards will play a key role in developing any data management and sharing plan and will highlight the variety of data types that consider disciplines and transcends them. Standards would also include complete methodology and

metadata formats including information on software used to create or capture data; and highlight the importance of creating citable datasets and software.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

Start with educational efforts and soft policy and move quickly into emphasis on standards and reproducibility (phased implementation) with increased infrastructure, incentives, then increased enforcement and consequences.

Support / Education

There should be more education, support resources and standards for documentation such as readme files and detailed protocols done either by grant agencies as agencies roll out policies or establish or use a system similar to the Network of National Libraries data management training program <https://nlnm.gov/classes/biomedical-and-health-research-data-management-librarians>. The type of training programs could vary and include options such as:

Repository of training components - i.e. earth science has Dataone - <https://www.dataone.org/data-management-planning>

Training could be live online - BD2K series <https://www.youtube.com/channel/UCKIDQOa0JcUd3K9C1TS7FLQ/videos>

Training could be a virtual module similar to John Hopkins Open Science trainings - <http://dms.data.jhu.edu/training/online-training/open-science-online-training/>

Trainings that travels (similar to NCBI) for people involved in data management and for the researchers (this would ensure consistency across organizations)

Pull resources and trainers from libraries across US

ACRL roadshow - <http://www.ala.org/acrl/conferences/roadshows/rdmroadshow>

Standards

NIH should align with what are other institutions are doing, and seek a common model that all agencies can follow so that there is consistency as it relates to standards. At the higher level there should be cross-talk between agencies such as NIH, NSF, NLM, DOE, and others. These standards should drive educational efforts and support documents such as this example from NASA. https://smd-prod.s3.amazonaws.com/science-red/s3fs-public/atoms/files/Data_Mgmt_Plan_guidelines-20110111.pdf

Also given the increase in academic institutional repositories that are available as a preservation option for data, there should be a focus on standards for all repositories,

particularly for institutional repositories and for data types. Research reproducibility is a major goal and is dependent on establishing standards for the various data types that consider disciplines and transcends them, complete methodology and complete metadata including information on software used to create or capture data.

Additional Infrastructure

The establishment of a national data catalog or data aggregator would let researchers know what research data is currently shared or will be generated and could be linked to open grants. There are great examples of federal government supported disciplinary repositories such as data.gov, Genbank, and clinicaltrials.gov. Also the national data storage platform (OSN) being developed by the NSF is a great initiative. There should be increased efforts such as this across agencies.

INCENTIVES:

We assume that in order for any type of data sharing process to be effective a defined policy must include clear guidelines that researchers can follow which include data standards; criteria surrounding reusability; ethical concerns surrounding data collection; and information of where current and future researchers can locate examples of what a good body of data management looks like. Processes should be put into place whereby researchers can gain clear feedback on their data management plans from NIH/NLM especially as it relates to how best improve future submissions with revisions to plans conditional to funding. Incentives should be considered when institutions have successfully implemented data management and sharing policies and are willing to share their techniques and processes with other institutions.

Form of incentives could be recognition and badges for individual researchers and institutions (<https://www.datasealofapproval.org/en/information/requirements/>) or supplemental awards and continued access to funding.

Goals

The major goal is the effective use of valuable resources that results in data worthy to preserve and share.

This will be aided by consistency across government organizations (NIH, NSF, DOE, OSTP, etc.) in support of best practices for research data especially in support of trans disciplinary research, long term access & preservation, and research reproducibility.

Kristen Burgess, Special Assistant to the Senior Associate Dean

Jane Combs, Associate Director UCIT-Research & Development

Lori Harris, Assistant Director Health Sciences Library

Amy Koshoffer, Science Informationist

Rebecca Olson, Social Science and Business Informationist

Submission #128**Date:** 12/10/2018**Name:** Daniel Shriner**Name of Organization:** National Human Genome Research Institute**Type of Organization:** Government Agency**Role:** Scientific Researcher**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Genetic epidemiology

I. The definition of Scientific Data

In III (Scope and Requirements) line 2, it is not clear whether “scientific data” includes data collected prior to NIH funding but analyzed with NIH funding. What about software code (see 3.2) used in the analysis? Both also relate to additional data to be included (see same paragraph).

II. The requirements for Data Management and Sharing Plans

It would be good to have some guidelines when to terminate storage. One opinion is that data preservation/archiving does not have an expiration date. Otherwise, at some future date, somebody could go looking for the data and not find it; this would be detrimental to reproducibility. Given that the NIH is publicly funded by taxpayer dollars, there is a fiduciary responsibility to preserve data. On the other hand, it is not possible for individual researchers or labs to commit to supporting data forever, and in any case the cost of storing everything forever could reduce funds to do new science. Researchers should state their timelines and give reasons. The time to terminate storage is clearly not retirement of the original researcher. It should be somehow bounded to the purpose of storage, not the purpose of the original project. The goal should be to encourage scientific sharing of data to minimize cost by not using NIH funded dollars to repeat the same/similar studies. It would make sense for data generated by NIH funds to be stored by NIH repositories, though it needs to be a priority to ensure that data are consistently formatted, processed, documented, and made available to the broader public. One problem with an NIH repository is that it is not exactly available to all researchers and what exactly is available is not well curated/documentated. A non-NIH repository could also be problematic given personal data.

In section 1.1 of Plan Elements, specific guidelines regarding raw or processed data could help to ensure that analyses can be replicated and data combined across studies.

In section 3 of Plan Elements, the document mentions standards and other data documentation. How far should this go? Some data in dbGaP seem to be rather limited in purpose to restrict use.

NIH is not international and thus does not follow European Union data protection laws. That means there are automatically restrictions on data originated from Europe to be stored there.

Submission #129**Date:** 12/07/2018**Name:** Felice J. Levine**Name of Organization:** American Educational Research Association**Type of Organization:** Professional Org/Association**Role:** Member of the Public**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

education research; social and behavioral sciences

I. The definition of Scientific Data

AERA recommends the inclusion of metadata in the definition of scientific data. In addition, we encourage including data documentation in the definition of a data management plan. An explicit reference to data documentation would require researchers to document actions such as the coding of variables and methods of analysis. Such documentation is central to research transparency, reproducibility, and replication research. Doing so will also allow other investigators to build upon and extend studies in their research.

II. The requirements for Data Management and Sharing Plans

AERA supports the proposals for NIH data management and sharing plan requirements. We encourage NIH to provide flexibility for individual institutes and centers (IC) to supplement (but not override) the NIH-wide policy with data management and sharing plans that apply to the particular aspects of the scientific disciplines and fields supported by each IC. One model for implementation is the National Science Foundation (NSF), where several directorates – including Education and Human Resources and Social, Behavioral, and Economic Sciences – have included data management plan guidelines in their funding solicitations that supplement the NSF-wide data sharing policy. Grant applicants must include a data management plan per the NSF Proposal & Award Policies & Procedures Guide but also adhere to the guidelines and expectations of the directorates.

AERA recommends the following guidance for specific plan elements, noted by the numerical item in the proposed requirements:

4.1 : Provide examples to grant applicants of existing repositories supported by NIH (e.g., NIH Data Commons, NICHD DASH) and external to NIH (e.g., Inter-university Consortium for Political and Social Research, Databrary).

4.2 : Require use of a persistent identifier such as a Digital Object Identifier. Researchers conducting reproducibility or replication studies would easily find and cite data being used in their published work, providing attribution to the NIH-supported grantees for the initial data collection and analysis.

4.5 and 4.7: Clarify the goals for both items to encourage sharing data in archives for access on restricted-use basis, so long as confidentiality and privacy protections are in place through a memorandum of understanding or other similar agreements.

5: Encourage researchers to include any costs for data management and preservation in their budget proposals. The cost of data management should be regarded as essential to the research and be evaluated for adequacy and reasonableness along with other proposal costs.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

AERA would support the adoption of the final Data Management and Sharing Policy for NIH Funded or Supported Research to be applied to the subsequent cycle of Funding Opportunity Announcements. The 2003 NIH Data Sharing Policy expanded awareness and promoted a culture of change across the scientific and health communities to document data processes and share and archive data, but the policy only applies to a subset of NIH funding. To educate the entire NIH community about developing data management plans, we encourage updates to the documents listed as guidance for submitting a data sharing plan under the 2003 policy (https://grants.nih.gov/grants/policy/data_sharing/) to be issued prior to or at the same time as the initial cycle for which the new policy would apply.

Attachment:

December 10, 2018

Dr. Carrie Wolinetz
Acting Chief of Staff and Associate Director for Science Policy
Office of Science Policy
National Institutes of Health
6705 Rockledge Drive, Suite 750
Bethesda, MD 20892

Dear Dr. Wolinetz,

Thank you for the opportunity to comment on the Request for Information on Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research. The American Educational Research Association (AERA) is the major national scientific association of more than 25,000 faculty, researchers, graduate students, and other distinguished science professionals dedicated to advancing knowledge about education, encouraging scholarly inquiry related to education, and promoting the use of research to improve education and serve the public good. Many of our members receive funding from the National Institutes of Health (NIH) for fundamental research in areas such as understanding learning processes and the intersection of health and education outcomes.

We commend NIH for developing these proposed provisions as part of the agency's commitment to open science and as a steward of the federal investment in scientific research. NIH has been at the forefront of promoting the sharing and use of scientific data, and we appreciate the work that NIH is undertaking to continue building a culture of data sharing consistent with the FAIR (findable, accessible, interoperable, and reusable) principles.

The following responses were also submitted to the corresponding fields on the [RFI website](#):

I. The definition of Scientific Data

AERA recommends the inclusion of metadata in the definition of scientific data. In addition, we encourage including data documentation in the definition of a data management plan. An explicit reference to data documentation would require researchers to document actions such as the coding of variables and methods of analysis. Such documentation is central to research transparency, reproducibility, and replication research. Doing so will also allow other investigators to build upon and extend studies in their research.

II. The requirements for Data Management and Sharing Plans

AERA supports the proposals for NIH data management and sharing plan requirements. We encourage NIH to provide flexibility for individual institutes and centers (IC) to supplement (but not override) the NIH-wide policy with data management and sharing plans that apply to the particular aspects of the scientific disciplines and fields supported by each IC. One model for implementation is the National Science Foundation (NSF), where several directorates – including Education and Human Resources and Social, Behavioral, and Economic Sciences – have included data management plan guidelines in their funding solicitations that supplement the NSF-wide data sharing policy. Grant applicants must include a data management plan per the NSF Proposal & Award Policies & Procedures Guide but also adhere to the guidelines and expectations of the directorates.

AERA recommends the following guidance for specific plan elements, noted by the numerical item in the proposed requirements:

4.1: Provide examples to grant applicants of existing repositories supported by NIH (e.g., NIH Data Commons, NICHD DASH) and external to NIH (e.g., Inter-university Consortium for Political and Social Research, Databrary).

4.2: Require use of a persistent identifier such as a Digital Object Identifier. Researchers conducting reproducibility or replication studies would easily find and cite data being used in their published work, providing attribution to the NIH-supported grantees for the initial data collection and analysis.

4.5 and 4.7: Clarify the goals for both items to encourage sharing data in archives for access on restricted-use basis, so long as confidentiality and privacy protections are in place through a memorandum of understanding or other similar agreements.

5: Encourage researchers to include any costs for data management and preservation in their budget proposals. The cost of data management should be regarded as essential to the research and be evaluated for adequacy and reasonableness along with other proposal costs.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards.

AERA would support the adoption of the final Data Management and Sharing Policy for NIH Funded or Supported Research to be applied to the subsequent cycle of Funding Opportunity Announcements. The 2003 NIH Data Sharing Policy expanded awareness and promoted a culture of change across the scientific and health communities to document data processes and share and archive data, but the policy only applies to a subset of NIH funding. To educate the

entire NIH community about developing data management plans, we encourage updates to the documents listed as guidance for submitting a data sharing plan under the 2003 policy (https://grants.nih.gov/grants/policy/data_sharing/) to be issued prior to or at the same time as the initial cycle for which the new policy would apply.

Additional feedback

In addition to the comments above, AERA encourages NIH to develop guidelines for data management and sharing plans consistent with other federal agencies as applicable. NIH serves as a vital source of grant funding for education researchers along with the Institute of Education Sciences (IES) and NSF. Both agencies have recently begun implementing data management plan requirements in their funding opportunities to promote research transparency. We recommend that NIH consider the IES and NSF guidelines and harmonize areas where burden could be reduced for grant applicants.

Please do not hesitate to call on AERA if we can further help with this effort or provide additional information. I can be reached directly at the contact information below.

Sincerely,

A handwritten signature in black ink, appearing to read "Felice J. Levine". The signature is fluid and cursive, with the first name being the most prominent.

Felice J. Levine, PhD
Executive Director
flevine@aera.net
202.238.3201 (office)
202.262.7189 (cell)

Submission #130

Date: 12/10/2018

Name: Vince Mor and Julie Lima

Name of Organization: Brown University

Type of Organization: University

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

health services research

Attachment:

Responses to NOT-OD-19-014_ Proposed Provisions for a Draft Data Management and Sharing Policy

Submitted by Vince Mor and Julie Lima
Brown University
Center for Gerontology and Health Care Research
December 10, 2018

Definition of Scientific Data -- The recorded factual material commonly accepted in the scientific community as necessary to validate and replicate research findings including, but not limited to, data used to support scholarly publications. Scientific data do not include laboratory notebooks, preliminary analyses, completed case report forms, drafts of scientific papers, plans for future research, peer reviews, communications with colleagues, or physical objects, such as laboratory specimens. For the purposes of a possible Policy, scientific data may include certain individual 2 level and summary or aggregate data, as well as metadata.⁷ NIH expects that reasonable efforts should be made to digitize all scientific data.

No suggested edits.

Draft NIH Data Management and Sharing Policy

Plan Review and Evaluation

Comments

It was proposed that peer reviewers might consider the appropriateness of the plan during the grant review process. We recommend that NIH provide detailed guidance to peer reviewers about the standards for appropriate data management plans when projects have data use agreements with CMS or other such entities that restrict re-use and data sharing so that secondary data projects are not "penalized" in the review process.

Plan Elements

2. Related Tools, Software, and/or Code - Indicate what software/computer code will be used to process or analyze the scientific data (the inclusion of scripts may be helpful), why the software/code was chosen, and whether it is free and open source. If software/code that is not free and open source is needed to access or further analyze the scientific data, briefly describe why this particular software/code is needed. Describe whether there is an alternative free and open source software/code that may be used to further analyze the scientific data.

Comments

A great many researchers use commercial packages such as SAS and Stata, in part because their background, training, and current set of duties do not include maintaining an open source

platform. Is NIH working towards a policy requiring the use of Open Source software in any broad area where it exists, regardless of existing infrastructure and practice built around commercial software? If so, this poses its own complications (discussion for another time), but even if not we believe the proposed requirement to describe whether there is an alternative free and open source software/code that may be used to access or further analyze the scientific data goes beyond the responsibilities of the researcher doing the original analyses. Without a substantial investment in time, it would be very hard to know what different software packages are available and what specifically they can do. What seems reasonable is to point to the general specs, licensing and purchasing details for the commercial product that was used (e.g. by providing the product's website) in order for new users to examine their own feasibility of acquiring it for their own use. It would allow them the opportunity to include the cost in the budgets of their own grant applications seeking to re-use the scientific data in question.

4. Data Preservation and Access...

Comment to section 4 in general

We seek clarification as to whether the analysis of administrative and other data collected apart from the research enterprise is considered data to be preserved, and would suggest that it is not the responsibility of the researcher to preserve such data, but rather the data provider. This interacts well with HIPAA requirements that would disallow direct sharing of such data, since ownership and control necessarily remains with the data provider. We would suggest that the requirements be amplified to include analysis code and other research methodology documentation as part of the data sharing plan, especially in cases where the underlying data cannot be shared due to privacy or other concerns. (E.g., IP rights of the data collector.) We would also suggest that it be a requirement of federal agencies providing sensitive data released under a Data Use Agreement that they preserve the technique used to provide data for any research such that the same data in the format could be provided to other researchers at a future point in time. In the case of non-public data not provided by a federal agency we would suggest that the original data sharing plan include a statement as to how long the source data can be expected to be preserved by the data provider, and how to it was contacted and the data acquired when the research was performed.

4.4 Describe alternative plans for maintaining, preserving, and providing access to scientific data should the original Plan not be achieved.

Comments to 4.4

It would be reasonable to require researchers to say something generic such as that efforts will be made to explore new options at NIH's request if the original plan should not be achieved, but giving concrete alternative plans is difficult without knowing the reasons for the original failure.

6. Data Sharing Agreements, Licensing, and Intellectual Property

Comments

For data that simply cannot be re-used (per HIPAA requirements, e.g.), the process is straightforward – the researcher can refer new users to the covered entities that provided the original data (e.g., CMS) new users can undergo their own access request process.

For other types of data such as primary data collected under an NIH grant that should be made publically available to others, we suggest that the burden lies with NIH to enter into data sharing/licensing agreements with new users to ensure data confidentiality and security are being maintained by new users.

7. Oversight of Data Management

Comments

Once data are submitted into a repository by the original researchers, we believe that data sharing and data distribution should not be the responsibility of the original researcher. NIH should either take ownership of the process within its own repository option, or else, researchers should ensure that the repository chosen for storage and distribution is equipped to handle requests for future download and use.

Submission #131**Date:** 12/10/2018**Name:** Erin Garrison**Name of Organization:** Tribal Nations Research Group**Type of Organization:** Nonprofit Research Organization**Role:** Institutional Official**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Native American research.

I. The definition of Scientific Data

The Collaborative Research Center for American Indian Health has reviewed the proposed changes along with tribal nation collaborators and researcher. The consensus in terms of the definition of scientific data was that it was lacking in addressing cultural knowledge, Indigenous ways of knowing, and plans to address protected populations (i.e. pregnant women, fetus). Tribal nations already see researchers coming in and disregarding cultural knowledge, and we feel that the NIH definition of scientific data MUST address this aspect. A statement about tribal sovereignty and a tribe's ability to define scientific data how they see applicable is also suggested. It should also be clear that any data collected with a tribe belongs to the tribe and they have the right to dictate how it is used.

II. The requirements for Data Management and Sharing Plans

A statement about tribal sovereignty and a tribe's ability to require more in terms of data management and sharing plans should be included. Tribal research data is unique and needs to have additional protections and allowances for tribes to exercise data sovereignty. An additional unique aspect of tribal data is in regard to identifiable data. Tribal nations must be allowed to approve/disapprove of having their tribe identified via data in (not limited to) publications, presentations, reports, etc. When it comes to multi-center projects, there must be a way for tribes to exclude any identifiable data (if it is collected - such as tribal nation, district, village, city, reservation) from reports, presentations, publications, etc. Any restrictions to data sharing should be made up front and be known to both the researcher and the tribal nation and if a researcher plans to share any data in any way, that must be disclosed up front or when the researcher becomes aware of intent to share. No data should be shared without prior tribal approvals. Regarding large tribal data-sets - the tribe/tribal IRB may make the decision to house

the data, and if this is the case, it should be the first option for location to house the data; to access these data-sets, the tribe/tribal IRB must give permission to researchers for secondary data analyses. A tribal nation also has the right to enforce fines, fees, and restrictions of researchers that do not follow the tribe's requirements and this should be specifically stated. The required page limit may need to be increased. Management of genomic, not identifiable data should be addressed.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and

Any implementations must keep tribal nations in mind. Any new requirements could put new burdens on tribal nations, requiring more time to and resources to prepare for the changes. It is vital to work with tribal nations throughout this process.

Submission #132**Date:** 12/10/2018**Name:** Dushanka Kleinman and Mary Shelley**Name of Organization:** University of Maryland School of Public Health**Type of Organization:** University**Role:** Institutional Official**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Biomedical, behavioral, and public health

I. The definition of Scientific Data

The definition of scientific data is necessarily broad and should cover all information used to address a proposed scientific question or aim.

NIH's Data Management and Sharing Policy should distinguish between primary and secondary data and provide corresponding guidelines and requirements for dissemination. Increasingly, investigators integrate and synthesize diverse forms of data from multiple sources (including secondary sources), and investigators may not have the rights to redistribute data which they themselves did not collect but which are integral to a study's findings.

II. The requirements for Data Management and Sharing Plans

Since an investigator may not be allowed to share all the data for a given study due to confidentiality concerns, data use agreements, etc., documenting which data sources are used and how they are combined in a given study is key to ensuring reproducibility of study results and reusability of materials. Thus, sharing metadata and computer code related to data management and analysis should be keystone requirements of any data management plan in order to document provenance of results.

Data, metadata, and code should be shared in a machine readable format.

NIH guidelines for data management and sharing plans should be broad and flexible enough to allow researchers from a range of institutions with a range of technical capacities to compete for most awards.

NIH should provide guidance and resources to enable investigators to share data, code, and metadata in a comprehensive archive with minimal ongoing financial or time burden to

investigators to comply with the mandate. Examples of forms this support could take include provision of a centralized repository for archiving funded studies, additional funds for depositing study data in existing repositories, and maintaining a list of partner repositories.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

Phasing of data management and sharing plan requirements is essential. Standards themselves should be phased in with attention to at least two dimensions: time and scale of research. As technological capabilities evolve, so will standards. Knowing what NIH expects in the short, medium-term, and longer-term will greatly help institutional planning regarding resources and infrastructure. At the same time, large projects, especially those involving large and/or heterogeneous and or CUI data and/or computationally intensive research, will require different standards than smaller-scale projects using unprotected data conducted by smaller teams. NIH should provide a checklist of requirements and anticipated timings at the intersection of these dimensions to assist institutions in scoping and deploying the technical resources and training investigators will need to meet evolving requirements for multiple projects at a variety of scales.

Submission #133

Date: 12/10/2018

Name: Denis Wirtz

Name of Organization: Johns Hopkins University

Type of Organization: University

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

Research

Attachment:

December 10, 2018

Carrie D. Wolinetz, Ph.D.
Office of Science Policy
National Institutes of Health
6705 Rockledge Drive, Suite 750
Rockville, MD 20892

RE: Response to Request for Information (RFI) on Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research (NOT-OD-19-104)

Dear Dr. Wolinetz:

Please accept this letter as the response of the Johns Hopkins University (“Johns Hopkins”) to the Request for Information (RFI) on the above captioned proposed provisions. Johns Hopkins understands the value and importance of making NIH funded research results available to members of the research community when consistent with other laws and obligations (including to research subjects) and is committed to pursuing effective practices for sharing data resulting from NIH funded research to the research community and the public.

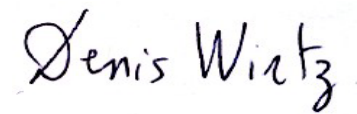
Improved access to research data offers enormous benefits to science and society, but must be managed in a way that retains the public’s confidence and willingness to participate in research studies. Properly curating and sharing data with appropriate protections for privacy is key to assuring that the fruits of the taxpayer’s funds are maximized. We commend NIH’s efforts to ensure public access through this RFI and support several of the suggestions made in the draft data management and sharing policy. The Association of American Medical Colleges (AAMC), Association of American Universities (AAU) and the Council on Governmental Relations (COGR) have submitted insightful responses and we strongly endorse their suggestions. In addition, we ask NIH to consider the following:

1. Prior to implementing the proposed changes, NIH should select a few of the changes to pilot test to assure that they have the desired effects and that there are no unexpected problems implementing such requirements system wide. As in other fields of scientific endeavor, testing in a small sample can yield better design.
2. There are other NIH grant provisions or other federal policy that may impact data sharing. We urge the NIH to consider how it will balance concerns it has expressed about foreign collaborations with NIH funded research. If data sharing will be subject to any restrictions for foreign requesters and collaborators then NIH should explicitly state what approvals will be required to share the data.

Johns Hopkins appreciates the opportunity to provide input on the RFI and hopes that you will work with universities and university associations, such as AAMC, COGR and AAU, to ensure

that any data sharing provision implemented proves to be manageable and truly effective in achieving the intended purpose.

Sincerely,

A handwritten signature in black ink that reads "Denis Wirtz". The signature is written in a cursive, slightly slanted style.

Denis Wirtz

Vice Provost for Research, Johns Hopkins University

Theophilus H. Smoot Professor in Engineering Science

Departments of Chemical and Biomolecular Engineering, Oncology and Pathology

Director, Johns Hopkins Physical Sciences in Oncology Center

Director, NCI postdoctoral training program

Director, NCI predoctoral training program

Submission #134**Date:** 12/10/2018**Name:** Karen Estlund, Robyn Reed, Cynthia Hudson Vitale,**Name of Organization:** Pennsylvania State University Libraries**Type of Organization:** University**Role:** Institutional Official**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Pennsylvania State University Libraries feels strongly that data that results from scholarship and research produced on the Pennsylvania State University campuses are an intellectual output that must be curated, preserved, and annotated in such a manner as it is findable, accessible, interoperable, and reusable (FAIR). There is not a specific research area that is most important; the importance is the ongoing accessibility of datasets themselves, which benefit the commonwealth as part of Penn State's land grant mission. Penn State University has, as part of its strategic plan, the foundational categories of enhancing global engagement and driving economic development. Public access to datasets and scholarship facilitates global equity to information and can drive innovation, job creation, and spur entrepreneurship. Through these categories, Penn State Libraries, in collaboration with other institutional partners, have made a commitment to ensuring the intellectual record of the institution, including research data and other digital assets, is appropriately curated and made accessible.

I. The definition of Scientific Data

The definition of scientific data written in the Proposed Provisions for a Draft NIH Data Management and Sharing Policy aligns with similar definitions provided by the National Science Foundation, the scientific community, and other funding bodies. The NIH definition is somewhat limited, though, and lacks a clear definition, does not provide a description of a complete data package, and does not have specific timelines or guidelines for making research data available publicly.

We strongly recommend that the NIH thoroughly articulate their definition of data, including whether or not the data they expect to be shared includes null results. Failed experiments are rarely, if ever, published, yet, having a public record of work that was tried, but failed or produced null results, still has much intellectual merit and can potentially reduce scientific inefficiencies. Knowing what combination of variables were applied or attempted in a scientific

experiment allows future researchers to avoid those same null results or further modify the variables to test other components of the experiment. Thus, we would recommend that NIH be as specific as possible in how it defines data.

Penn State Libraries also recommends that NIH require a publicly available data package from funded researchers to further support the transparency and reproducibility of funded research. While sharing data and publications is useful, it is still an incomplete record of the research. For research to be verifiable and reusable, the analyzable dataset, protocols, metadata, data dictionary, statistical analysis plan, and analytic code should be made available in a packaged format or the full executable computational experiment be made available through a containerization application (such as Docker or ReproZip).

Further, we encourage the NIH to articulate a timeframe for making data publicly available. Currently the definition states in a “timely manner.” While flexible, it allows for too broad of an interpretation and could be misused. Rather, the NIH should be clear about their expectations for when data are to be made available and when delay is allowed. Penn State Libraries recommends that the NIH adopt a data release policy that is similar to the Patient-Centered Outcomes Research Institute (PCORI) or the NIH Public Access Policy. PCORI requires that funded researchers deposit their data in a repository on the same day that they submit their final report. Further, PCORI encourages awardees to curate their datasets in collaboration with repository staff. Curation is an essential component of the data publication workflow that ensures that the data are FAIR for the long-term. Many institutions have library staff and faculty who specialize in appropriately curating research data. Penn State Libraries recommends that the NIH support faculty in working with repository staff, whether inside the institution or at the repository organization, to meet curation requirements.

II. The requirements for Data Management and Sharing Plans

The Proposed Provisions for a Draft NIH Data Management and Sharing Policy recommends a variety of different ways that a NIH IC should consider for plan review and evaluation. Ultimately, none of the recommendations for review impact the overall award score of the scientific project. Rather, the data management and sharing plans are evaluated independently of the scientific project and in three out of the four ways, are reviewed not by peers, but by NIH administrators, technical specialists, and in the context of other awards. While the remaining recommendation includes a review in the larger scientific review process, it doesn’t factor into the award impact score. This recommendation, currently listed in Extramural Grants, states that NIH staff will work with the awardee to address any reviewer concerns with compliance integrated into the terms and conditions. While this is a wonderful aspiration, we worry about the administrative overhead this would place on NIH staff. Rather, we would recommend that the NIH have PI’s work with their institutions, and specifically with data curators or data management specialists, to develop a reasonable and appropriate data management plan. This would leverage local expertise, lessen NIH administrative burden, and ensure that the plan reflects local capabilities for proper stewardship and preservation.

DMP Components

Data Type

Penn State Libraries appreciates that the NIH has required that researchers not only articulate the types of data that will be produced, but also an estimate size of the amount of data that will be produced. As institutions have become an archive for data created on their campuses, having a method to estimate how much data faculty may need to archive or steward from the outset of the project is valuable planning information.

Data Preservation & Access

Regarding data preservation, we suggest requiring specific file types, pointing to directory of allowed types, and/or specifying that they are non-proprietary and widely-used formats. It is important to encourage the use of machine readable/machine actionable formats for storage.

A few more details about data preservation and access should be included in these NIH guidelines. If NIH would like to encourage or require a data repository-generated permanent identifier or a Digital Object Identifier (DOI), please include this information in NIH guidance. The NIH may have concerns about a repository no longer being available over time and should make suggestions of acceptable repositories or guidance on what to consider when selecting a repository. Additionally, the NIH should include information about assigning metadata to datasets and require the tagging of datasets with Medical Subject Headings (MeSH), as this will aid in the organization of the data. Widely used metadata schemas that address discoverability, preservation, or domain interoperability should be required.

Data Preservation & Access Timeline

One possible concern in data sharing is the timeline. We would suggest being more specific in defining when data are required to be available publicly. A logical timeline would be at the time of final report submission but NIH should consider allowing an embargo period following publication dates.

Data Sharing Agreements, Licensing, and Intellectual Property

Penn State Libraries recommends that the NIH require funded researchers to release datasets, if copyrightable, with Creative Commons licenses to support the broadest impact of the research. We recommend that researchers release datasets with CC BY 4.0, CC BY-NC 4.0, or CC0 licenses.

Compliance and Enforcement

We appreciate the additional details that the NIH has provided in regards to compliance and enforcement. The recommendations support the importance of proper data management

during the research project and upon its completion. We recommend that the NIH create mechanisms for researchers to be able to modify the contents of the data management and sharing plan. Throughout the course of a research project unforeseen changes may occur which would make adherence to the original plan difficult or impossible. These kind of exceptions, while not expected or frequent, should be planned for.

We also encourage NIH to adopt guidance similar to the Wellcome Trust's policy on data sharing by encouraging researchers to review their data management plan throughout the research project lifecycle, as well as set the expectation that, "all users of research data, software and materials cite the source, and abide by the terms and conditions under which they were accessed." For data and related research outputs to be considered first-class research assets, tracking impact and reuse is essential – and will only come about through a normalization of data citation.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

We would encourage NIH to have at least one sample of an acceptable data management plan or template on their data management plan requirements website. Additionally, researchers or academic research office staff would benefit from having contact information for further guidance when questions arise.

We encourage the NIH to become leaders in the broad sharing and dissemination of data management plans and funded research proposals. Rather than being static documents, we recommend that the NIH mandate machine-actionable data management plans. Through machine-actionable plans, many compliance checks would be automated, author and institutional disambiguation would be eased, and many components of the DMP would be standardized.

Lastly, we appreciate how the proposed draft noted that associated costs with meeting these requirements could be offset by grant funding. This is a common question that researchers would ask from the onset. We encourage NIH to leave a statement about the associated costs in their final proposal.

Submission #135

Date: 12/10/2018

Name: Laure Haak

Name of Organization: ORCID, Inc

Type of Organization: Other

Other Type of Organization: Research Infrastructure provider

Role: Institutional Official

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

All areas. Interested in the process of information sharing and attribution regardless of research area.

II. The requirements for Data Management and Sharing Plans

Plan elements - section 2. Related Tools, Software and/or Code. Include reference to the event where custom tools, software and/or code was developed for the purpose of analyzing or interpreting the data. Include reference to the individuals that authored these custom tools, including the ORCID iD of these contributors. (The inclusion of contributors of custom software will aid in achieving the goals for sharing scientific data, by increasing the possibility to consult with tool contributors, an important capability since the documentation accompanying custom software/tools can include less robust documentation than that found for commercially- or open source-available items.)

Plan elements - section 7. Oversight of Data Management. Include reference to unique person identifiers to precisely identify those involved and eliminate confusion. For example, "Indicate the individual(s) with their ORCID iDs or other unique identifier(s) who will execute various components of the Plan..." If not included in section 2, include reference to individuals that may have contributed custom software or tools.

Submission #136**Date:** 12/10/2018**Name:** James Glazier**Name of Organization:** Indiana University**Type of Organization:** University**Role:** Scientific Researcher**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Developmental Diseases

I. The definition of Scientific Data

Some of the most productive and exciting areas of contemporary science and engineering result from the increasingly tight integration of classically defined data with tools for data processing (in biology we would think especially of bioinformatics) and simulation (i.e., mechanistic predictive models of living or designed systems). Such integration is central to new technologies that blend data-sources, modeling and delivery to deliver assisted navigation via cell phones, self-driving cars, and other smart systems. Similar approaches could enable development of technologies like virtual digital twins that monitor an individual's health state and predict adverse health events before they occur. All these technologies depend on the treatment of computer programs, abstract models and workflows (which I will generally refer to as modules) as types of data, and similarly, on the assembly of classical 'data' into higher level modular structures that can be flexibly reused and combined. Treating models, workflows and software (modules) as data, with appropriate standards for specification, interconnection, reuse, adaptation and extension greatly facilitates knowledge capture and generation. Appropriate harmonization of standards for 'data' and 'module' description can enable, e.g., automated generation of executable computer simulations from experimental or clinical data (e.g. to predict the rate of growth of a tumor observed using MRI subject to various treatment options). Restricting our definition of data to be only the raw and processed outputs of experiments limits our ability to apply modern intelligent-systems approaches in biomedicine. Defining data broadly to include, conceptual models, software tools for data science, mechanistic simulations and workflows (modules) has significant payoffs. Ultimately, we can describe this philosophy as "annotated data are modules—modules are data."

II. The requirements for Data Management and Sharing Plans

Current NIH-required sharing plans for software and computational models are inadequate to enable the transformative data-software technologies under development in other areas. Publishing computer code (open sourcing) is a necessary start, but does not address the encapsulation and standardization of functionality that an integrated approach requires. At a minimum, if NIH wants to develop software tools and models that have value beyond the laboratory of the inventor (or beyond a small circle of experts), modules need to be treated as data—annotated, modularized and presented in a way that they can be searched, underlying concepts mapped and extracted, components extracted, reused, combined, and extended:

- 1) Follow standards for software and model specification when they already exist.
- 2) Be modularized (broken down into the smallest functional units) with well-defined APIs (connection standards for each module defining inputs and outputs). That is, tools should be conceived as sets of component modules individually designed to support reuse.
- 3) Embed descriptions of their underlying conceptual models (annotated using the appropriate ontologies) so that the conceptual models which gave rise to each module are recoverable from the module. These annotations should be searchable, so that modules don't need to live in repositories ("the web is the repository").
- 4) Specify purpose and range of validity of the modules and appropriate data types and limitations (units, expected and allowed ranges, default values,...).
- 5) Embedded parameters and concepts should have their sources (journal articles, experiments,...) specified in the modules that use them.
- 6) Two key guiding principles to promote reusability and knowledge capture are to describe function at the most human-intelligible level possible (declarative or functional specification in preference to object-oriented specification in preference to procedural code, visual representations and controlled natural language representations in preference to computer code).
- 7) Documentation and validation should be an integral part of all module development. Modules without documentation are no more useful than a spreadsheet with numbers and formulae without descriptions of the meaning and significance of the contents. Ultimately, modules should include embedded functional validation and test information to allow users to trust their function.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

The key problem is bootstrapping: we lack sufficient standards and tools for modularization, annotation and distribution, so proper development of modules as data is difficult. However,

the motivation for standard and tool development is lacking until there is a base of researchers to use them. Focused efforts like the human genome project, which led to the development of the standards and support tools that enable this approach in –omics and the current Human Cell Atlas projects, can provide this bootstrapping. In areas which are not subject to such large-scale coordinated effort, there is still the possibility of phased development:

- 1) Require open-source software development with embedded standards-based (here human-readable and web searchable ontology-based) descriptions of function to enable the use of modules as data.
- 2) Rework innovation criteria in review to support module reuse and extension rather than reinvention (current innovation guidelines penalize those who follow best practice in software development which maximizes reuse and adaptation).
- 3) Encourage modularization and the employment of standard description languages for module development.
- 4) Develop and enforce a standard of best-practice for module development and distribution.
- 5) Develop and enforce a standard for rigor and reproducibility for modules (embedded testing).

Attachment:

Section I —

Define Data — Some of the most productive and exciting areas of contemporary science and engineering result from the increasingly tight integration of classically defined data with tools for data processing (in biology we would think especially of bioinformatics) and simulation (*i.e.*, mechanistic predictive models of living or designed systems). Such integration is central to new technologies that blend data-sources, modeling and delivery to deliver computer-aided design of airplanes, assisted navigation via cell phones, self-driving cars, and other smart systems. Similar approaches could enable development of technologies like microscopes that capture tissue images, simulate the development of the tissue and compare the experimentally observed development to the predicted development, virtual digital twins that monitor an individual's health state and predict adverse health events before they occur, allowing health maintenance rather than reactive medicine, platforms for the development of optimized personalized therapies and tools for the design of maximally informative experiments. All these technologies depend on the treatment of computer programs, abstract models and workflows (which I will generally refer to as software modules or just *modules*) as types of data, and similarly, on the assembly of classical 'data' into higher level modular structures that can be flexibly reused and combined. Treating models, workflows and software (modules) as data, with appropriate standards for specification, interconnection, reuse, adaptation and extension greatly facilitates knowledge capture and generation. Appropriate harmonization of standards for 'data' and 'module' description can enable, *e.g.*, *automated* generation of executable computer simulations from experimental or clinical data (*e.g.* to predict the rate of growth of a tumor observed using MRI subject to various treatment options). Restricting our definition of data to be only the raw and processed outputs of experiments limits our ability to apply modern intelligent-systems approaches in biomedicine. Defining data broadly to include, conceptual models, software tools for data science, mechanistic simulations and workflows (modules) has significant payoffs. Ultimately, we can describe this philosophy as "annotated data are modules—modules are data." Where such approaches have been followed (as in the workflows for processing high-throughput -omics data), they have been transformative in biology.

Section II —

Current NIH-required sharing plans for software and computational models are inadequate to enable the transformative data-software technologies under development in other areas. Publishing computer code (open sourcing) is a necessary start, but does not address the encapsulation and standardization of functionality that an integrated approach requires. At the moment, most of the knowledge that goes into developing modules is lost, because the knowledge remains in the mind of the developer—it is not embedded in the module itself. A key philosophical change is to try to minimize the information loss at each step in the translation from an idea in the mind of a scientist into its software instantiation as a module. At a minimum, if NIH wants to develop software tools and models that have value beyond the laboratory of the inventor (or beyond a small circle of experts), modules need to be treated as data—annotated, modularized and presented in a way that they can be searched, underlying concepts mapped and extracted, components extracted, reused, combined, and extended. Developing modules in this way represents a significant break from current practice in the biomedical research community, but is common in large-scale industrial development (and is the reason that software ecosystems around

Google Android or Apple standards for applications and data flow or middleware standards for cloud computing are so productive). At a minimum, NIH-supported computational models and software need to:

- 1) Follow standards for software and model specification when they already exist.
- 2) Be modularized (broken down into the smallest functional units) with well-defined APIs (connection standards for each module defining inputs and outputs). That is, tools should be conceived as sets of component modules individually designed to support reuse.
- 3) Embed descriptions of their underlying conceptual models (annotated using the appropriate ontologies) so that the conceptual models which gave rise to each module are recoverable from the module. These annotations should be searchable, so that modules don't need to live in repositories ("the web is the repository").
- 4) Specify purpose and range of validity of the modules and appropriate data types and limitations (units, expected and allowed ranges, default values,...).
- 5) Embedded parameters and concepts should have their sources (journal articles, experiments,...) specified in the modules that use them.
- 6) Two key guiding principles to promote reusability and knowledge capture are to describe function at the most human-intelligible level possible (declarative or functional specification in preference to object-oriented specification in preference to procedural code, visual representations and controlled natural language representations in preference to computer code). As an example, Systems Biology Markup Language allows the description of chemical reactions as chemical reactions, which can carry with them the graphical representation of biochemical networks familiar from biology literature. Such descriptions retain the conceptual model of the biological process they represent, while a mathematical description of the same reactions in the form of ordinary differential equations, or a computational description in the form of Python or Matlab code do not. While an SBML description of a biological network can be compiled to generate a set of mathematical formulae or computer code, the mathematical formulae or computer code cannot be uncompiled to recover the biological networks that they represent.
- 7) Documentation and validation should be an integral part of all module development. Modules without documentation are no more useful than a spreadsheet with numbers and formulae without descriptions of the meaning and significance of the contents. Ultimately, modules should include embedded functional validation and test information to allow users to trust their function.

Section III —

Many researchers developing modules in the area of high-throughput –omics technologies already follow many of these guidelines. However, outside the –omics world, this more efficient approach to treating modules as data requires greater change. One problem is bootstrapping: we lack sufficient standards and tools for modularization, annotation and distribution, so proper development of modules as data is difficult. However, the motivation for standard and tool development is lacking until there is a base of researchers to use them. Focused efforts like the human genome project, which led to the

development of the standards and support tools that enable this approach in –omics and the current Human Cell Atlas projects, can provide this bootstrapping. In areas which are not subject to such large-scale coordinated effort, there is still the possibility of phased development:

- 1) Require open-source software development with embedded standards-based (here human-readable and web searchable ontology-based) descriptions of function to enable the use of modules as data.
- 2) Rework innovation criteria in review to support module reuse and extension rather than reinvention (current innovation guidelines penalize those who follow best practice in software development which maximizes reuse and adaptation).
- 3) Encourage modularization and the employment of standard description languages for module development.
- 4) Develop and enforce a standard of best-practice for module development and distribution.
- 5) Develop and enforce a standard for rigor and reproducibility for modules (embedded testing).

James A. Glazier, PhD
Director Biocomplexity Institute
Professor of Intelligent Systems Engineering
School of Informatics, Computing and Engineering
Indiana University
Bloomington, IN USA
Office: 812-855-3735 Cell: 812-391-2159 jaglazier@gmail.com

Submission #137

Date: 12/10/2018

Name: Anonymous

Name of Organization:

Type of Organization: University

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

Epidemiology

II. The requirements for Data Management and Sharing Plans

Major concerns are:

1. It is not at all clear whether participants in existing cohorts have actually provided informed consent to share their individual data in controlled-access or public access databases. Re-consenting may result in withdrawal from existing studies.
2. Collaborative data sharing models work well and are useful given the complex nature of epidemiological data. These model allows investigators who collect the data and use the data regularly to provide training and instruction on data analysis and interpretation. They also provide the added opportunity for training & mentoring of graduate students and postdoctoral fellows.
3. Making epidemiological data available through public or controlled-access databases could jeopardize funding opportunities for investigators, especially early stage investigators, and career opportunities. The presumption is that such data would be available to lower resource entities, but the reality is likely that well-funded and well-resourced entities would be able to more readily capitalize on the data.
4. For existing cohorts with ongoing data collection, the data are not static, so it is unclear when sharing should occur. How would changes to data be handled? For ongoing cohorts, there is never a "project completion date."
5. Who is responsible and what are the consequences for actual intentional or unintentional breach of privacy or confidentiality?
6. Misuse of data in public access or controlled access databases could jeopardize subject participation in existing cohorts.

Submission #138**Date:** 12/10/2018**Name:** Julie Palmer**Name of Organization:** Boston University**Type of Organization:** University**Role:** Scientific Researcher**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Epidemiology

II. The requirements for Data Management and Sharing Plans

I believe that sharing of scientific data is valuable and should be encouraged. We will advance science by maximizing the knowledge that can be gained from data produced by a single study. Because the field of epidemiology relies in large part on information obtained from study participants through interviews or questionnaires on a wide range of exposures, including some very personal and sensitive information, there are specific concerns that are not relevant to animal or basic science data which must be taken into account when setting guidelines for data sharing.

One example is related to informed consent. The NIH has funded a number of large cohort and case-control studies, some of which began 15 or more years ago, before advances in genetics, technology, and data science made possible widespread sharing of data through controlled-access repositories. While new epidemiologic studies will obtain written informed consent for data to be widely shared on publicly available databases, ongoing studies that began 10 to 40 years ago did not obtain informed consent for this type of activity. Requiring existing cohort studies to re-consent participants to allow wider sharing will likely result in only a fraction of the cohort giving written consent, at the cost of having another fraction of participants withdrawing altogether out of concern that their privacy will be jeopardized. To preserve continued active follow-up in the existing cohorts, it may be best to permit the cohorts to continue sharing through methods that are consistent with the original informed consent

Most epidemiologic studies, including the large cohort studies, have systems whereby outside investigators can access study data. An advantage of current systems in which a given study directly provides an outside investigator with an analytic dataset is that the outside investigator is more likely to receive help from study investigators in analyzing and interpreting the data,

something that would not occur under a system in which all data are stored on an NIH controlled-access database.

Submission #139

Date: 12/10/2018

Name: Lois Brako

Name of Organization: University of Michigan HRPP

Type of Organization: University

Role: Institutional Official

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

All health and behavioral research

I. The definition of Scientific Data

See attached file

II. The requirements for Data Management and Sharing Plans

See attached file

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

See attached file

Attachment:



12/10/18

Francis S. Collins, MD, PhD
National Institutes of Health
Bethesda, MD

Submitted electronically: <https://osp.od.nih.gov/provisions-data-managment-sharing/>

Re: RFI on Proposed Provisions for Draft Data Management and Sharing Policy

Dear Dr. Collins:

I am writing on behalf of the University of Michigan's Human Research Protections Program (HRPP). We acknowledge the challenges that NIH faces in proposing provisions for data management and sharing and we appreciate the opportunity to comment.

As a recipient of a significant amount of NIH support for our biomedical and social and behavioral human research across a broad range of disciplines, we can only provide comments to this RFI based on our experience with privacy and confidentiality concerns for research participants and our appreciation of the diverse methods used to collect, maintain, analyze, store, and share research data.

We are concerned that the current proposal is too general and broadly inclusive that it will be difficult for NIH to achieve the goal of advancing science without providing clear definitions of the data to be shared and examples to learn from, reasonable and flexible time frames for compliance, developing tools to assist with the process, providing guidance and supplemental funding to address new costs associated with compliance, and presenting a vision for adapting to changes in technologies associated with the collection, storage, and sharing of data. We highly recommend that NIH consider delaying this initiative until a broad group of NIH-funded investigators and data managers meet with the NIH policy leaders to revise and refine this proposal.

We provide comments on specific sections of the proposed policy below:

The Definition of Scientific Data

We strongly agree that preliminary analyses, lab notebooks, case report forms, and other early outputs of the research process should not be included in the definition of scientific data and that the timing for sharing data be flexible. Guidance on standards for data and metadata should be provided by NIH.

The Requirements for Data Management and Sharing Plans

We suggest that any requirements for a Data Management and Sharing Plans (DMSP) be harmonized across federal science agencies, or at least across NIH funding opportunities, with the exception of NIH funding mechanisms that would not generate scientific data (e.g., shared instrumentation awards, training grants, and conference support). While a brief description of the DMSP could be provided at the proposal stage, the NIH should allow for changes to this plan

as the project progresses. This might be monitored through Progress Reports. Templates for the original DMSP and Progress Report updates should be provided by NIH.

We agree with the comments submitted by AAMC that to help improve data sharing and licensing agreements, NIH might consider providing "a template licensing agreement that lays out sharing terms for data as it has previously done for biological materials."

We also share AAMC's and PRIM&R's concerns regarding the need for coordination between the informed consent process and the mandate for sharing data. For data collected prior to the implementation of the new policy, terms and conditions for data sharing must honor any promises of confidentiality made to participants in the consent document. As the AAMC letter mentioned, deidentification of data from human research may present additional costs to study teams or, in some cases, may not be possible. Regarding any additional costs to study teams, such costs should be provided as a supplement to the funding provided by NIH for the research, not under the budget for the proposed project.

Compliance and Enforcement

We feel strongly that NIH should develop the mechanism for data preservation and storage to allow the data to be as useful as possible to the scientific community without creating new regulatory burdens on the research community. We encourage the creation of a federal government repository for data preservation and sharing. Submission and storage of data through a central repository with a well-defined framework for data and metadata would allow NIH to more easily monitor and enforce compliance. We recommend that NIH clarify expectations for the timeline for data preservation and sharing to not exceed the funded project period.

Thank you for the opportunity to provide comments on the proposed draft policy.

Sincerely,

Assistant Vice President for Research - Regulatory and Compliance Oversight
HRPP Director

Submission #140

Date: 12/10/2018

Name: Margaret McCarthy

Name of Organization: Yale Collaboration for Research Integrity and Transparency

Type of Organization: University

Role: Bioethicist/Social Science Researcher

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

Integrity and transparency in medical product regulation, data sharing.

Attachment:



Collaboration for Research Integrity and Transparency

A PROGRAM OF YALE LAW SCHOOL, YALE SCHOOL OF MEDICINE AND THE YALE SCHOOL OF PUBLIC HEALTH

December 10, 2018

Via Electronic Submission

Office of Science Policy
National Institutes of Health
6705 Rockledge Drive, Suite 750
Bethesda, MD 20892

Yale Collaboration for Research Integrity and Transparency Comments on NIH's Proposed Provisions for Draft NIH Data Management and Sharing Policy, NOT-OD-19-014

The Yale Collaboration for Research Integrity and Transparency (CRIT) supports the effort to update the NIH's 2003 Data Sharing Policy. The NIH has been a leader in requiring data sharing for funded projects, for sharing data, and for creating and implementing standards that can be used across fields.

Below are our comments on the three specific areas on which comment is sought.

Section I The Definition of Scientific Data

We are concerned that the proposed definition of scientific data varies from the definition in the current *NIH Data Sharing and Implementation Guidance* (October 2003), which requires sharing of "final research data." Current NIH policies, including the *Frequently Asked Questions (FAQs) on Data Sharing* (February 16, 2004), the *NIH Grants Policy Statement: Access to Research Data* (2010), and the *National Institutes of Health Plan for Increasing Access to Scientific Publications and Digital Scientific Data Form HHF Funded Scientific Research* (February 2015), all refer to the requirement to



Collaboration for Research Integrity and Transparency

A PROGRAM OF YALE LAW SCHOOL, YALE SCHOOL OF MEDICINE AND THE YALE SCHOOL OF PUBLIC HEALTH

share “final research data” rather than “scientific data.” We urge the NIH to modify all current NIH policies to adopt the term scientific data in place of final research data. This will more closely mirror current NIH data sharing policy, and clarify that data sharing of final data, rather than preliminary or incomplete data, is required. We also urge the NIH to clarify that data sharing requirements for longitudinal and cohort studies begin with the completion of data collection for each wave or module, and that data sharing is required for clinical trials within 1 year of primary outcome ascertainment.

Section II The requirements for Data Management and Sharing Plans

We are fully in support of requirements for data management and sharing plans, and in particular, of the proposed requirement that a data management and sharing plan be made as part of the funding application process. Inclusion of a detailed data management and sharing plan at the proposal phase, at the level of detail contained in the proposed definition, is necessary to ensure that data is collected, managed and preserved using commonly accepted standards to ensure its eventual availability and usability by secondary users. The requirement contained in the proposed definition for identification of a data archive at the proposal stage is also welcome, and should be included in the final definition.

Section III The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards.



Collaboration for Research Integrity and Transparency

A PROGRAM OF YALE LAW SCHOOL, YALE SCHOOL OF MEDICINE AND THE YALE SCHOOL OF PUBLIC HEALTH

The current NIH general policy, in effect since 2004, requires data sharing for proposals seeking \$500,000 per year in direct costs. Although various NIH subsidiaries have included proposals seeking lower direct costs when imposing data sharing requirements, the current proposal marks a significant change by requiring data sharing for all proposals, regardless of annual direct cost amount. We applaud the NIH for removing the minimum of \$500,000 per year in direct costs from the requirement that data be shared. However, we recommend that the NIH carefully consider the best way to phase in data sharing requirements and ensure that already funded studies are eligible to request reasonable costs associated with data management and sharing.

The National Institutes of Health Plan for Increasing Access to Scientific Publications and Digital Scientific Data Form HIH Funded Scientific Research and the *NIH Strategic Plan for Data Science* contain information that can be used to develop detailed standards for implementation of the new data management and sharing policy.

Our experience with data archiving leads us to propose that the NIH institute specific requirements to ensure that data collected is of high enough quality to be acceptable to data archives, and to be usable by secondary analysts. Proposals should be required to include a budget for implementation of the data preservation and access requirements of this policy, including both ongoing work to create and preserve materials to be archived, as well as final preparation of the data for archiving.



Collaboration for Research Integrity and Transparency

A PROGRAM OF YALE LAW SCHOOL, YALE SCHOOL OF MEDICINE AND THE YALE SCHOOL OF PUBLIC HEALTH

We also urge the NIH to include a provision for the final payment of grant funds to be made upon the deposit and acceptance by the archive of the final research data from the project.

All too often, researchers may exhaust their grant funds prior to archiving data, resulting in either archiving of data that does not meet the standards for use by secondary researchers, or that does not meet archive standards for data deposit.

Finally, we urge the NIH create data repositories across all of its institutes to ensure that all NIH-funded research is shared by repositories employing best practices for data archiving, safeguarding and sharing, such as those employed by NHLBI's BioLINCC.

We appreciate the opportunity to comment on the draft policy.

Sincerely,

Margaret E. McCarthy
Executive Director

Submission #141

Date: 12/10/2018

Name: Sharad Verma

Name of Organization: NTAP at Johns Hopkins

Type of Organization: Nonprofit Research Organization

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

clinical, translational, discovery, NF1

II. The requirements for Data Management and Sharing Plans

1) That the policy for sharing also list SPORE grants amongst the groups. The language presented in paragraph 1 of section 3, and in the 4th bullet of section 4, seems to imply this, but it should explicitly state that the data management and sharing policy will apply to SPORE grants, which are major generators of data.

2) In section 4 under plan elements, it should be captured how data is collected (i.e. paper lab notebooks or e-notebooks), and details regarding data traceability back to original source.

Submission #142

Date: 12/10/2018

Name: Laura Thornhill

Name of Organization: Alzheimer's Association

Type of Organization: Patient Advocacy Organization

Role: Patient Advocate

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

Alzheimer's and related dementias

I. The definition of Scientific Data

The Alzheimer's Association applauds NIH's ongoing efforts to facilitate and improve the sharing of scientific data and supports the proposed definitions and requirements of a data management and sharing policy for all NIH-funded research.

II. The requirements for Data Management and Sharing Plans

We support the requirements for data management and sharing plans, but we note that NIH proposes only to require explanations of perceived barriers and not inclusion of possible steps to overcome those barriers or explanations of why they cannot be addressed. Powerful incentives to retain or withhold data exist in the current landscape such as career advancement and lucrative proprietary information, among others. We encourage NIH to require investigators and applicants to more carefully consider the barriers to sharing their information at the time of application.

Attachment:

Office of Science Policy
National Institutes of Health
6705 Rockledge Drive
Suite 750
Bethesda, Maryland 20892

December 10, 2018

Re: Request for Information on Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research

To Whom It May Concern:

The Alzheimer's Association appreciates the opportunity to respond to the National Institutes of Health's (NIH) Request for Information on Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research.

The Alzheimer's Association is the leading voluntary health organization in Alzheimer's care, support, and research. The Alzheimer's Impact Movement (AIM), the sister organization of the Alzheimer's Association, is a nonpartisan, nonprofit advocacy organization and works in strategic partnership with the Alzheimer's Association to make Alzheimer's a national priority. Today, there are more than 5 million Americans living with Alzheimer's, and it is the only cause of death among the top 10 without a way to prevent, cure, or even slow its progression. As the size and proportion of the United States population age 65 and older continue to increase, the number of Americans with Alzheimer's and other dementias will grow: without a disease-modifying therapy, as many as 16 million people may have the disease by 2050.¹ It is imperative that researchers have access to as much relevant data as possible to accurately diagnose individuals and to find a treatment.

The Alzheimer's Association applauds NIH's ongoing efforts to facilitate and improve the sharing of scientific data and supports the proposed definitions and requirements of a data management and sharing policy for all NIH-funded research.

We support the requirements for data management and sharing plans, but we note that NIH proposes only to require explanations of perceived barriers and not inclusion of possible steps to overcome those barriers or explanations of why they cannot be addressed. Powerful incentives to retain or withhold data exist in the current landscape such as career advancement and lucrative proprietary information, among others. We encourage NIH to require investigators and applicants to more carefully consider the barriers to sharing their information at the time of application.

The Alzheimer's Association appreciates NIH's direct linkage of the submission of data sharing plans to funding and support decisions. We recommend that reviewers have opportunities to be trained on how to appropriately evaluate these plans. In 2015, we launched the Global Alzheimer's Association Interactive

¹ Alzheimer's Association. (2015). *Changing the Trajectory of Alzheimer's Disease: How a Treatment by 2025 Saves Lives and Dollars*.

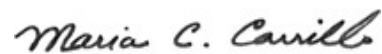
Network (GAAIN) (gaain.org), the first operational online integrated research platform, which links scientists, shared data, and sophisticated analysis tools. GAAIN has data partners and affiliates worldwide and over 500,000 unique clinical records from nearly 40 clinical studies available for interrogation. We would be pleased to partner with NIH and share in detail what we have learned from GAAIN to ensure that reviewers are well-qualified to evaluate data management and sharing plans alongside applications for funding.

Thank you for the opportunity to comment. The Alzheimer's Association and AIM would be pleased to work with NIH as it continues to promote data science and sharing. Please contact Laura Thornhill, Senior Associate Director, Regulatory Affairs, at 202-638-7042 or lthornhill@alz-aim.org if you have questions or if we can be of additional assistance.

Sincerely,



Robert Egge
Chief Public Policy Officer



Maria C. Carrillo, Ph.D.
Chief Science Officer

Submission #143

Date: 12/10/2018

Name: Fernando Rios, Chris Kollen

Name of Organization: University of Arizona - Office of Digital Innovation and Stewardship

Type of Organization: University

Role: Institutional Official

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

The Office of Digital Innovation and Stewardship supports research data management, including data management plans, across the university. The most important health-related research area in the organization is cancer research

I. The definition of Scientific Data

See attached

II. The requirements for Data Management and Sharing Plans

See attached

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

See Attached

Attachment:

University of Arizona – Office of Digital Innovation and Stewardship

Response to: RFI on Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research

Prepared by Fernando Rios and Chris Kollen

Our office provides research data management support – including preparing data management plans, compliance with funder data sharing requirements, and supporting FAIR research practices – to researchers across the University of Arizona. We provide feedback to the NIH RFI on Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research from that point of view.

I. The definition of Scientific Data

Comment on Scientific Data and Research Software

The definition of Scientific Data as stated does not include software since presumably, software is not generally considered “factual material”. However, Scientific Data as defined also mentions the notion of validating and replicating research findings. An important but often overlooked aspect of validation and replication is the software needed to carry out such replication and validation. It is the combination of data and the necessary analysis that address the reusability aspect of FAIR. Therefore, specifically including software as a sub-type of Scientific Data (or alternatively, as a fully separate and distinct kind of research output) would not only elevate it to the same level as data which, due to its importance is warranted, but would also serve to provide clearer guidance for what to do with software in terms of the proposed Data Management and Sharing policy. Although Section IV (2) of the proposed requirements addresses software, we believe it is insufficient.

A term “Scientific Software” could be defined and scoped so that it only includes software that is a research output (e.g., custom code implementing a novel algorithm, code for loading and data cleaning, statistical analysis scripts, simulation code, etc.) as opposed to commercial software or other software which was simply used as part of the research. By explicitly treating scientific software as another kind of output in addition to Scientific Data, it would become much clearer to researchers how to include this important research output in the scope of the Requirements (Section III) and under Section IV dealing with the requirements to describe and share these outputs. Section IV (2) dealing with related software and/or code should therefore be modified to give further requirements for describing and sharing novel or bespoke software. See Comments on the Requirements for Data Management and Sharing Plans for more explanation on these suggested modifications.

Comment on Definitions of Additional Terms

In addition, standard definitions of terms like “preserved”, “archived”, “persistent unique identifier”, and “metadata standard” should be included as these are often sources of confusion. For example, what many researchers consider “archiving” likely does not match what is meant in the context of the Data Management and Sharing Plan requirements.

II. The requirements for Data Management and Sharing Plans

Comments on the Requirements for Data Management and Sharing Plans

Comments are provided on specific aspects including research software (also see comments given in the previous section on Scientific Data), metadata and documentation, alternative plans for data management, and archiving.

1. SOFTWARE. In relation to the treatment of research software, it would be beneficial to expand the requirements to specifically mention software, thereby reflecting the importance of this research artifact. For example Section IV (2) does not appear adequate for dealing with software that was produced as part of the research (e.g., novel algorithm implementations, statistical analysis scripts, etc.) as this small section does not cover what is needed to make it FAIR. For example. Section IV (2) could be expanded to require description of the anticipated software that will be produced and how it will be documented. Section IV (4)-(7) would be expanded to explicitly mention software-related aspects. If software is included as a sub-type of Scientific Data as suggested previously, little actual modification to the text would be required. See the NASA Planetary Sciences DMP requirement for an example of what a more in-depth software requirements look like.

2. METADATA AND DOCUMENTATION. Throughout the Requirements, metadata standards and documentation should be made more explicit as these are a very important part of making Scientific Data FAIR. Places where more explicit mention is warranted include: Section IV (iii) to make reference to the term “metadata” in order to link with the mention of metadata in the definition of Scientific Data. Section IV (4.2) should include a reference to metadata as well.

3. ALTERNATIVE PLANS. Section IV (4.4). It should be made clearer under what circumstances it is acceptable that the original plan was not achieved, given the Plan Review, Evaluation, and Compliance requirements. Examples would be useful here.

4. ARCHIVING. Section IV (7) encourages the use of free-to-use repositories. We take the position that this could greatly harm the sustainability of these repositories in the long term. The impact may be especially felt by smaller, disciplinary repositories. Given that the requirements include language that allows for budgeting for plan compliance, encouraging depositing in repositories with good sustainability plans (which may or may not include payment for deposit) would be a more prudent suggestion. Another issue to address related to archiving is to provide guidance of what could be done when a repository (free or not) is no longer available. If this is one situation that falls under Section IV (4.4), it would be beneficial to mention the need to identify alternative data repositories as part of fulfilling the Plan requirements

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards.

No Comment

Submission #144

Date: 12/10/2018

Name: Luba Smolensky

Name of Organization: The Michael J. Fox Foundation

Type of Organization: Nonprofit Research Organization

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

Parkinson's disease (clinical, genomics, neuroscience, wearables, imaging, pre-clinical)

I. The definition of Scientific Data

The definition for Scientific Data is reasonable and the inclusion of metadata is very helpful.

II. The requirements for Data Management and Sharing Plans

The requirements for the Plan are broad and flexible. Requesting for researchers to detail how they intend to address perceived barriers, in effort to ultimately share and preserve scientific data, could enhance these requirements. Similarly, providing more specificity around “adequate data security” and committing to concrete timelines (i.e., within 18 months of publication) would help researchers manage expectations, funding, and resources. NIH can further define “broadest sharing” including examples of appropriate open access versus restricted access data sets. Succinct 1-page summaries of available resources (i.e., NIH common data elements, data curation, library archival frameworks) would be helpful reference tools for researchers to quickly onboard to data sharing best practices.

Enabling researchers to include “reasonable costs” for data sharing within funding requests is laudable and would have a positive impact in actualizing scientific data sharing. Furthermore, encouraging thought exercises of alternate data management plans is a very responsible request as researchers would be more prepared to manage changing technology environments.

As an extension of this proposed provisions, it would be helpful to understand how these improvements will be consistently implemented across NIH institutes and how these provisions align with journal mandates for data sharing. Similarly, Data Sharing Plan templates would be a great community resources as funders can integrate them into their own grant awards requirements.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

The optimal timing to implement these provisions would be as soon as possible regardless of phased adoption.

Submission #145**Date:** 12/10/2018**Name:** Anurupa Dev**Name of Organization:** Association of American Medical Colleges**Type of Organization:** Professional Org/Association**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

medical research

I. The definition of Scientific Data

The Association of American Medical Colleges (AAMC) is pleased to provide input here on the specific questions posed by the NIH, as well as overarching comments in the attached letter.

The AAMC generally agrees with the proposed definition that scientific data should not include preliminary analysis, lab notebooks, and other early outputs of the research process, and should primarily consist of “individual level and summary or aggregate data, as well as metadata.” NIH might consider including not only metadata in this definition but also code (e.g., SQL, R, Python) used to interact with the data.

We note that without the use of standards for data and metadata, data sharing will not be useful for secondary analysis or results replication and is unlikely to achieve the stated goal of increased scientific progress. Thus, AAMC urges the NIH to provide as much guidance as possible on the necessary elements for both data and metadata, and wherever possible, provide examples of accepted standards to increase the usability and interoperability of the data reported to NIH. For certain common data types, it would be helpful if NIH specified the metadata necessary for further analytic processing. We also recommend that NIH continue to invest in projects such as the NCATS Biomedical Data Translator and PhenX Toolkit, which harmonize metadata and promote cross-study comparisons and analyses.

The AAMC supports the usage of common data elements and discipline-specific schema for metadata; however, researchers have expressed that resources such as the NIH Common Data Element (CDE) portal can be difficult to navigate without significant knowledge of data science. Additionally, investigators who use a core for data analysis (sometimes at a different institution) may not have adequate expertise to identify the necessary metadata if it is outside of their primary research domain. We recommend that NIH explore ways to make it easier for researchers to identify the appropriate metadata for their data type, update the CDE portal so

it is more user-friendly and continue to fund and incorporate lessons from community-based tools such as the NCBO BioPortal.

II. The requirements for Data Management and Sharing Plans

The development of a Data Management and Sharing Plan (“DMSP”) is a critical element in integrating data-specific considerations into the research lifecycle. The AAMC supports a requirement to include a DMSP as a part of applications and proposals for NIH-funded research, both for its role in promoting future data sharing as well as its utility as a planning and compliance document. We have also commented in this section about current challenges to carrying out proposed elements of the DMSP which the NIH would have to address prior to putting a data sharing policy in place.

We agree with the current proposal that the DMSP be evaluated as an Additional Review Consideration but not factored into the overall impact score for extramural grants. Particularly in the early stages of policy implementation, the focus should be on education and guidance for investigators on how to correctly formulate a DMSP. Once a plan is agreed upon and approved by NIH staff and grantees, compliance with the DMSP could be integrated into award terms and conditions. Formal inclusion of the DMSP in the annual Research Performance Progress Report would likely facilitate adherence to the plan within the proposed timeframe, as well as an opportunity to address any challenges or barriers that have emerged.

The AAMC encourages the NIH to harmonize its DMSP requirements with other federal science agencies whenever appropriate. We also recommend that NIH provide a template or recommendations as well as sample DMSPs which would clearly state the necessary components of a plan as well as the appropriate level of detail. If NIH has specific expectations for investigators with regard to data management, it is important that the DMSP be comprehensive enough to define these requirements at the start of the research funding period.

With regard to data preservation and storage, the timeline for how long data should be maintained will depend on several factors, including the potential long-term utility of the data. Most institutions currently use grant funding in real-time for data collection and curation. It is unclear who bears the responsibility for funding the continued curation and storage of data after a grant has ended. One solution might be to require the researcher to save data for a fixed period of time, determined by the initial grant so that it varies appropriately with the science. For fields where the mandated time is very long (e.g. >5 years), NIH should consider a federally funded repository, possibly run by the National Library of Medicine.

Additionally, as it can be very expensive for institutions to retain data locally, cloud-based platforms from NIH would create long-term capability for data storage and help alleviate some of this financial burden. We appreciate and encourage NIH’s ongoing effort to expand resources in this space, including the development of the NIH Data Commons as well as the STRIDES Initiative to make use of commercial cloud computing.

In relation to issues of data discoverability and access, the proposed provisions suggest that the data should follow the FAIR principles, and we also commend to NIH for consideration the “FAIR-TLC” principles proposed by the Monarch Initiative, TransMed NCATS Data Translator projects, and International Society for Biocuration, which posit that in addition to being FAIR, data should also be traceable, licensed, and connected. From AAMC’s work on data sharing, we have found that if datasets are not assigned a persistent identifier when shared, it is impossible to fully track data usage, and subsequently, appropriately credit researchers for their contributions.

We also recommend that NIH provide a list of the desired characteristics of a repository, using standards for discoverability and archiving developed by the research data community and groups such as the Research Data Alliance and FORCE11, as well as specific recommendations for common data types (as in the case of DbGaP for genomic data). In terms of a timeline for when the data needs to be made accessible after the research concludes, we believe it will create significant confusion and lack of consistency if this element is proposed anew by the investigator for each grant and recommend instead that NIH suggest in the policy a specific length of time within which the data should be made accessible, with the option for the researcher to request a longer timeframe and provide justification.

Under the section on data sharing agreements and licensing, the proposal states that “NIH encourages terms that provide for the broadest use of data resulting from NIH-funded or -supported research, consistent with privacy, security, informed consent, and proprietary issues.” Institutions have shared with AAMC that the lack of standardization in licensing agreements greatly complicates and hinders inter-institutional sharing of NIH-funded data. This process could be facilitated if NIH were to provide a template licensing agreement that lays out sharing terms for data as it has previously done for biological materials.

To better recognize the unique aspects of research with human subjects and to distinguish research with data derived from biospecimens, we urge NIH to revise the proposed provisions’ approach to requirements for data sharing, access, and privacy as they relate to these types of research. We appreciate NIH’s recognition that data sharing aspects of a proposal (including standards for data and metadata collection) need to be developed by an applicant in parallel with informed consent considerations but urge that the data sharing policy itself not appear to dictate informed consent expectations or requirements. We recommend that the draft policy instead remind institutions and investigators that where informed consent is required, the processes of developing the DMSP and the informed consent language must be coordinated instead of suggesting that the data sharing considerations drive the promises made to subjects in the informed consent document. We further recommend that NIH publish guidance to ensure that Program Officers and awardees alike are attuned to the set of issues that will dictate these elements of the DMSP.

The proposed provisions include the statement: “Data may be shared across institutions and repositories to maximize utility, and informed consent should permit broad sharing wherever possible.” While an understandable ideal, the language in the informed consent may have been developed and administered long before the application (in the case of existing data), may not be required at all (in the case of unidentified biospecimens), or may be substantially modified by the IRB after the grant application to account for local context, vulnerable populations, or other considerations. For research that involves human subjects, it will be critical for awardees to connect the informed consent development and institutional review board (IRB) approval process with the negotiated elements of the DMSP to ensure that objectives of the DMSP are realistic, given any constraints of state or federal regulations, institutional policies, or the nature of the data being collected. Even when the research will not involve direct interaction with human subjects or with biospecimens, the policy should explicitly acknowledge the limitations that may be placed on data sharing when using data derived from participants or biosamples from certain populations, such as Native American tribes.

On the issue of identifiability, it is important to recognize that de-identification of data from human subjects is neither simple nor inexpensive and requires additional resources, as further discussed below.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

Compliance with a new data management and sharing policy may require significant changes to the research process and additional time, infrastructure, and resources both on the part of the institution and the individual investigator. AAMC encourages NIH to include a proposed implementation timeline in the forthcoming draft policy to provide stakeholders with an opportunity to comment and suggests that, at a minimum, the policy should be applicable to applications submitted one year from the publication of the final policy. For further insight into the needs of the extramural community in this space, we also refer NIH to the “Accelerating Public Access to Research Data” initiative from the Association of American Universities (AAU) and Association of Public and Land-grant Universities (APLU), which is working with institutions on strategies to facilitate data sharing and management.

The proposal states that “reasonable costs” associated with data management and sharing could be requested under an award budget, and we believe this is essential for investigators to be able to comply with a data sharing policy. As this number may vary widely depending on the data type and volume, we would like to request additional estimates from NIH on allowable expenses. It would also be helpful if the cost of data sharing was included as a standard line item in grant budgets, as it is now for certain types of research. We realize that NIH has funding mechanisms outside of the Research Project Grant, such as training or infrastructure grants for libraries and core facilities, that could supplement funding for data sharing efforts. While this is

helpful for institutions, we would encourage the NIH to disseminate any lessons learned from these funding mechanisms to the broader extramural community, so that the benefit of this investment is distributed beyond the individual institution.

We are in an era of data-centric approaches to understanding biomedical problems, and the AAMC shares the NIH's goal of increased scientific data sharing. However, we would emphasize that a policy alone will not be sufficient to reach this objective. The agency must ensure that it is providing adequate training, education, and guidance, increasing available financial resources, and leading the development of tools and infrastructure in order to enable and facilitate policy implementation. Currently, there are not enough data scientists or informaticists to curate data in the way it will be required by the NIH. While this issue can be partially addressed in the long-term through efforts in training and curriculum changes, a different solution is needed for the implementation timeframe of this policy. Scientists need to be able to identify high-value data and appropriately annotate that data before it is shared. NIH's corresponding investment in research and development will create the toolbox that makes this possible.

Attachment:

December 10, 2018

Office of Science Policy
National Institutes of Health
6705 Rockledge Drive, Suite 750
Bethesda, MD 20892

Re: NOT-OD-19-014 “Request for Information (RFI) on Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research”

The Association of American Medical Colleges (AAMC) appreciates the opportunity to comment on NIH’s request for information regarding proposed provisions for a draft data management and sharing policy. The AAMC is a not-for-profit association representing all 152 accredited U.S. medical schools, nearly 400 major teaching hospitals and health systems, and more than 80 academic and scientific societies. Through these institutions and organizations, the AAMC represents nearly 173,000 faculty members, 89,000 medical students, 129,000 resident physicians, and more than 60,000 graduate students and postdoctoral researchers in the biomedical sciences. These comments on NIH’s proposed provisions consider feedback provided by AAMC-member institutions on their data sharing practices as well as broader standards in the scientific community.

The AAMC supports efforts to increase sharing and re-use of scientific data generated through NIH-funded research and for the agency to develop a clearly defined policy on data sharing. While there has been ambiguity in this space that has impeded investigators from meeting sharing aspirations, there is also recognition that the benefits and utility of data sharing vary significantly across datasets. Any policy developed should reflect this understanding.

In addition to responding to the specific questions for which NIH has requested information, AAMC provides the following high-level recommendations as NIH develops its data sharing policy and the processes it will use to implement and enforce that policy:

- Given the constantly evolving nature of scientific data, we encourage NIH to ensure that the new policy be evaluated regularly and be sufficiently flexible to keep pace with rapidly changing technologies. **Plans to evaluate the impact of the policy should be described and implemented prior to its effective date** to align agency and community expectations about the metrics that will be evaluated. At a minimum, these metrics should include the percentage

of the awarded grants' budgets that are designated for data management and sharing activities and a mechanism to determine whether the shared data have been accessed or re-used.

- **The data sharing policy should be applied consistently regardless of which institute is providing the funding.** The proposed provisions allow for individual Institutes or Centers (ICs) to set differing requirements for data management and sharing, but we urge NIH to harmonize requirements and develop an agency-wide policy that takes into account data type and scientific discipline. Given that institutions and investigators receive funding from multiple ICs, a more consistent policy would facilitate and simplify the development of data management and sharing processes across grants and projects. Where IC-specific variations are necessary, this guidance should be clear and readily available to researchers.
- We recommend that NIH give **explicit guidance on how the new policy would “establish expectations for other NIH policies,”** particularly the NIH Policy on the Dissemination of NIH-funded Clinical Trial information and the NIH Genomic Data Sharing Policy.
- The proposed provisions suggest that NIH is intending to create a draft policy that lists only broad principles and requirements, and then asks investigators to develop and propose an application-specific data management and sharing plan. While the flexibility is appreciated, **if NIH does have unstated expectations for what types of data must be shared or how accessible that data should be, those expectations should be included in the policy.**

In furtherance of the fourth recommendation, we suggest that NIH provide more specific guidance about the agency's expectations about which data investigators should share and for how long that data should be available. Perpetual storage for all data is not feasible or helpful; it would be hugely expensive to hold all data to the same standard and also make the most useful data harder to find. If the policy is too broad, it will not achieve the desired goals of producing meaningful shared data, but an overly prescriptive policy might not be able to keep up with the state of the science. It is hard to know in advance what data will be valuable, although useful benchmarks for a middle ground include: any data that underlies a publication or the minimum dataset that is required to reproduce the analyses that leads to the conclusions of a research project, as well as data which are readily deposited into well described, curated, funded repositories.

Effective data sharing will also be aided by a greater understanding of the use of shared data, and we are pleased that the NIH is involved in the AAMC's "Credit for Data Sharing"¹ initiative, a collaborative project with the New England Journal of Medicine and the MRCT Center, which is working to promote a validated, systematic pathway to link datasets to publications, allow academic researchers to obtain credit for shared datasets, and incentivize and promote data sharing in accordance with FAIR (Findable, Accessible, Interoperable, Reusable) data principles.

¹ www.aamc.org/datasharing; <http://www.nejm.org/doi/full/10.1056/NEJMs1616595>

In response to the specific questions posed by NIH:

I. The definition of Scientific Data:

The AAMC generally agrees with the proposed definition that scientific data should not include preliminary analysis, lab notebooks, and other early outputs of the research process, and should primarily consist of “individual level and summary or aggregate data, as well as metadata.” NIH might consider including not only metadata in this definition but also code (e.g., SQL, R, Python) used to interact with the data.

We note that without the use of standards for data and metadata, data sharing will not be useful for secondary analysis or results replication and is unlikely to achieve the stated goal of increased scientific progress. Thus, AAMC urges the NIH to provide as much guidance as possible on the necessary elements for both data and metadata, and wherever possible, provide examples of accepted standards to increase the usability and interoperability of the data reported to NIH. For certain common data types, it would be helpful if NIH specified the metadata necessary for further analytic processing. We also recommend that NIH continue to invest in projects such as the NCATS Biomedical Data Translator² and PhenX Toolkit³, which harmonize metadata and promote cross-study comparisons and analyses.

The AAMC supports the usage of common data elements and discipline-specific schema for metadata; however, researchers have expressed that resources such as the NIH Common Data Element (CDE) portal can be difficult to navigate without significant knowledge of data science. Additionally, investigators who use a core for data analysis (sometimes at a different institution) may not have adequate expertise to identify the necessary metadata if it is outside of their primary research domain. We recommend that NIH explore ways to make it easier for researchers to identify the appropriate metadata for their data type, update the CDE portal so it is more user-friendly and continue to fund and incorporate lessons from community-based tools such as the NCBO BioPortal⁴.

II. The requirements for Data Management and Sharing plans

The development of a Data Management and Sharing Plan (“DMSP”) is a critical element in integrating data-specific considerations into the research lifecycle. The AAMC supports a requirement to include a DMSP as a part of applications and proposals for NIH-funded research, both for its role in promoting future data sharing as well as its utility as a planning and compliance document. We have also commented in this section about current challenges to carrying out proposed

² <https://ncats.nih.gov/translator>

³ <https://www.phenxtoolkit.org/>

⁴ <https://bioportal.bioontology.org/>

elements of the DMSP which the NIH would have to address prior to putting a data sharing policy in place.

We agree with the current proposal that the DMSP be evaluated as an Additional Review Consideration but not factored into the overall impact score for extramural grants. Particularly in the early stages of policy implementation, the focus should be on education and guidance for investigators on how to correctly formulate a DMSP. Once a plan is agreed upon and approved by NIH staff and grantees, compliance with the DMSP could be integrated into award terms and conditions. Formal inclusion of the DMSP in the annual Research Performance Progress Report would likely facilitate adherence to the plan within the proposed timeframe, as well as an opportunity to address any challenges or barriers that have emerged.

The AAMC encourages the NIH to harmonize its DMSP requirements with other federal science agencies whenever appropriate. We also recommend that NIH provide a template or recommendations as well as sample DMSPs which would clearly state the necessary components of a plan as well as the appropriate level of detail. If NIH has specific expectations for investigators with regard to data management, it is important that the DMSP be comprehensive enough to define these requirements at the start of the research funding period.

With regard to data preservation and storage, the timeline for how long data should be maintained will depend on several factors, including the potential long-term utility of the data. Most institutions currently use grant funding in real-time for data collection and curation. It is unclear who bears the responsibility for funding the continued curation and storage of data after a grant has ended. One solution might be to require the researcher to save data for a fixed period of time, determined by the initial grant so that it varies appropriately with the science. For fields where the mandated time is very long (e.g. >5 years), NIH should consider a federally funded repository, possibly run by the National Library of Medicine.

Additionally, as it can be very expensive for institutions to retain data locally, cloud-based platforms from NIH would create long-term capability for data storage and help alleviate some of this financial burden. We appreciate and encourage NIH's ongoing effort to expand resources in this space, including the development of the NIH Data Commons as well as the STRIDES Initiative to make use of commercial cloud computing.

In relation to issues of data discoverability and access, the proposed provisions suggest that the data should follow the FAIR principles⁵, and we also commend to NIH for consideration the "FAIR-TLC" principles⁶ proposed by the Monarch Initiative, TransMed NCATS Data Translator projects, and International Society for Biocuration, which posit that in addition to being FAIR, data should

⁵ <https://www.nature.com/articles/sdata201618>

⁶ <https://zenodo.org/record/203295#.XAgMEuInaUk>

also be traceable, licensed, and connected. From AAMC's work on data sharing, we have found that if datasets are not assigned a persistent identifier when shared, it is impossible to fully track data usage, and subsequently, appropriately credit researchers for their contributions.

We also recommend that NIH provide a list of the desired characteristics of a repository, using standards for discoverability and archiving developed by the research data community and groups such as the Research Data Alliance⁷ and FORCE11⁸, as well as specific recommendations for common data types (as in the case of DbGaP for genomic data). In terms of a timeline for when the data needs to be made accessible after the research concludes, we believe it will create significant confusion and lack of consistency if this element is proposed anew by the investigator for each grant and recommend instead that NIH suggest in the policy a specific length of time within which the data should be made accessible, with the option for the researcher to request a longer timeframe and provide justification.

Under the section on data sharing agreements and licensing, the proposal states that "NIH encourages terms that provide for the broadest use of data resulting from NIH-funded or -supported research, consistent with privacy, security, informed consent, and proprietary issues." Institutions have shared with AAMC that the lack of standardization in licensing agreements greatly complicates and hinders inter-institutional sharing of NIH-funded data. This process could be facilitated if NIH were to provide a template licensing agreement that lays out sharing terms for data as it has previously done for biological materials.

To better recognize the unique aspects of research with human subjects and to distinguish research with data derived from biospecimens, we urge NIH to revise the proposed provisions' approach to requirements for data sharing, access, and privacy as they relate to these types of research. We appreciate NIH's recognition that data sharing aspects of a proposal (including standards for data and metadata collection) need to be developed by an applicant in parallel with informed consent considerations but urge that the data sharing policy itself not appear to dictate informed consent expectations or requirements. We recommend that the draft policy instead remind institutions and investigators that where informed consent is required, the processes of developing the DMSP and the informed consent language must be coordinated instead of suggesting that the data sharing considerations drive the promises made to subjects in the informed consent document. We further recommend that NIH publish guidance to ensure that Program Officers and awardees alike are attuned to the set of issues that will dictate these elements of the DMSP.

The proposed provisions include the statement: "Data may be shared across institutions and repositories to maximize utility, and informed consent should permit broad sharing wherever possible." While an understandable ideal, the language in the informed consent may have been

⁷ <https://www.rd-alliance.org/>

⁸ <https://www.force11.org/>

developed and administered long before the application (in the case of existing data), may not be required at all (in the case of unidentified biospecimens), or may be substantially modified by the IRB after the grant application to account for local context, vulnerable populations, or other considerations. For research that involves human subjects, it will be critical for awardees to connect the informed consent development and institutional review board (IRB) approval process with the negotiated elements of the DMSP to ensure that objectives of the DMSP are realistic, given any constraints of state or federal regulations, institutional policies, or the nature of the data being collected. Even when the research will not involve direct interaction with human subjects or with biospecimens, the policy should explicitly acknowledge the limitations that may be placed on data sharing when using data derived from participants or biosamples from certain populations, such as Native American tribes.

On the issue of identifiability, it is important to recognize that de-identification of data from human subjects is neither simple nor inexpensive and requires additional resources, as further discussed below. Institutions have also expressed the need for allowances with respect to data sources that are not easily de-identifiable (for example, certain images or results from clinical equipment that do not support de-identification) and would like the NIH to provide additional guidance on how to assure appropriate use of research data made publicly available. While the policy frequently references “researchers and the broader public,” AAMC emphasizes that these are two very different end-users, particularly in the case of clinical data, which are more likely to be shared under a controlled-access model.

III. The optimal timing, including phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy.

Compliance with a new data management and sharing policy may require significant changes to the research process and additional time, infrastructure, and resources both on the part of the institution and the individual investigator. AAMC encourages NIH to include a proposed implementation timeline in the forthcoming draft policy to provide stakeholders with an opportunity to comment and suggests that, at a minimum, the policy should be applicable to applications submitted one year from the publication of the final policy. For further insight into the needs of the extramural community in this space, we also refer NIH to the “Accelerating Public Access to Research Data” initiative⁹ from the Association of American Universities (AAU) and Association of Public and Land-grant Universities (APLU), which is working with institutions on strategies to facilitate data sharing and management.

The proposal states that “reasonable costs” associated with data management and sharing could be requested under an award budget, and we believe this is essential for investigators to be able to comply with a data sharing policy. As this number may vary widely depending on the data type and

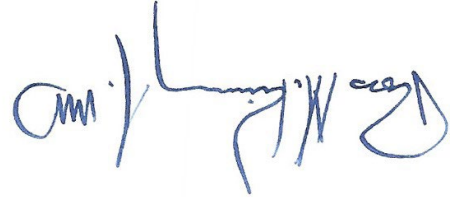
⁹ <https://www.aau.edu/key-issues/aau-aplu-public-access-working-group-report-and-recommendations>

volume, we would like to request additional estimates from NIH on allowable expenses. It would also be helpful if the cost of data sharing was included as a standard line item in grant budgets, as it is now for certain types of research. We realize that NIH has funding mechanisms outside of the Research Project Grant, such as training or infrastructure grants for libraries and core facilities, that could supplement funding for data sharing efforts. While this is helpful for institutions, we would encourage the NIH to disseminate any lessons learned from these funding mechanisms to the broader extramural community, so that the benefit of this investment is distributed beyond the individual institution.

We are in an era of data-centric approaches to understanding biomedical problems, and the AAMC shares the NIH's goal of increased scientific data sharing. However, we would emphasize that a policy alone will not be sufficient to reach this objective. The agency must ensure that it is providing adequate training, education, and guidance, increasing available financial resources, and leading the development of tools and infrastructure in order to enable and facilitate policy implementation. Currently, there are not enough data scientists or informaticists to curate data in the way it will be required by the NIH. While this issue can be partially addressed in the long-term through efforts in training and curriculum changes, a different solution is needed for the implementation timeframe of this policy. Scientists need to be able to identify high-value data and appropriately annotate that data before it is shared. NIH's corresponding investment in research and development will create the toolbox that makes this possible.

The AAMC is very appreciative that NIH is engaging stakeholders at this early stage of the policy process and looks forward to continued engagement on this issue as the data sharing and management draft policy and other guidance are developed. Please feel free to contact me or my colleagues Anurpa Dev, PhD, Lead Specialist for Science Policy (adev@aamc.org) and Heather Pierce, JD, MPH, Senior Director for Science Policy and Regulatory Counsel (hpierce@aamc.org) with any questions about these comments.

Sincerely,



Ross E. McKinney, Jr., MD
Chief Scientific Officer

Submission #146**Date:** 12/10/2018**Name:** Peter Schiffer**Name of Organization:** Yale University**Type of Organization:** University**Role:** Institutional Official**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Yale University has important work in a broad range of scientific research areas.

II. The requirements for Data Management and Sharing Plans

We agree with the NIH's plan to establish responsible management and sharing of scientific data to be managed, preserved, and made accessible in a timely manner for appropriate use by the research community and the broader public. This requirement will likely place a heavier burden on researchers and institutions by expecting added and/or more sophisticated tracking of components of the research process. Institutions may need to scale up their capacity to track research and NIH should be prepared to offer guidance and funding for meeting the new standards.

We comment on three essential components (Submission, Evaluation, Enforcement) as well as the timing for implementation of the new policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards.

Section II. The requirements for Data Management and Sharing Plans (DMP)

Plan elements:

NIH should update DMP templates to help guide researchers toward what they need to consider and what is appropriate and acceptable to be compliant. We anticipate, however, that PIs will need help and guidance on filling out such plans. This will require training and outreach to assist researchers, and potentially the development of technological solutions embedded in service provision by the institution. It may require more robust university resources and services to support researchers on meeting requirements.

1. Data type: Controlled vocabularies should be used to describe the types and amounts of scientific data to avoid reporting inconsistencies. Vocabularies should be readily expandable to

accurately describe research as methods and fields of study change over time.

2. Tools, software, code: Data management plans should include both hardware and software aspects that could impact the output of research pipelines. Ideally, plans should include information about the computing environment that will be used during the active research phase (where will the data be stored and computed) ---these are key to reproducibility of data processing and analysis. The failure to consider this would leave out critical information about how the computing environment may affect the other plan elements (e.g., security). It also misses an opportunity to ask researchers to use systems and tools that will facilitate the required data sharing and preservation at the end of the project.

3. Standards: Yale advocates for data standards and, as an example, has led the creation of a public web resource which provides templates for multiple data types for high quality data collection and reporting <http://ec2-52-7-115-95.compute-1.amazonaws.com/ribeiro/templates.jsf>

4. Preservation and access: The DMP requires that PIs indicate where the data will be archived and how they will be made discoverable. NIH should provide guidelines on certified repositories, expectations for local archiving, and preferred metadata schema. The proposed provisions also require the DMP include information about how the data will be made discoverable, underscoring a preference for open science, which Yale supports (see Yale policy <https://your.yale.edu/policies-procedures/policies/6001-research-data-materials-policy#6001.4>).

5. Preservation and access timeline: Including this element in the plan may place additional burden on PIs to provide highly specialized information about anticipated timeframes for scientific data storage and accessibility, and criteria for how decisions affecting scientific data storage and accessibility will be made throughout the course of the study. NIH should consider funding the resources that may be required at the institutional level in order to provide this information.

6. Legal considerations: Including this element in the plan will place additional burden on institutions' legal departments to develop guidelines and advise on appropriate terms for data reuse consistent with privacy, security, informed consent, and proprietary issues.

Compliance and Enforcement:

1. Who enforces?

In the proposal, all compliance is placed with NIH Institute or Center. However, in addition, the NIH should plan collaborative initiatives, partnering with institutions on how to ensure compliance consistent with how the institution would classify the risk of the data.

2. Non-compliance:

We agree that non-compliance with the NIH-approved Plan would be taken into account by the funding Institute or Center for future funding or support decisions.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

Section III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

Propose phased implementation of the policy:

1. Comment Phase: Prior to publication of a final policy, NIH should issue a draft implementation timeline to provide stakeholders the opportunity to comment.

2. Phase 1: Release new data management and sharing plan requirements (early 2020). NIH will communicate to PIs the new requirements and work with institutions to distribute helpful guidance to researchers. NIH should work with institutions to develop processes to effectively capture the information provided in the plans for the purpose of internal tracking, evaluation, and compliance. Also consider collaboration with research publishers to increase findability of comprehensive data packages, e.g., through links in the Supporting Information section.

3. Phase 2: Simultaneously or shortly thereafter, implement data management and sharing plan evaluation by NIH IC (late 2020). NIH will work with institutions and PIs to design a fair, expedited, and thorough evaluation process. NIH proposes a variety of ways to implement the determination of whether a plan is acceptable:
 - Extramural Grants: Plans could be evaluated as an Additional Review Consideration, i.e., evaluated as acceptable or unacceptable by reviewers, but not be factored into the overall impact score through the peer review process. This allows for NIH staff to work with potential awardees to ensure that any reviewer concerns regarding the Plan could be addressed for meritorious applications as a contingency of NIH funding. Plan compliance would be integrated into terms and conditions as appropriate.

 - Contracts: Plans could be included as part of the technical evaluation performed by NIH staff and incorporated in the subsequent terms of the contract.

 - NIH Intramural Research Projects: Plans could be reviewed by the Scientific Director (or designee) or Clinical Director (or designee) of the researcher's funding IC and integrated into approval conditions as appropriate.

- Other funding/support agreements: Plans could be evaluated in the context of other funding/support agreement mechanisms, e.g., CRADA, Other Transactions, and integrated into the terms and conditions as appropriate.

4. Phase 3: Implement enforcement including establishing some “best methods for meeting requirements” (late 2022). The time lag here is intended to allow revising requirements if necessary based on evaluation, and to allow institutions to establish internal processes for data management and sharing plan evaluation and enforcement.

Attachment:

Response

We agree with the NIH's plan to establish responsible management and sharing of scientific data to be managed, preserved, and made accessible in a timely manner for appropriate use by the research community and the broader public. This requirement will likely place a heavier burden on researchers and institutions by expecting added and/or more sophisticated tracking of components of the research process. Institutions may need to scale up their capacity to track research and NIH should be prepared to offer guidance and funding for meeting the new standards.

We comment on three essential components (Submission, Evaluation, Enforcement) as well as the timing for implementation of the new policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards.

Section II. The requirements for Data Management and Sharing Plans (DMP)

Plan elements:

NIH should update DMP templates to help guide researchers toward what they need to consider and what is appropriate and acceptable to be compliant. We anticipate, however, that PIs will need help and guidance on filling out such plans. This will require training and outreach to assist researchers, and potentially the development of technological solutions embedded in service provision by the institution. It may require more robust university resources and services to support researchers on meeting requirements.

1. Data type: Controlled vocabularies should be used to describe the types and amounts of scientific data to avoid reporting inconsistencies. Vocabularies should be readily expandable to accurately describe research as methods and fields of study change over time.
2. Tools, software, code: Data management plans should include both hardware and software aspects that could impact the output of research pipelines. Ideally, plans should include information about the computing environment that will be used during the active research phase (where will the data be stored and computed) ---these are key to reproducibility of data processing and analysis. The failure to consider this would leave out critical information about how the computing environment may affect the other plan elements (e.g., security). It also misses an opportunity to ask researchers to use systems and tools that will facilitate the required data sharing and preservation at the end of the project.
3. Standards: Yale advocates for data standards and, as an example, has led the creation of a public web resource which provides templates for multiple data types for high quality data collection and reporting <http://ec2-52-7-115-95.compute-1.amazonaws.com/ribeiro/templates.jsf>
4. Preservation and access: The DMP requires that PIs indicate where the data will be archived and how they will be made discoverable. NIH should provide guidelines on certified repositories, expectations for local archiving, and preferred metadata schema. The proposed provisions also require the DMP include information about how the data will be made discoverable, underscoring a preference for open science, which Yale supports (see Yale policy <https://your.yale.edu/policies-procedures/policies/6001-research-data-materials-policy#6001.4>).

5. Preservation and access timeline: Including this element in the plan may place additional burden on PIs to provide highly specialized information about anticipated timeframes for scientific data storage and accessibility, and criteria for how decisions affecting scientific data storage and accessibility will be made throughout the course of the study. NIH should consider funding the resources that may be required at the institutional level in order to provide this information.
6. Legal considerations: Including this element in the plan will place additional burden on institutions' legal departments to develop guidelines and advise on appropriate terms for data reuse consistent with privacy, security, informed consent, and proprietary issues.

Compliance and Enforcement:

1. Who enforces?

In the proposal, all compliance is placed with NIH Institute or Center. However, in addition, the NIH should plan collaborative initiatives, partnering with institutions on how to ensure compliance consistent with how the institution would classify the risk of the data.

2. Non-compliance:

We agree that non-compliance with the NIH-approved Plan would be taken into account by the funding Institute or Center for future funding or support decisions.

Section III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

Propose phased implementation of the policy:

1. Comment Phase: Prior to publication of a final policy, NIH should issue a draft implementation timeline to provide stakeholders the opportunity to comment.
2. Phase 1: Release new data management and sharing plan requirements (early 2020). NIH will communicate to PIs the new requirements and work with institutions to distribute helpful guidance to researchers. NIH should work with institutions to develop processes to effectively capture the information provided in the plans for the purpose of internal tracking, evaluation, and compliance. Also consider collaboration with research publishers to increase findability of comprehensive data packages, e.g., through links in the Supporting Information section.
3. Phase 2: Simultaneously or shortly thereafter, implement data management and sharing plan evaluation by NIH IC (late 2020). NIH will work with institutions and PIs to design a fair, expedited, and thorough evaluation process. NIH proposes a variety of ways to implement the determination of whether a plan is acceptable:
 - Extramural Grants: Plans could be evaluated as an Additional Review Consideration, i.e., evaluated as acceptable or unacceptable by reviewers, but not be factored into the overall impact score through the peer review process. This allows for NIH staff to work with potential awardees to ensure that any reviewer concerns regarding the Plan could be addressed for meritorious applications as a contingency of NIH funding. Plan compliance would be integrated into terms and conditions as appropriate.

- Contracts: Plans could be included as part of the technical evaluation performed by NIH staff and incorporated in the subsequent terms of the contract.
 - NIH Intramural Research Projects: Plans could be reviewed by the Scientific Director (or designee) or Clinical Director (or designee) of the researcher's funding IC and integrated into approval conditions as appropriate.
 - Other funding/support agreements: Plans could be evaluated in the context of other funding/support agreement mechanisms, e.g., CRADA, Other Transactions, and integrated into the terms and conditions as appropriate.
4. Phase 3: Implement enforcement including establishing some "best methods for meeting requirements" (late 2022). The time lag here is intended to allow revising requirements if necessary based on evaluation, and to allow institutions to establish internal processes for data management and sharing plan evaluation and enforcement.

Submission #147

Date: 12/10/2018

Name: Ara Tahmassian

Name of Organization: Harvard University

Type of Organization: University

Role: Institutional Official

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

Biomedical Research

I. The definition of Scientific Data

See Attached Letter

II. The requirements for Data Management and Sharing Plans

See Attached Letter

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

See Attached Letter

Attachment:



Harvard University
Office of the Vice Provost for Research

Ara Tahmassian
Ullil'mi(* Chief Resemrh Collipliallce QJi,ir
ara_tahmassian@harvard.edu
vpr.harvard.edu

Richard A. and Susan F. Smith Campus Center
1350 tvlassachusetts Avenue, 836
Cambridge, l'vIA 02138
t. 617.495.9797 f. 617.495.8051

Office of Science Policy
National Institutes of Health
6705 Rockledge Drive, Suite 750
Bethesda, MD 20892

RE: NOT-OD-19-014 Request for Information (RFI) on Proposed Provisions for a Draft Data Management and Sharing Policy for National Institutes of Health (NIH) Funded or Supported Research

On behalf of the President and Fellows of Harvard College ("Harvard"), we thank you for the opportunity to comment on the NIH's Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research ("Policy"). In addition to the comments included below, we wish to extend our endorsement to the comments submitted by the Association of American Medical Colleges (AAMC), and the joint comments submitted by the Association of American Universities (AAU), the Association of Public and Land Grant University (APLU) and the Council on Government Relations (COGR).

Overall, Harvard supports the objectives and motivations described in the draft Policy. Our University has always encouraged scholarly exchange and emphasized the importance of generating and disseminating research data for public use. In an effort to further promote these ideals and the Policy's aims in a practical and comprehensive fashion, we respectfully provide the following recommendations:

Definition of Scientific Data:

We agree that the definition of Scientific Data should not apply to physical materials, "laboratory notebooks, preliminary analysis, completed case report forms, drafts of scientific papers, plans for future research, peer reviews, or communications with colleagues." We do however suggest that NIH consider drafting the Policy so that it explicitly requires Metadata, code (e.g. software, scripts), and workflows to be made reasonably available with the Scientific Data it relates to. Without these supporting data elements, it may be impossible for Scientific Data to be assessed, reused or replicated, defeating the overarching purpose of the Policy.

The Policy and definitions should, however, acknowledge that access to certain sensitive or third-party data will need to be controlled. For example, proprietary software or identifiable information provided by a hospital can be made available upon request, pursuant to the terms of a data use or nondisclosure agreement. In these cases, the Plan should include an explanation for why the Scientific Data cannot be made publicly available, and the Policy should recommend that the researcher endeavor to make the

Metadata (at least descriptive Metadata) available in a repository, with the Scientific Data restricted and only accessible if permissions are granted under the applicable terms and conditions.

Harvard agrees that Scientific Data and the supporting information necessary to replicate results must be made available, but also believe the level of accessibility must be a project specific determination. Developing guidance on the scope of the definition, such as whether it includes background data or third-party data, and to what extent this is within the researchers' discretion will greatly enhance universities' ability to comply with the Policy.

Additionally, we believe it may be more accurate to reference "Research Data", as opposed to "Scientific Data". Certain data may not be categorized as scientific (e.g. statistical data, market data) depending on the researcher, discipline or specifics of the project. Using "Research Data" will allow the Policy to apply to the broader research community funded by NIH, as well as be consistent with international regulations such as the EU Commission.

Requirements for Data Management and Sharing Plans ("Plans"):

Evaluation Criteria: Section *N* states "Plans could be evaluated as an Additional Review Consideration, i.e., evaluated as acceptable or unacceptable by reviewers, but not be factored into the overall impact score through the peer review process." We agree that Plans should not factor into the overall impact score. Especially as this is a new Policy, and the data landscape is constantly changing, researchers and universities will require clear guidance as to what elements are required to be included in a Plan, what elements may simply be preferred, and what elements will be detrimental to include or omit. Without specific guidance from the NIH, it will be difficult for researchers to develop consistent and informed Plans, which will ultimately detract from this Policy's objectives.

Data Management and Preservation: In addition to comprehensive guidance on expected content, ideally with input from other federal agencies, but at least in concurrence with the ICs, Harvard requests that the NIH develop a standardized template for Plans for illustrative purposes. The template could be specific to the funding mechanism, discipline and/or type of data, but each template should consistently interpret and apply the Policy. Use of the templates wouldn't need to be mandatory, but templates would provide a valuable reference and standard for both the researchers and institutions drafting the Plans, as well as the peer reviewers and NIH personnel reviewing the proposals. As mentioned above, the more information researchers have on the expectations for their Plans and the criteria by which they will be assessed, the more effective and successful this Policy will be.

As noted above, data management can be quite costly. Certain researchers may include data costs in their proposed budget, but many researchers may be, at least initially, unaware of the costs of deidentifying and properly storing data, or potentially unaware of how to appropriately reflect those costs in the proposal. To further compound the issue, the responsibility for supporting data access after a project has ended has historically fallen to the institution. As the NIH noted in its Webinar, some data ages like a fine wine, while other data may only be useful for a matter of weeks, resulting in an unpredictable strain on institutional resources.

Potential solutions may include a federally funded repository for Scientific Data, accessible to all recipients of federal funding. Similar to the Plan template discussed above, institutions would not be required to utilize the federal repository, however it would provide a feasible option for those researchers and institutions who may lack data resources, or are concerned they may not be in compliance with all federal requirements applicable to data storage and accessibility.

It would also be helpful for the NIH to provide references to existing acceptable repositories, and to establish minimum standards for compliant repositories, generally. For example: Should a repository require persistent identifiers? Is there specific language that must be reflected in user licenses? Are there specific usability requirements that must be met? Must a repository be precisely aligned with the FAIR Principles? Detailed guidance on these issues should be incorporated into the Policy to better allow for researchers and institutions to make informed decisions when deciding on where to store their data.

It is important that the Policy not only reflects the variability in the useful life of data and the prospective costs, as well as any guidance on how best to comply with applicable standards, but also provides a specific process for modifying Plans and budgets so institutions aren't put in the position of terminating a project due to lack of internal funding.

Modifications: Basic and fundamental research is inherently unpredictable, and often requires researchers to react to results and adjust their strategies in real time. Because of this, experiments, analyses and results may deviate from researchers' preliminary expectations. Consequently, it will be important for the NIH to develop guidance establishing the level of specificity expected in Plans, and equally important, a process for modifying Plans throughout the course of the research. With regard to the process for updating Plans, Harvard suggests allowing researchers to include substantive revisions in their Progress Reports. This would allow for the NIH to be notified of adjustments to a Plan, and give feedback, but not require a formal amendment process.

Overarching Policy Recommendations

Flexibility and Adaptability: The research community has embraced data sharing as a staple of innovation and progress across all research disciplines. As researchers and institutions develop strategies and technologies for utilizing and managing data, successful regulations and standards will need to adapt to the changing culture. For this Policy to remain as valuable and relevant as it is today, we request that the NIH draft the Policy and Plan requirements with sufficient flexibility to account for the rapidly developing data landscape. The terms of the Policy should require periodic reassessment and communication with stakeholders to ensure that its application remains consistent with the intended principles and objectives.

Relationship Between Other Rules and Regulations: We strongly encourage the NIH to consider the interplay between this Policy and other NIH policies, such as the Genomic Data Sharing Policy, as well as the potential implications for research projects that are funded by multiple federal sponsors with overlapping requirements. To the extent practicable, it would be advantageous for those interpreting and implementing the Policy if the IC' s, as well as other government agencies supporting fundamental research, adopted compatible standards.

Thank you again for engaging the research community in developing this important and exciting Policy. We very much look forward to continued conversations and collaboration throughout the process of finalizing the Policy and associated requirements and standards, developing relevant resources, trainings and guidance, and the eventual implantation.

Sincerely,

A stylized, blue handwritten signature that appears to read 'Ara Tahmassian'. The signature is composed of several connected strokes, including a large initial 'A', followed by a series of loops and horizontal lines.

Ara Tahmassian, PhD
Harvard University Chief Research Compliance Officer

Submission #148**Date:** 12/10/2018**Name:** Tom Sellers, PhD, Center Director**Name of Organization:** H. Lee Moffitt Cancer Center & Research Institute, Inc.**Type of Organization:** Nonprofit Research Organization**Role:** Institutional Official**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Clinical, genomics, neuroscience, infectious disease, epidemiology, cancer

I. The definition of Scientific Data

The H. Lee Moffitt Cancer Center & Research Institute (“Moffitt”) team members appreciate that the scope proposed by NIH does not include preliminary and other types of data in this policy. However, in the definition of Scientific Data, NIH has not considered or addressed several key concerns. We summarize these and make recommendations below.

- **Sharing requirements for data and Protected Health Information (PHI) across hospitals, healthcare systems and research institutions:** Expectations of sharing scientific data seem to conflict with HIPAA and privacy protections related to PHI. We recommend that NIH issue clearer and more specific guidelines clarifying that “scientific data” specifically exempt patient datasets, as well as data elements and biospecimens that fall outside the scope of published datasets. Clinical datasets, even when de-identified, often have strict sharing and access requirements between researchers and healthcare systems. For example, data generated by clinical trials can include genomic and specimen data that is part of clinical care as well as research. Many healthcare systems have, or are developing, their own biobanks for clinical care with secondary research use. This, further complicates the conflicting need to protect patient privacy with the mandate to share data (and the associated issue of funding source particularly if the funding for such data emanates from both NIH and a third-party source). We are concerned that such strict language may have a negative impact on institutional support, collaboration and/or subject participation in research.
- **Assessing data sharing requirements for work supported by funding from both NIH and non-NIH sources:** It is quite common for NIH grants to leverage research resources funded through non-NIH sources (e.g., industry, institutional funds, other foundations, philanthropy) to supplement or fully support the cost of completing study specific aims or related work. In fact,

NIH often encourages and evaluates projects based on additional institutional support and secondary data analysis or sample usage for many of its funding opportunities. Therefore, while federal funding is the largest sponsor of research, research projects are rarely only supported by NIH. In fact, economic pressure and scientific complexity necessitate multiple funding sources that may have different interests and may span institutions within and outside of the United States. Examples include academia-industry partnerships and large-scale consortia to accelerate the discovery and translation of, as well as data or samples (e.g., for GWAS studies). These partnerships and collaborations are often initially supported through non-NIH funding sources, later leading to NIH support to expand scientific impact; conversely, NIH may initially fund work that is later extended or continued through other funding sources. However, the current document treats all data as funded exclusively from federal sources -- alternate funding sources or collaborations with researchers in other countries with different data sharing policies is not discussed. We recommend that NIH issue more clear and specific guidelines regarding requirements for work supported by multiple funding sources, as well as international collaborations that lead to scientific data. We recommend that data sharing be limited to the federally-supported portion. As above, we are concerned that these requirements may negatively impact science by creating a disincentive for non-NIH sources to provide support.

- Defining scope of required data sharing for work building on past scientific data or serving as the foundation for future scientific data: Similar to the point above, we recommend greater clarity in defining the scope of scientific data considered as the “project.” Specifically, a key question surrounds whether the scope of the scientific data to be shared is based on the data needed to answer the specific aims of the NIH grant or all data collected under a specific study that may involve other institutions, data, or funding sources. This is particularly relevant when collecting data to support a federal grant application with non-NIH funding or for consortia projects that are funded by multiple sources (e.g., genomics studies or industry partnerships). We recommend that data sharing be limited to the federally-supported portion. As above, we are concerned that these requirements may negatively impact science by providing a dis-incentive towards non-NIH sources to provide support.
- Reporting of individual or aggregated data for projects partially supported by NIH: As noted above, in cases where data involves multiple institutions or funding sources, it is unclear if individual participant data must be shared or if aggregated data sharing is acceptable. Given the interests represented by partnering institutions or networks that do not have uniform federal support, we recommend requiring individual-level data for NIH-supported work, and allowing aggregated data for participants. This approach maximizes the utility while fostering collaboration.

II. The requirements for Data Management and Sharing Plans

Regarding the requirements of the plans, we recommend the following:

- Rename “Perceived Barriers” as “Reported Barriers to Sharing:” We recommend changing the term “perceived barriers” to sharing to “reported barriers to sharing” as many research institutions utilize restricted and heavily regulated datasets that cannot be shared widely or in the open manner described by the BD2K initiative. The term “perceived barriers” implies that there are no real (e.g., legal, ethical) barriers. “Reported” barriers to sharing are a direct result of the institution’s responsibility and duty to uphold legal contracts and agreements, such as compliance with IRB decisions regarding the manner and extent of data sharing, requirements for controlled or limited access, and/or requirement to uphold agreements with third party organizations, pharmaceutical partners, and consortia. Many of these barriers exist to meet competing federally mandated requirements, such as HIPAA (described in section 1). International collaboration may also result in barriers to sharing, particularly with the advent of the General Data Protection Regulation (GDPR) recently passed by the European Union. Institutions cannot forgo their legal obligations and as such, collaborative science funded and supported by NIH may, paradoxically, be limited by strict data sharing policies. We agree that investigators should outline efforts to maximize use and sharing of data collected under NIH funds within legal and ethical constraints; however, the requirements as currently proposed? may hinder collaboration.
- Increase Page Limit to 4-5 Pages: The recommended limitation of two pages for the Data Sharing Plan is not practical, given the requirements. Considering likely obligations and restrictions, it would be more prudent of NIH to allow for at least a four- to five-page data sharing plan consisting of documentation outlining how scientific data will be managed and preserved, while sufficiently identifying restrictions and barriers to sharing certain information. Given widely publicized data security measures, a specific section should be added to describe security strategies used by the submitting institution to provide adequate data security. Our concern is that the two-page length negatively impacts transparency and clarity.
- Identify Response Times as 45-60 Days: With the inclusion of the Data Sharing Plan as an “Additional Review Consideration” for Extramural Grants, there is no clear response time to address “reviewer concerns” regarding a proposed data sharing plan. This is a particular concern given the growing complexity of this work (noted above). The data in question may be under heavy restrictions due to external agreements with industry partnerships, pharmaceutical agreements, IRB rulings, data sharing agreements, international laws, etc. Providing 45-60 days we feel is a reasonable time period for more complicated arrangements; however, with funding held contingent on resolving these concerns, institutions have a natural desire to expedite responses.
- Provide an Opportunity for Mid-Award/Renewal Data Type Revision or Data Sharing Revision: We request that NIH provide clarification on the subject of “Data Type” for Extramural Applications that are submitted without specifically knowing the amount and exact type of data to be generated over the course of the application. This is particularly important if the Data Sharing Plan is included in the Notice of Award. The rapid pace of scientific advancement often

leads to real-time changes in research methods and approaches to use the most robust tools and assays. A major benefit to NIH funding mechanisms is the ability to provide flexibility to complete scientific aims using the best possible approaches. We thus recommend that a mechanism be developed to provide a revised Data Sharing Plan with description of Data Type at year 3 of an award, or at each competitive renewal, as long as the project met initial goals of the project or negotiated with the Program Officer, if changes were necessary. This would address situations such changes in patient accrual targets are altered or assay technology.

- Specify a Defined Data Retention Time (Six Years or 10 Years): There are inconsistent not data retention standards. In the past, many institutions relied upon patent law to drive the requirements; however, recent changes to the law have left a void in which numerous bodies have issued non-binding recommendations ranging from two to ten years. Journal and research integrity standards vary.

A key correlate of retention is the data that should be retained (i.e., the dataset[s] driving published work and/or the associated but not published secondary data). Furthermore, if data are stored within a public repository such as dbGaP, ProteomXchange, or Dataverse, there is no clear guidance for the retention of onsite records after submission, which could result in duplicative efforts to maintain data?.

Special consideration should be given towards consortia and community projects in which datasets are distributed amongst private entities within a closed healthcare system, and/or partnership and may be subject to heavy restrictions towards sharing access, and costly upkeep. Likewise, little guidance has been given on large-scale data generating projects, which suffer from a variety of barriers to sharing; the least of which is the fact that it is simply not known what type and form of data will ultimately be generated until the project has concluded.

We recommend setting a period; six years would be our preference, but 10 years would be a second logical period.

- Allow Direct or F&A Support of Costs - As noted in some of the comments above, with uncertainty in the details of the policy, this places greater financial burden on institutions. We specifically request that NIH support storage costs (i.e., taxpayer) in accordance with the final scope (i.e., if limited to NIH support as we recommend above, or more broadly if the scope is not limited). We request that data storage for federal awards be either specifically allowable as a direct cost in projects with data sharing provisions, or as a specific Facilities & Administration element. While these do not address out-year costs, this does contribute to the substantial cost.
- Support of Data Repositories - Many publishers now require that scientific data be stored in federal repositories even if the data was not generated from federal funds. As noted above, this provides a mismatch of requirements and resources. We recommend that either: 1) data sharing requires data from non-federal sources, that federal repositories be open to data

generated from non-NIH sources for consistency, or 2) NIH mandate to publishers that federal repositories not be required for non-federally supported projects.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

- **Revise to Rolling Implementation Timeline:** We recommend that NIH adopt a rolling implementation date for these requirements, similar to how the Public Access Compliance regulation and uptake was administered. All projects created before a certain date would be grandfathered into provisions exempting them from the new plan requirements, and projects founded after the date, would be subject to the requirements.
- **Develop Model Templates:** We recommend that NIH provide guidance, education, and specific guidelines towards developing robust data plan templates and boilerplate language that can serve as a model for a variety of institutions and project types.

Submission #149**Date:** 12/10/2018**Name:** Adam Thomas**Name of Organization:** NIMH IRP**Type of Organization:** Government Agency**Role:** Scientific Researcher**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Neuroscience, Neuroimaging, and Mental Health

I. The definition of Scientific Data

Regarding the definition of scientific data, the proposed provisions succeed in outlining the key principle: "all materials needed to validate and replicate a research finding". This includes the standards and protocols used, the metadata, and the software and analysis code.

II. The requirements for Data Management and Sharing Plans

Regarding the requirements of the data sharing plans, the proposed provisions adequately cover the seven major topics that should be included in this short document: (i) data types, (ii) related tools and software, (iii) data standards, (iv) data preservation, access (including timelines) and discoverability, (v) terms for re-use and redistribution, (vi) limitations on access, and (vii)

oversight of data management.

The compliance and enforcement section would benefit from additional details. It is often the case that scientist seeking to replicate or verify a finding are unable to obtain the necessary data or metadata to do so. (Stodden et al. 2018, PNAS DOI:10.1073/pnas.1708290115). The NIH should provide a mechanism to mediate issues between data producers and scientists seeking to obtain these data for secondary analysis.

The IRP section of the compliance provisions is vague. Compliance with submitted data sharing plans should be specifically evaluated by the Board of Scientific Counsellors at intramural investigators' quadrennial review. A data sharing section should also be added to investigators' annual reports.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

Nearly six years have passed since the White House released its memo instructing the NIH to increasing access to the results of federally funded scientific research. It is critical for NIH to draft and release this new policy without further delay. Hard deadlines need to be set for the creation of needed infrastructure and standards.

Attachment:

Comments on Proposed Provisions for a Draft NIH Data Management and Sharing Policy

I strongly support the outlined provisions for a Draft NIH Data Management and Sharing Policy. Nearly six years have passed since the White House released its memo instructing the NIH to increasing access to the results of federally funded scientific research. It is critical for NIH to draft and release this new policy without further delay. Hard deadlines need to be set for the creation of needed infrastructure and standards.

Regarding the definition of scientific data, the proposed provisions succeed in outlining the key principle: "all materials needed to validate and replicate a research finding". This includes the standards and protocols used, the metadata, and the software and analysis code.

Regarding the requirements of the data sharing plans, the proposed provisions adequately cover the seven major topics that should be included in this short document: (i) data types, (ii) related tools and software, (iii) data standards, (iv) data preservation, access (including timelines) and discoverability, (v) terms for re-use and redistribution, (vi) limitations on access, and (vii) oversight of data management.

The compliance and enforcement section would benefit from additional details. It is often the case that scientist seeking to replicate or verify a finding are unable to obtain the necessary data or metadata to do so. (Stodden et al. 2018, PNAS DOI:10.1073/pnas.1708290115). The NIH should provide a mechanism to mediate issues between data producers and scientists seeking to obtain these data for secondary analysis.

The IRP section of the compliance provisions is vague. Compliance with submitted data sharing plans should be specifically evaluated by the Board of Scientific Counsellors at intramural investigators' quadrennial review. A data sharing section should also be added to investigators' annual reports.

Adam Thomas adamt@nih.gov
Data Science and Sharing Team
NIMH Intramural Research Program
10 Center Dr., Room 1D80
Bethesda MD. 20892-1148
Phone: 301-402-6351

Submission #150**Date:** 12/10/2018**Name:** Santiago Schnell**Name of Organization:** STRENDA Commission**Type of Organization:** Other**Other Type of Organization:** International Consortium of Standards for Reporting Enzymology Data**Role:** Scientific Researcher**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Protein function and enzymology

I. The definition of Scientific Data

We suggest to define formally scientific data as the results of observations performed under specified conditions. This definition is in accordance with the NIH principle that scientific data should consist of “individual level and summary or aggregate data, as well as metadata.” We also recommend to incorporate as part of the scientific data information about the mathematical, statistical and computational approaches used to analyze and interpret the data (statistical protocols, algorithms, software, etc).

For the specific case of enzyme-function data, the STRENDA Commission presently defines data at two levels:

- List Level 1A. Data required for a complete description of an experiment. For more details: DOI: [10.3762/strenda.17](https://doi.org/10.3762/strenda.17)
- List Level 1B. Description of enzyme-activity data. For more details: DOI: [10.3762/strenda.27](https://doi.org/10.3762/strenda.27)

II. The requirements for Data Management and Sharing Plans

The STRENDA Commission supports a requirement to include Data Management and Sharing Plans as a part of proposals for NIH-funded research, both for its role in promoting future data sharing as well as its utility as a planning and compliance document.

We recommend that the NIH request their extra- and intramurally funded scientists adopt the STRENDA Guidelines for reporting kinetics and equilibrium data. In the STRENDA Guidelines, "All reports of kinetic and binding data must include a description of the identity of the catalytic or binding entity (enzyme, protein, nucleic acid or other molecule). This information should include the origin or source of the molecule, its purity, composition, and other characteristics such as post-translational modifications, mutations, and any modifications made to facilitate expression or purification. The assay methods and exact experimental conditions of the assay must be fully described if it is a new assay or provided as a reference to previously published work, with or without modifications.

The temperature, pH and pressure (if other than atmospheric) of the assay must always be included, even if previously published. In instances where catalytic activity or binding cannot be detected, an estimate of the limit of detection based on the sensitivity and error analysis of the assay should be provided. Ambiguous terms such as "not detectable" should be avoided. A description of the software used for data analysis should be included along with calculated errors for all parameters.

First-order and second-order rate constants should be reported in units of s^{-1} and $M^{-1} \cdot s^{-1}$, respectively. Equilibrium binding constants should normally be reported as dissociation constants with units of concentration (M, mM, μ M, nM). The values k_{cat} , k_{cat}/K_m and K_m from steady-state enzyme kinetics should be reported in units of s^{-1} , $M^{-1} \cdot s^{-1}$ and concentration (mM, μ M, nM), respectively. The steady-state specific activity of an enzyme should normally be reported as a k_{cat} . If there is considerable uncertainty in the molar concentration of the catalyst, the specific activity should be reported as a V_{max} (nmol, μ mol) of product formed per amount of protein per unit time (e.g. μ mol \cdot mg $^{-1} \cdot$ s $^{-1}$).

With regard to data preservation and storage, the STRENDA Commission recommends that scientists deposit their enzyme assay conditions and protein-functional data in the STRENDA Database (STRENDA DB, <https://www.beilstein-strenda-db.org/strenda/index.xhtml>). The STRENDA Commission is currently working with other repositories (SABI-RK, Brenda, BioCatNet) to maximize utility of enzyme data across different platforms.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

We urge the NIH to consider that a new data management and sharing policy will require changes to the research process and additional time for preparing and writing grants. Data management and sharing policies maybe also require changes in infrastructure, and resources available to individual investigators through their laboratories or institutions.

NIH should provide resources to cover any reasonable cost to implement data management and sharing policies, which could include training, education, guidance, financial resources for and the development of the tools needed to enable and facilitate policy implementation. For example, there are not enough data scientists in the protein-function field to curate the data available in the literature, or to manage the data being generated by scientists.

We encourage the NIH to ensure that the Data Management and Sharing Plans policy be evaluated regularly and be sufficiently flexible to keep pace with rapidly changing technologies as science is continuously evolving.

The STRENDA Commission recommends the immediate adoption of the STRENDA Guidelines for reporting enzyme functional experiments and data as part of the Data Management and Sharing Plan. We also recommend the deposition of enzyme assays protocols and results in STRENDA DB.

Attachment:

Department of Molecular & Integrative Physiology
1137 E. Catherine St, Med. Sci. II 7744
Ann Arbor, MI, 48109-5622, USA
medicine.umich.edu/physiology

Interim Department Chair
John A. Jacquez Professor of Physiology
Professor of Molecular & Integrative Physiology
and Computational Medicine & Bioinformatics
William K. Brehm Investigator

Tel: +1 734 615 8733
E-mail: schnells@umich.edu

10th December 2018

Office of Science Policy
National Institutes of Health
6705 Rockledge Drive, Suite 750
Bethesda, MD 20892

To Whom It May Concern

In response to: NOT-OD-19-014 “Request for Information (RFI) on Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research”

As representatives of the STRENDA (*Standards for the Reporting of Enzyme Data*)¹ Commission, we appreciate the opportunity to respond to the NIH’s request for information regarding proposed provisions for a draft data management and sharing policy. The STRENDA Commission was constituted in 2004 under the auspices of the Beilstein-Institute and is comprised of an international panel of scientists who bring in diverse expertise in biochemistry, enzyme nomenclature, analytical biochemistry, measurement science, mechanistic enzymology and systems biology.² Our comments on NIH’s proposed provisions consider the work of the commission over the last 14 years on the current data sharing practices as well as existing standards in the enzyme function data community.

The volume of protein-function information annually entering the literature far outstrips the capacity of the human mind to house it. Collectively, this ‘big dataset’ represents an enormous, continually refreshed set of interconnected information that cannot be mined effectively with current tools; consequently, scientists expend significant effort compiling small subsets of these data and working to decipher trends within them. What if a comprehensive repository that contained both protein-function data and the conditions under which they were acquired were readily available to the public? The ability to instantly summon from the Web all relevant initial-rate, binding, allostery, stability, pH and ionic strength dependencies, substrate specificities, equilibrium data... would be a tremendous advantage to experimentalists working with a particular enzyme, and perhaps even more so to those attempting to integrate across these

¹ <https://www.beilstein-institut.de/en/projects/strenda>

² <https://www.beilstein-institut.de/en/projects/strenda/commission>

findings to describe complex biological processes. Extant websites that incorporate protein-function data, such as UniProt and the PDB, could augment their protein-function platforms by interfacing with the repository. The simple act of establishing uniform data-submission practices for published protein structures and making the structures available to all at the “push-of-a-button” spawned a transformation of the biological sciences that included genesis of the *Protein Structure Initiative* and gave traction to burgeoning scientific cottage industries including molecular dynamics and protein-structure predication and visualization tools. The relationship of structure-to-function is among the most fundamental in science; yet, access to large-scale, well-organized protein-function data lags far behind that of structure.

Over the past 14 years, the STRENDA Commission has worked with scientists and publishers in the protein-function community to develop a data deposition model that can electronically captures protein-function data as it enters the literature, and ensures that the conditions under which the data were acquired are fully described.³ While more than 55 journals currently recommend the guidelines to their authors, funding agencies have not yet recommended the guidelines in their data management and sharing policies. Transparent reporting of experimental methods to ensure the reproducibility of results is particularly crucial in enzyme kinetics, given the sensitivity of the measured parameters to small changes in experimental conditions.⁴

STRENDA supports efforts to increase sharing and re-use of scientific data generated through NIH-funded research, and for the agency to develop a set of clearly-defined policies on data sharing. Our interpretation of the information provided through NOT-OD-19-014 is that NIH is aiming to create a policy that lists only broad principles and requirements, and leave to the investigators to develop and propose application-specific data management and sharing plans. While as scientists we appreciate the flexibility of this position, the absence of clearly stated data-sharing guidelines could hinder investigators from meeting the NIH data-sharing aspirations.

In response to the specific questions posed by NIH, the STRENDA Commission has the following recommendations:

I. The definition of Scientific Data

We suggest to define formally *scientific data* as *the results of observations performed under specified conditions*. This definition is in accordance with the NIH principle that scientific data

³ K. F. Tipton, R.N. Armstrong, B. M. Bakker, A. Bairoch, A. Cornish-Bowden, P. J. Halling, J.-H. Hofmeyr, T. S. Leyh, C. Kettner, F. M. Raushel, J. Rohwer, D. Schomburg, C. Steinbeck (2014) Standards for Reporting Enzyme Data: The STRENDA Consortium: What it aims to do and why it should be helpful. *Perspectives in Science* **1**, 131-137. DOI: [10.1016/j.pisc.2014.02.012](https://doi.org/10.1016/j.pisc.2014.02.012)

⁴ P. Halling, P. Fitzpatrick, F. M. Raushel, J. Rohwer, S. Schnell, U. Wittig, R. Wohlgenuth and C. Kettner (2018). An empirical analysis of enzyme function reporting for experimental reproducibility: missing/incomplete information in published papers. *Biophysical Chemistry* **242**, 22-27. DOI: [10.1016/j.bpc.2018.08.004](https://doi.org/10.1016/j.bpc.2018.08.004)

should consist of “individual level and summary or aggregate data, as well as metadata.” Scientific data should not include lab notebooks, and other early outputs of the research process. However, it should incorporate information about the mathematical, statistical and computational approaches used to analyze and interpret the data (statistical protocols, algorithms, software, etc).

For the specific case of enzyme-function data, the STRENDA Commission presently defines data at two levels:

- **List Level 1A.** Data required for a complete description of an experiment. For more details: DOI: [10.3762/strenda.17](https://doi.org/10.3762/strenda.17)
- **List Level 1B.** Description of enzyme-activity data. For more details: DOI: [10.3762/strenda.27](https://doi.org/10.3762/strenda.27)

We want to empathize that without the use of standard guidelines (similar to those listed above), data sharing will not be useful for secondary analysis or replication of results. The goal of standard guidelines for sharing data is to increase scientific progress. Therefore, the STRENDA Commission urges the NIH to provide as much guidance as possible on the necessary elements for both data and metadata, and wherever possible, provide examples of accepted standards – like the STRENDA Guidelines – to increase the usability and interoperability of the data reported as outcomes of NIH funded research.

II. The requirements for Data Management and Sharing Plans

The STRENDA Commission supports a requirement to include Data Management and Sharing Plans as a part of proposals for NIH-funded research, both for its role in promoting future data sharing as well as its utility as a planning and compliance document.

We recommend that the NIH request their extra- and intramurally funded scientists adopt the STRENDA Guidelines for reporting kinetics and equilibrium data. In the STRENDA Guidelines,

All reports of kinetic and binding data must include a description of the identity of the catalytic or binding entity (enzyme, protein, nucleic acid or other molecule). This information should include the origin or source of the molecule, its purity, composition, and other characteristics such as post-translational modifications, mutations, and any modifications made to facilitate expression or purification. The assay methods and exact experimental conditions of the assay must be fully described if it is a new assay or provided as a reference to previously published work, with or without modifications.

The temperature, pH and pressure (if other than atmospheric) of the assay must always be included, even if previously published. In instances where catalytic

activity or binding cannot be detected, an estimate of the limit of detection based on the sensitivity and error analysis of the assay should be provided. Ambiguous terms such as “not detectable” should be avoided. A description of the software used for data analysis should be included along with calculated errors for all parameters.

First-order and second-order rate constants should be reported in units of s^{-1} and $M^{-1}\cdot s^{-1}$, respectively. Equilibrium binding constants should normally be reported as dissociation constants with units of concentration (M, mM, μ M, nM). The values k_{cat} , k_{cat}/K_m and K_m from steady-state enzyme kinetics should be reported in units of s^{-1} , $M^{-1}\cdot s^{-1}$ and concentration (mM, μ M, nM), respectively. The steady-state specific activity of an enzyme should normally be reported as a k_{cat} . If there is considerable uncertainty in the molar concentration of the catalyst, the specific activity should be reported as a V_{max} (nmol, μ mol) of product formed per amount of protein per unit time (e.g. μ mol \cdot mg $^{-1}\cdot$ s $^{-1}$).

With regard to data preservation and storage, the STRENDA Commission recommends that scientists deposit their enzyme assay conditions and protein-functional data in the STRENDA Database (STRENDA DB, <https://www.beilstein-strenda-db.org/strenda/index.xhtml>).⁵ The STRENDA Commission is currently working with other repositories (SABI-RK, Brenda, BioCatNet) to maximize utility of enzyme data across different platforms.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards.

We urge the NIH to consider that a new data management and sharing policy will require changes to the research process and additional time for preparing and writing grants. Data management and sharing policies maybe also require changes in infrastructure, and resources available to individual investigators through their laboratories or institutions.

NIH should provide resources to cover any reasonable cost to implement data management and sharing policies, which could include training, education, guidance, financial resources for and the development of the tools needed to enable and facilitate policy implementation. For example, there are not enough data scientists in the protein-function field to curate the data available in the literature, or to manage the data being generated by scientists.


⁵ N. Swainston, A. Baici, B. M. Bakker, A. Cornish-Bowden, P. F. Fitzpatrick, P. Halling, T. S. Leyh, C. O'Donovan, F. M. Raushel, U. Reschel, J. M. Rohwer, S. Schnell, D. Schomburg, K. F. Tipton, M.-D. Tsai, H. V. Westerhoff, U. Wittig, R. Wohlgemuth, and C. Kettner (2018). STRENDA DB: enabling the validation and sharing of enzyme kinetics data. *FEBS Journal* **285**, 2193-2204. DOI: [10.1111/febs.14427](https://doi.org/10.1111/febs.14427)

We encourage the NIH to ensure that the Data Management and Sharing Plans policy be evaluated regularly and be sufficiently flexible to keep pace with rapidly changing technologies as science is continuously evolving.

The STRENDA Commission recommends the immediate adoption of the STRENDA Guidelines for reporting enzyme functional experiments and data as part of the Data Management and Sharing Plan. We also recommend the deposition of enzyme assays protocols and results in STRENDA DB.

On behalf of the STRENDA Commission, we want to express our gratitude to NIH for engaging stakeholders at this early stage of the policy process. We look forward to continued engagement on this important issue. Please feel free to contact us, Thomas S. Leyh, Ph.D., Professor of Immunology & Microbiology at Albert Einstein College of Medicine (tom.leyh@einstein.yu.edu) and Santiago Schnell, PhD, John A. Jacquez Collegiate Professor of Physiology at the University of Michigan Medical School (schnells@umich.edu) with any questions about the STRENDA Commission comments.

Yours faithfully,


Santiago Schnell, DPhil (Oxon), FRSC
John A. Jacquez Professor of Physiology
University of Michigan Medical School

Digitally signed by Tom Leyh
Thomas S. Leyh, Ph.D.
Professor of Immunology & Microbiology
Albert Einstein College of Medicine

Attachments:

1. STRENDA Background Information
2. STRENDA Guidelines

Background Information

The STRENDA Commission is formed by an international panel of highly-regarded scientists who bring in diverse expertises such as biochemistry, enzyme nomenclature, bioinformatics, systems biology, modelling, mechanistic enzymology and theoretical biology.

The current members are:

Barbara M. Bakker	University Medical Center Groningen, The Netherlands
Antonio Baici	University of Zurich, Switzerland
Athel Cornish-Bowden	CNRS-BIP, Marseille, France
Paul F. Fitzpatrick	University of Texas Health Science Center, San Antonio, TX, USA
Peter Halling	University of Strathclyde, Glasgow, United Kingdom
Carsten Kettner	Beilstein-Institut, Frankfurt am Main, Germany
Thomas S. Leyh	The Albert-Einstein-College, Bronx, NY, USA
Claire O'Donovan	European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Cambridge, United Kingdom
Frank M. Raushel	Texas A&M University, College Station, TX, USA
Johann M. Rohwer	University of Stellenbosch, South Africa
Santiago Schnell	University of Michigan, Ann Arbor, MI, USA
Dietmar Schomburg	Technical University Braunschweig, Germany
Santiago Schnell	University of Michigan, Ann Arbor, MI, USA
Neil Swainston	The University of Manchester, United Kingdom
Ming-Daw Tsai	Academia Sinica, Taipei, Taiwan
Ulrike Wittig	Heidelberg Institute for Theoretical Studies, Germany
Roland Wohlgemuth	Sigma-Aldrich, Buchs, Switzerland

The STRENDA Commission, founded under the auspices of the Beilstein-Institut (www.beilstein-strenda.org), was set up with the aim of improving the reporting quality of functional enzyme data in the literature. STRENDA stands for “Standards for Reporting Enzymology Data”.

Enzyme activity data can be found in large quantities in the scientific literature and publicly available databases. However, detailed examination shows quickly that enzyme kinetics data are routinely measured under different experimental conditions, including temperature, pH, ionic strength, enzyme and substrate concentration, and the presence of activators and inhibitors. In order for readers to interpret the meaning of the kinetic data, the unambiguous and comprehensive definition and reporting of such accompanying information (metadata) is essential. However, in publications, this metadata and kinetic data itself is often incompletely or ambiguously reported.

Such omissions inhibit the comparison of enzyme activity data between publications and also prevent the interpretation, corroboration and reproduction of findings from single papers.

The Commission has drawn up the STRENDA Guidelines, in a consensus-driven process and in close consultation with the wider community. These Guidelines have been designed to ensure that published descriptions of enzymatic assays contained all of the information necessary for replication of the results. Over 50 journals that publish enzymological data recommend that authors adhere to these guidelines (see below).

To further support the author in adhering to these standards and to facilitate readers and reviewers in interpreting, evaluating and corroborating the experimental findings, the Beilstein-Institut, together with the STRENDA Commission, have developed and released STRENDA DB (<http://www.strenda-db.org>).

The database contains enzyme functional data in a standard and searchable format and will be a resource for enzymologists and systems biologists, greatly simplifying the task of accessing published kinetics data for enzymes along with the corresponding experimental conditions.

The following journals have adopted the Guidelines:

- ACS Catalysis
- Antimicrobial Agents and Chemotherapy
- Applied and Environmental Microbiology
- Archives in Biochemistry and Biophysics
- Biochemical Journal
- Biochemical and Biophysical Research Communications
- Biochimica et Biophysica Acta (alle nine sections)
- Biochemistry
- Biophysical Chemistry
- Biophysical Journal
- Clinical and Vaccine Immunology
- Enzyme Engineering
- FEBS Journal
- Free Radical Research
- Genome Announcements
- Infection and Immunity
- Journal of Clinical Microbiology
- Journal of the American Chemical Society
- mBio
- Molecular and Cellular Biology
- mSphere
- Proceedings of the National Academy of Sciences USA
- The Journal of Bacteriology
- The Journal of Biological Chemistry
- The Journal of Virology
- Trends in Biotechnology

The following publishers, journals and initiatives recommend the authors to use the normative checklists from the BioSharing.org portal when publishing their research results.

Publishers

- BioMedCentral (e.g. BMC Biochemistry, BMC Bioinformatics, BMC Biological Research, BMC Biology, BMC Biotechnology, BMC Cell Biology, BMC Systems Biology, ...)
- PLoS (e.g. PLoS One, PLoS Biology, PLoS Computational Biology, PLoS Genetics, PLoS Medicine, ...)

Journals

- eLife
- FEBS Lett.
- Journal of Biomedical Science
- Nature Chemical Biology
- OMICS - A Journal of Integrative Biology
- Science

The following journals and publishers recommend their authors to share their data using STRENDA DB:

- Archives in Biochemistry and Biophysics
- Beilstein Journal of Organic Chemistry
- eLife
- Nature (including Biotechnology, Chemistry, Microbiology, Pharmacology, Systems Biology)
- PLoS (relevant journals, e.g. One, Biology, Computational Biology, Medicine)
- Scientific Data
- The Journal of Biological Chemistry

Further reading (selection):

Swainston, N., Baici, A., Bakker, B.M., Cornish-Bowden, A., Fitzpatrick, P.F., Halling, P., Leyh, T.S., O'Donovan, C., Raushel, F.M., Reschel, U., Rohwer, J.M., Schnell, S., Schomburg, D., Tipton, K.F., Tsai, M.-D., Westerhoff, H.V., Wittig, U., Wohlgemuth, R. and Kettner, C. (2018) STRENDA DB: enabling the validation and sharing of enzyme kinetics data. *The FEBS J.* doi:10.1111/febs.14427.

Tipton, K.F., Armstrong, R.N., Bakker, B.M., Bairoch, A., Cornish-Bowden, A., Halling, P.J., Hofmeyr, J.-H.S., Leyh, T.S., Kettner, C., Raushel, F.M., Rohwer, J., Schomburg, D., Steinbeck, C. (2014) Standards for Reporting Enzyme Data: The STRENDA Consortium: What it aims to do and why it should be helpful. *Persp. in Sci.* 1(1-6):131-137. doi:10.1016/j.pisc.2014.02.012.

Apweiler, R., Armstrong, R., Bairoch, A., Cornish-Bowden, A., Halling, P.J., Hofmeyr, J.-H.S., Kettner, C., Leyh, T.S., Rohwer, J., Schomburg, D., Steinbeck, C. and Tipton, K.T. (2010) A large-scale protein-function database. *Nature Chem. Biol.* 6: 785. DOI:10.1038/nchembio.460.

STRENDA Guidelines Level 1A

Version 1.7 doi:10.3762/strenda.17

as approved on 22nd September 2016

The STRENDA Commission (Standards for Reporting Enzymology Data) compiled the following Guidelines, as a service to the community, to define the minimum amount of information that should accompany any published enzyme activity data.

The current STRENDA Guidelines (List Level 1A) was reviewed on the STRENDA meeting in September 2016 in terms of consistency of form and content, as well as of the order and plausibility of the list entries.

List Level 1A

defines data that are recommended for the methods section for publishing enzyme data.

This information should allow the reproducibility of the results.

Data	Comments
Identity of the enzyme	
Name of reaction catalyst	name, preferably the accepted name from the IUBMB Enzyme List
EC number	
Sequence accession number	
Organism/species & strain	NCBI Taxonomy ID
Additional information on the enzyme	
Isoenzyme	naturally occurring variant
Tissue	
Organelle	
Localization	within cell. Specify what localization is based on
Post-translational modification	add only when determined
Preparation	
Description	<i>e.g.</i> , commercial source, procedure used or reference along with modifications
Artificial modification	<i>e.g.</i> , truncated, His-tagged, fusion protein, lacking native glycosylation

STRENDA Guidelines Level 1A

Data	Comments
enzyme or protein purity	purity defined by which criteria. Specify whether protein or enzyme was purified. <i>e.g.</i> , apparently homogeneous by PAGE, crude mitochondrial fraction, determined by MS
Metalloenzyme	Mutant, content, cofactors
Storage Conditions	
Storage temperature	If frozen, freezing method, <i>e.g.</i> , -20 °C flash
Atmosphere if not air	
pH	<i>e.g.</i> , pH 7.0
At which temperature was the pH measured?	<i>e.g.</i> , 25 °C
Buffer & concentrations (including counter-ion)	<i>e.g.</i> , 200 mM potassium phosphate, 100 mM HEPES-KOH
Metal salt(s) & concentrations	<i>e.g.</i> , 10 mM KCl, 1.0 mM MgSO ₄
Other components	<i>e.g.</i> , 1.0 mM EDTA, 1.0 mM dithiothreitol, 10% glycerol, 20% DMSO, 1 mg/ml PEG2000, 2 mg/ml BSA, peptidase inhibitors
Enzyme/protein concentration	Molar concentration if known, otherwise mass concentration, <i>e.g.</i> , mg ml ⁻¹ or better: μM
Optional:	<i>e.g.</i> , less than 10% loss after 1 month
Statement about observed loss of activity under the above conditions	
Statement about the thawing procedure	<i>e.g.</i> , on ice
Assay Conditions	
Substrate purity	Origin of substrate
Measured reaction	as a stoichiometrically balanced equation <i>e.g.</i> , 2 mol substrate oxidized per mol O ₂ consumed
Assay temperature	
Assay pressure	if it is not atmospheric; indicate if not aerobic
Atmosphere if not air	

STRENDA Guidelines Level 1A

Data	Comments
Assay pH	How was it measured?
Buffer & concentrations	<i>e.g.</i> , 100 mM Tris-HCl, 200 mM potassium phosphate, including counter-ion
Metal salt(s) & concentrations	<i>e.g.</i> , 10 mM KCl, 1.0 mM MgSO ₄
Other assay components	<i>e.g.</i> , 1.0 mM EDTA, 1.0 mM dithiothreitol
Coupled assay components	if relevant
Substrate & concentration ranges	<i>e.g.</i> , 1 - 100 mM glucose, 5 mM ATP
Enzyme/protein concentration	Molar concentration if known, otherwise mass concentration. <i>e.g.</i> mg ml ⁻¹ or better: μM
Varied components	<i>e.g.</i> inhibitor concentration
Total assay mixture ionic strength	
Activity	
Initial rates of the reaction measured	determine how established. <i>e.g.</i> true initial tangent or average over specified time
Proportionality between initial velocity and enzyme concentration	if available
Enzyme activity	Ideally k_{cat} otherwise expressed as amount product formed per amount enzyme protein present - sometimes referred to as enzyme unit or international unit (1 IU = 1 μmol min ⁻¹). The katal (mol/s) may alternatively be used as a unit of activity (conversion factor 1 unit = 16.67 nkat).
Methodology	
Assay method	a literature reference may suffice for an established procedure but any modification should be detailed
Type of assay	<i>e.g.</i> , continuous or discontinuous, direct or coupled
Reaction stopping procedure	in the case of discontinuous assays
Direction of the assay	with respect to the reaction equation provided <i>e.g.</i> , NAD reduction by alcohol dehydrogenase; alcohol + NAD ⁺ → aldehyde or ketone + NADH + H ⁺

STRENDA Guidelines Level 1A

Data	Comments
Reactant determined	<i>e.g.</i> , NADH formation, O ₂ utilization
Additional material desirable	
Free metal cation concentrations	<i>e.g.</i> , of Mg ²⁺ and Ca ²⁺ , specify how calculated
Reaction equilibrium constant	define conditions and reaction direction

About the STRENDA Commission:

The Commission was founded in 2003 and is supported by the Beilstein-Institut since then. Members of the Commission are: R.N. Armstrong† (Vanderbilt University, Nashville, TN, USA), A. Bairoch (University of Geneva, Switzerland), Barbara M. Bakker (University Medical Center Groningen, The Netherlands), A. Cornish-Bowden (CNRS-BIP, Marseilles, France), P. Fitzpatrick (University of Texas Health Science Center at San Antonio, San Antonio, TX, USA), P. Halling (University of Strathclyde, Glasgow, UK), T.S. Leyh (The Albert Einstein College of Medicine, Bronx, NY, USA), C. O'Donovan (EBI, Cambridge, UK), F. Raushel (Texas A&M University, College Station, TX, USA), J. Rohwer (University of Stellenbosch, South Africa), S. Schnell (University of Michigan, Ann Arbor, MI, USA), D. Schomburg (Technical University of Braunschweig, Germany), N. Swainston (The University of Manchester, UK), M.-D. Tsai (Academia Sinica, Taipei, Taiwan), K. Tipton (Trinity College, Dublin, Ireland), R. Wohlgemuth (Sigma-Aldrich, Buchs, Switzerland) and C. Kettner (co-ordination, Beilstein-Institut, Frankfurt, Germany).

More information: www.beilstein-strenda.org

STRENDA Guidelines Level 1B

Version 1.7

doi:10.3762/strenda.27

as approved on 22nd September 2016

The STRENDA Commission (Standards for Reporting Enzymology Data) compiled the following Guidelines, as a service to the community, to define the minimum amount of information that should accompany any published enzyme activity data.

The current STRENDA Guidelines (List Level 1B) was reviewed on the STRENDA meeting in September 2016 in terms of consistency of form and content, as well as of the order and plausibility of the list entries.

List Level 1B

defines those data that are required to allow a quality check on the data and to ensure their value to others. In principle, this is the minimum information to describe enzyme activity data.

Information required	Comments
Required data for all enzyme functional data	
Number of independent experiments	any problems of reproducibility should be stated
Precision of measurement	<i>e.g.</i> , standard error of the mean, standard deviation, confidence limits, quartiles
Specification whether relative to subunit or oligomeric form	
Data necessary for reporting kinetic parameters	
k_{cat}	V_{max} may be divided by the specific activity units, measured in s^{-1} or min^{-1}
V_{max}	V_{max} given as units, as defined in List Level 1A
$k_{\text{cat}}/K_{\text{m}}$	$k_{\text{cat}}/K_{\text{m}}$ given as concentration per time <i>e.g.</i> , $\text{mM}^{-1}\text{s}^{-1}$
K_{m}	units or concentration necessary, <i>e.g.</i> , mM
$S_{0.5}$	concentration, <i>e.g.</i> , mM
Hill coefficient, saturation ratio (RS) or other coefficients of cooperativity	

STRENDA Guidelines Level 1B

Information required	Comments
How was the given parameter obtained?	<i>e.g.</i> , non-linear curve fitting using least squares, non-parametric method such as direct linear plot, linear regression to transformed form of rate equation. Note: if commercial computer programs are used, determine which were used
Model used to determine the parameters	with explanation of why is the chosen model considered to be the “right” model
High-substrate inhibition, if observed, with K_i value	
Data required for reporting inhibition data	
Time-dependence and reversibility	with method described
Inhibition types:	K_i units necessary
reversible	<i>e.g.</i> , competitive, uncompetitive, etc., with units and how values were determined
tight-binding	association/dissociation rates
irreversible	<i>e.g.</i> , non-specific, mechanism-based, “suicide substrate”.
	There are too many alternative parameters to list here. The reference to a quite comprehensive source is recommended: Enzymes: Irreversible Inhibition. Tipton, K.F. In: Nature Encyclopedia of Life Sciences London (2001). http://www.els.net/ [doi:10.1038/npg.els.0000601]
	Note: IC_{50} values These have been used for both reversible or irreversible inhibition. However, the use is not recommended because these values are without a consistent meaning. The relationship of these values to inhibition constants is analysed in details, <i>e.g.</i> , by Cortes, A. <i>et al.</i> (2001) <i>Biochem. J.</i> 357:263-268.
Data required for reporting activation data	Similar to the requirements for inhibition data

STRENDA Guidelines Level 1B

About the STRENDA Commission:

The Commission was founded in 2003 and is supported by the Beilstein-Institut since then. Members of the Commission are: R.N. Armstrong† (Vanderbilt University, Nashville, TN, USA), A. Bairoch (University of Geneva, Switzerland), Barbara M. Bakker (University Medical Center Groningen, The Netherlands), A. Cornish-Bowden (CNRS-BIP, Marseilles, France), P. Fitzpatrick (University of Texas Health Science Center at San Antonio, San Antonio, TX, USA), P. Halling (University of Strathclyde, Glasgow, UK), T.S. Leyh (The Albert Einstein College of Medicine, Bronx, NY, USA), C. O'Donovan (EBI, Cambridge, UK), F. Raushel (Texas A&M University, College Station, TX, USA), J. Rohwer (University of Stellenbosch, South Africa), S. Schnell (University of Michigan, Ann Arbor, MI, USA), D. Schomburg (Technical University of Braunschweig, Germany), N. Swainston (The University of Manchester, UK), M.-D. Tsai (Academia Sinica, Taipei, Taiwan), K. Tipton (Trinity College, Dublin, Ireland), R. Wohlgemuth (Sigma-Aldrich, Buchs, Switzerland) and C. Kettner (co-ordination, Beilstein-Institut, Frankfurt, Germany).

More information: www.beilstein-strenda.org

Submission #151**Date:** 12/10/2018**Name:** Kitcki Carroll**Name of Organization:** United South and Eastern Tribes**Type of Organization:** Other**Other Type of Organization:** Tribal Nation**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

The United South and Eastern Tribes Sovereignty Protection Fund (USET SPF) is pleased to offer the following comments on the National Institutes of Health (NIH) Request for Information on Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research. USET SPF recognizes that sharing data among the scientific community is imperative for scientific discovery and advancement. However, it is critically important to recognize the historic relationship between scientific study and Tribal Nations, where researchers committed ethical violations against our communities and our people. American Indians and Alaska Natives (AI/AN) and Tribal communities have experienced negative impacts from the use of genomic data (*Arizona Board of Regents v. Havasupai Tribe*) without Tribal Nation informed consent. To ensure the privacy of Tribal Nation communities, as well as AI/AN individuals, USET SPF urges NIH to consult with Tribal Nations regarding its research, data sharing, and data management policies and offers the following recommendations in response to NIH's request for comment.

USET SPF is a non-profit, inter-tribal organization representing 27 federally recognized Tribal Nations from Texas across to Florida and up to Maine. Both individually, as well as collectively through USET SPF, our member Tribal Nations work to improve health care services for American Indians. Our member Tribal Nations operate in the Nashville Area of the IHS, which contains 36 IHS and tribal health care facilities. Our citizens receive health care services both directly at IHS facilities, as well as in Tribally Operated facilities operated under contracts with IHS pursuant to the Indian Self-Determination and Education Assistance Act (ISDEAA), P.L. 93-638.

I. The definition of Scientific Data

USET SPF agrees with the definition of Scientific Data as outlined in the Proposed Provisions for a Draft NIH Data Management and Sharing Policy. USET SPF also agrees that the definition of

Scientific Data should not include ‘laboratory notebooks, preliminary analysis, completed case report forms, drafts of scientific papers, plans for future research peer reviews, communications with colleagues, or physical objects, such as laboratory specimens’ (referred to hereafter as ‘work product’) and that researchers not be required to share this information with NIH. It is important to note, however, that researchers should be expressly prohibited from refusing to share work product (and Scientific Data) with any Tribal Nation impacted by the research project.

II. The requirements for Data Management and Sharing Plans

USET SPF appreciates NIH’s forethought of requiring a data management and sharing plan as part of future application processes, proposals, cooperative research and development agreements (CRADA), or other funding agreements, or intramural research reports. As part of the US’s trust obligation to federally recognized Tribal Nations, NIH has a duty to honor, protect and uphold Tribal Nation sovereignty in its efforts to ‘seek fundamental knowledge about the nature and behavior of living systems and the application of that knowledge to enhance health, lengthen life, and reduce illness and disability.’ Therefore, it is USET SPF’s recommendation that all submitted data management and sharing plans require an element entitled ‘Tribal Nation(s) Population’ that shows, first and foremost, evidence of Tribal Nation consent for data sharing.

No Tribal Nation data should be included in any level of access without explicit Tribal Nation consent. The consent mechanism varies from Tribal Nation to Tribal Nation and may take the form of Tribal Nation Council resolutions, signed MOU’s with the designated Tribal Nation leader, etc. In addition to documented Tribal Nation consent, the plan must address additional considerations between the researcher and the Tribal Nation such as:

- Data ownership and sovereignty
- Publication requirements and Tribal Nation consent procedures
- Specimen use, storage, and destruction policy
- Work product ownership and sovereignty
- Data use provisions for future studies
- Data sharing and use provisions for NIH-maintained databases (i.e. genomics)

USET SPF must note that the above list is not exhaustive, and that NIH must seek formal Tribal consultation on these recommendations, as well as on any future draft Data Management and Sharing Plan requirements. Much as ‘The Belmont Report’ and the National Research Act of 1974 have resulted in human subject protection as standard practice among researchers, USET SPF believes that such a required element for all NIH-funded research proposals will integrate Tribal Nation protection and sovereignty concerns into common research practice. USET SPF reminds NIH that it has an obligation to ensure that Tribal Nations are able to protect their citizens and data, and this obligation supersedes any data sharing interests.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

NIH must engage in formal Tribal consultation on the proposed policy provisions and proposed required elements for data management and sharing plans, as well as the subsequent draft NIH policy for data management and sharing. These should not be considered ready for drafting or adoption prior to thorough nationwide Tribal consultation in addition to guidance from NIH's Tribal advisory committee.

Attachment:

December 10, 2018

Office of Science and Policy
National Institutes of Health
6705 Rockledge Drive, Suite 750
Bethesda, MD 20892

Re: Request for Information on Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research

The United South and Eastern Tribes Sovereignty Protection Fund (USET SPF) is pleased to offer the following comments on the National Institutes of Health (NIH) *Request for Information on Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research*. USET SPF recognizes that sharing data among the scientific community is imperative for scientific discovery and advancement. However, it is critically important to recognize the historic relationship between scientific study and Tribal Nations, where researchers committed ethical violations against our communities and our people. American Indians and Alaska Natives (AI/AN) and Tribal communities have experienced negative impacts from the use of genomic data (*Arizona Board of Regents v. Havasupai Tribe*) without Tribal Nation informed consent. To ensure the privacy of Tribal Nation communities, as well as AI/AN individuals, USET SPF urges NIH to consult with Tribal Nations regarding its research, data sharing, and data management policies and offers the following recommendations in response to NIH's request for comment.

USET SPF is a non-profit, inter-tribal organization representing 27 federally recognized Tribal Nations from Texas across to Florida and up to Maine.¹ Both individually, as well as collectively through USET SPF, our member Tribal Nations work to improve health care services for American Indians. Our member Tribal Nations operate in the Nashville Area of the IHS, which contains 36 IHS and tribal health care facilities. Our citizens receive health care services both directly at IHS facilities, as well as in Tribally Operated facilities operated under contracts with IHS pursuant to the Indian Self-Determination and Education Assistance Act (ISDEAA), P.L. 93-638.

I. The Definition of Scientific Data

USET SPF agrees with the definition of Scientific Data as outlined in the *Proposed Provisions for a Draft NIH Data Management and Sharing Policy*. USET SPF also agrees that the definition of Scientific Data should not include 'laboratory notebooks, preliminary analysis, completed case report forms, drafts of scientific papers, plans for future research peer reviews, communications with colleagues, or

¹ USET SPF member Tribal Nations include: Alabama-Coushatta Tribe of Texas (TX), Aroostook Band of Micmac Indians (ME), Catawba Indian Nation (SC), Cayuga Nation (NY), Chitimacha Tribe of Louisiana (LA), Coushatta Tribe of Louisiana (LA), Eastern Band of Cherokee Indians (NC), Houlton Band of Maliseet Indians (ME), Jena Band of Choctaw Indians (LA), Mashantucket Pequot Indian Tribe (CT), Mashpee Wampanoag Tribe (MA), Miccosukee Tribe of Indians of Florida (FL), Mississippi Band of Choctaw Indians (MS), Mohegan Tribe of Indians of Connecticut (CT), Narragansett Indian Tribe (RI), Oneida Indian Nation (NY), Passamaquoddy Tribe at Indian Township (ME), Passamaquoddy Tribe at Pleasant Point (ME), Pamunkey Indian Tribe (VA), Penobscot Indian Nation (ME), Poarch Band of Creek Indians (AL), Saint Regis Mohawk Tribe (NY), Seminole Tribe of Florida (FL), Seneca Nation of Indians (NY), Shinnecock Indian Nation (NY), Tunica-Biloxi Tribe of Louisiana (LA), and the Wampanoag Tribe of Gay Head (Aquinnah) (MA).

physical objects, such as laboratory specimens' (referred to hereafter as 'work product') and that researchers not be required to share this information with NIH. It is important to note, however, that researchers should be expressly prohibited from refusing to share work product (and Scientific Data) with any Tribal Nation impacted by the research project.

II. The Requirements for Data Management and Sharing Plans

USET SPF appreciates NIH's forethought of requiring a data management and sharing plan as part of future application processes, proposals, cooperative research and development agreements (CRADA), or other funding agreements, or intramural research reports. As part of the US's trust obligation to federally recognized Tribal Nations, NIH has a duty to honor, protect and uphold Tribal Nation sovereignty in its efforts to 'seek fundamental knowledge about the nature and behavior of living systems and the application of that knowledge to enhance health, lengthen life, and reduce illness and disability.' Therefore, it is USET SPF's recommendation that all submitted data management and sharing plans require an element entitled 'Tribal Nation(s) Population' that shows, first and foremost, evidence of Tribal Nation consent for data sharing.

No Tribal Nation data should be included in any level of access without explicit Tribal Nation consent. The consent mechanism varies from Tribal Nation to Tribal Nation and may take the form of Tribal Nation Council resolutions, signed MOU's with the designated Tribal Nation leader, etc. In addition to documented Tribal Nation consent, the plan must address additional considerations between the researcher and the Tribal Nation such as:

- Data ownership and sovereignty
- Publication requirements and Tribal Nation consent procedures
- Specimen use, storage, and destruction policy
- Work product ownership and sovereignty
- Data use provisions for future studies
- Data sharing and use provisions for NIH-maintained databases (i.e. genomics)

USET SPF must note that the above list is not exhaustive, and that NIH must seek formal Tribal consultation on these recommendations, as well as on any future draft Data Management and Sharing Plan requirements. Much as 'The Belmont Report' and the National Research Act of 1974 have resulted in human subject protection as standard practice among researchers, USET SPF believes that such a required element for all NIH-funded research proposals will integrate Tribal Nation protection and sovereignty concerns into common research practice. USET SPF reminds NIH that it has an obligation to ensure that Tribal Nations are able to protect their citizens and data, and this obligation supersedes any data sharing interests.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards.

NIH must engage in formal Tribal consultation on the proposed policy provisions and proposed required elements for data management and sharing plans, as well as the subsequent draft NIH policy for data

management and sharing. These should not be considered ready for drafting or adoption prior to thorough nationwide Tribal consultation in addition to guidance from NIH's Tribal advisory committee.

IV. Compliance and Enforcement

USET SPF strongly suggests that an oversight mechanism, specific to Tribal Nation data be created by NIH. This mechanism would detail Tribal Nation data protection best practices, procedures, ensure researcher compliance, and recommend consequences for violations.

Conclusion

USET SPF appreciates this opportunity to provide comments on the NIH Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research. Because data management and sharing policies has significant implications for Tribal governments and their citizens, we urge NIH to seek formal Tribal Consultation on this issue. Should you have any questions or require additional information, please do not hesitate to contact Mr. Kitcki Carroll, USET Executive Director, at (615) 467-1540 or by e-mail at kcarroll@usetinc.org.

Sincerely,



Kirk Francis
President



Kitcki A. Carroll
Executive Director

Submission #152**Date:** 12/10/2018**Name:** Yvette Roubideaux MD MPH**Name of Organization:** NCAI Policy Research Center**Type of Organization:** Nonprofit Research Organization**Role:** Scientific Researcher**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

American Indian/Alaska Native Policy Research - all NIH research areas

I. The definition of Scientific Data

For the purpose of the NIH Data Management and Sharing Policy, in Section I, Definitions, the definition of “Scientific Data” should be modified to include.....”data used to support scholarly publications and scholarly presentations.” Often scientists conduct research that may not be submitted or accepted for publication and use scholarly presentations as a method to disseminate their data to advance scientific knowledge. Scholarly presentations share data to help advance the field of knowledge, which is fundamental to the definition of research. This definition should also note that American Indian and Alaska Native tribes, as sovereign nations, have the right to determine their own definition of data in their own data sharing and management plans and tribal research codes. The NIH Common Rule affirms that researchers with federal funding must follow tribal research codes, which can be more restrictive than the Common Rule. Nothing in this data management and data sharing policy should conflict with the ability of a tribe to establish its own definitions in its own research codes and data management and sharing policies for researchers it chooses to partner with on research. NIH should consult with American Indian and Alaska Native tribes on this draft NIH Data Management and Sharing Policy before a final policy is implemented and report back on how their input was incorporated into the final version.

II. The requirements for Data Management and Sharing Plans

For the purpose of the NIH Data Management and Sharing Policy, in Section IV, “Requirements for Data Management and Sharing Plans,” at the end of the first paragraph, the following should be added: “All Plans involving research and Scientific Data with American Indian and Alaska Native tribe(s) should include specific information on how the Plan complies with their tribal research codes, documentation of official tribal approval(s) for the Plan, and should

describe in detail how the Plan implements tribal requirements and preferences on data management and sharing to ensure that tribal nation(s) and their citizens, lands, and resources are protected, along with how the Plan will implement any tribal restrictions to data sharing.”

Under “Plan Review and Evaluation”, add a new bullet at the end of the list that states the following: “For all Extramural Grants, Contracts, NIH Intramural Research Projects, and Other funding/support agreements, the Plan should be determined to be “unacceptable” by reviewers or NIH staff if the Plan involves data from American Indian and Alaska Native tribes and does not include specific information on how the Plan complies with tribal research codes, documentation of official tribal approval(s) for the Plan, and detailed descriptions of how the Plan implements tribal requirements and preferences on data management and sharing to ensure that tribal nation(s) and their citizens, land and resources are protected, along with how the Plan will implement any tribal restrictions to data sharing.”

Under “Plan Elements”, add a new number 4 that states the following: “All Plans should indicate if their data includes data and information from American Indian and Alaska Native tribes or individuals, and if so, the Plan should include a one page addendum that describes in detail how the data was obtained, whether there is/was documentation of American Indian/Alaska Native tribal approval to obtain the data, and documentation of approval(s) from American Indian and/or Alaska Native tribes for the Plan for data sharing. The Plan should describe for each element listed above how the Plan will implement any tribal requirements or restrictions relevant to each element of the Plan. The Plan should have documentation that the tribe(s) affirmatively approve each element of the Plan.” Examples of how the tribe may approve each element of the Plan include the following: the tribal must approval all types of data that may be collected and shared, and reserves the right to not approve certain or any types of data to be shared; the tribe must approve all other information, including relevant associated data, that may be shared; the tribe must approve the methods of how the data will be processed or analyzed; the tribe must approve any standards used in data collection and sharing. Even though some tribes may value and encourage data sharing, any Plan should clearly include information that affirms any tribal approvals, requirements, restrictions, or denials of any or all elements of the Plan. NIH should consult with American Indian and Alaska Native tribes on this draft NIH Data Management and Sharing Policy before a final policy is implemented and report back on how their input was incorporated into the final version.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

NIH must consult as soon as possible with American Indian and Alaska Native tribes on any data management and sharing policy before implementation. According to the Department of Health and Human Services Tribal Consultation Policy, which applies to all Divisions in the Department including NIH, “Before any action is taken that will significantly affect Indian Tribes

it is the HHS policy that, to the extent practicable and permitted by law, consultation with Indian Tribes will occur. Such actions refer to policies that... have tribal implications and have substantial direct effects on one or more Indian Tribes..." NIH's draft Data Sharing and Management Policy meets this definition of an action that will significantly affect Indian tribes and thus requires tribal consultation before it is implemented.

To our knowledge, NIH has received a large amount of input from American Indian and Alaska Native tribes and other individuals with related expertise on the importance of developing its Data Sharing and Management plan in partnership with tribes and the importance of incorporating the input and recommendations of tribes to protect their sovereign rights to govern research and data that involves their citizens, lands, and resources. As of the date of this submission, American Indian and Alaska Native tribes and others have expressed in public forums their dissatisfaction with NIH's efforts to date to incorporate tribal recommendations into this draft policy.

We recommend that NIH immediately consider this draft NIH data sharing and management policy to represent a Critical Event, affirm that it impacts all 573 American Indian and Alaska Native tribes, and immediately initiate tribal consultation through a process that begins with a letter to all tribes with the current draft of the policy attached. NIH should understand that a Request for Information released to the public is not a form of tribal consultation. NIH should initiate a tribal consultation on this draft policy as soon as possible and should not wait for the draft NIH tribal consultation protocol under development by the NIH Tribal Research Office to be finalized. NIH should allow for in-person tribal consultation, should carefully review tribal input, and should communicate with all tribes how their recommendations will be reflected in the final draft of the NIH Data Sharing and Management Policy and give tribes one more opportunity at that point to consult that the final draft. All input received by tribes to date should be summarized by NIH and shared with tribes at the onset of this tribal consultation. NIH should understand that tribal consultation is a policy of the Department of Health and Human Services and that any policies developed without tribal consultation represent a litigation risk to the agency, the Department, and individual investigators. The topic of data sharing and management is a priority topic for American Indian and Alaska Native tribes who must have the opportunity to exercise their sovereignty over data management and data sharing, both of which have potential significant risks to their citizens, lands and resources.

Finally, the NCAI Policy Research Center submitted comments on the Draft NIH Genomic Data Sharing Policy in 2013 and in those comments the following five overarching principles were highlighted:

- Tribal nations have sovereignty over research conducted on tribal lands and with tribal citizens;

- Researchers must secure active tribal approval for the collection, use, and sharing of tribal data;
- There are successful models of tribally-driven data sharing that serve to both protect and benefit Native people;
- Research ethics need to acknowledge the importance of community consent alongside individual consent; and
- Research ethics need to include protections for biological samples collected from both living and deceased human beings.

Given that American Indian and Alaska Native tribes have experienced the harmful effects of inappropriate research and data sharing, and in some cases continue to experience those harmful effects, ensuring that a meaningful tribal consultation on the NIH data management and sharing policy is of utmost importance and urgency. Tribes understand the importance of research and data to eliminating disparities and can be helpful in developing recommendations and solutions to data management and sharing policies that respect their sovereignty, the government to government relationship between tribes and the federal government, and the rights of tribes to enter into respectful partnerships with researchers on data management and data sharing.

Submission #153

Date: 12/10/2018

Name: Jeffery Smith

Name of Organization: AMIA

Type of Organization: Professional Org/Association

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

Biomedical Informatics

I. The definition of Scientific Data

Please see the attached PDF.

II. The requirements for Data Management and Sharing Plans

Please see the attached PDF.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

Please see the attached PDF.

Attachment:



December 10, 2018

Carrie D. Wolinetz, Ph.D.
Associate Director for Science Policy
NIH Office of Science Policy
National Institutes of Health
6705 Rockledge Drive, Suite 750
Bethesda, MD 20892

Re: Proposed Provisions for a Draft NIH Data Management and Sharing Policy

Dr. Wolinetz:

AMIA's membership is comprised of informaticians across the spectrum of biomedical research, clinical care, public health, and consumer health, with backgrounds in medicine, biomedical sciences, and informatics. Our comments are rooted in this expertise and are representative of diverse and multidisciplinary stakeholders who are deeply experienced in the systematic collection, analysis, application and responsible sharing of data for health.

AMIA enthusiastically supports development of a pan-NIH Data Management and Sharing Policy (DMSP) and we commend the NIH for initiating this effort. We are pleased to see several elements of AMIA's Data Sharing Principles & Positions incorporated in the Proposed Provisions, including a reliance on FAIR (Findable, Accessible, Interoperable, and Re-usable) data principles, and acknowledgment that the DMSP should support underlying infrastructure and curation activities with funding.

We are especially pleased that the NIH envisions a DMSP that applies to "all intramural and extramural research, funded or supported in whole or in part by NIH, that results in scientific data, regardless of NIH funding level or mechanism." While this scope is ambitious, quality data management and sharing plans (Plans) are prerequisite to achieve the vision of FAIR data principles and such a scope should be the long-term goal of the NIH DMSP.

Recognizing the need to have all NIH-funded research comply with this DMSP, and with appreciation of what AMIA sees as necessary components of a data management and sharing plan, we recommend a phased compliance timeline based on funding levels. This phased implementation would only apply to new research funded after the DMSP is final. First, new research funded above \$500,000 per year and subject to the existing data sharing policy should comply with the final DMSP within one year of its adoption. Second, new research funded above \$250,000 per year should comply with the provisions of the DMSP within 2 years of its adoption, and finally, all grants funded below \$250,000 per year should comply with the DMSP within 3 years of adoption. This compliance approach would focus efforts on those grants that already must comply with the existing policy and

likely have the richest cache of scientific data, while giving smaller projects more time to become familiar with the DMSP.

Alongside this phased adoption timeline, the NIH should consider a graduated DMSP that appropriately calibrates requirements based on funding level and whether scientific data are deposited in an NIH-endorsed depository or knowledgebase. We strongly recommend that the draft DMSP encourage ICs to factor the quality of the Plan into the overall impact score through the peer review process for those grants that are supported at high levels or support programmatic priorities. We also recommend that NIH incentivize deposition of scientific data in NIH-endorsed databases and knowledgebases by allowing such Plans to comply with a streamlined DMSP.

We note several high-level observations and recommendations for which we provide additional detail and rationale in the enclosure of this comment letter:

- 1. The draft DMSP should improve data management and sharing of scientific data to facilitate learning health systems and continuous discovery.**
 - a. While we are supportive of a pan-NIH DMSP, subject to individual Institutes and Center (ICs) specific grant-types and awardees, AMIA recommends the DMSP encourage ICs to make Plans scorable elements of specific grants. This will improve Plans' quality and better ensure supplemental use of scientific data.
 - b. AMIA also recommends the DMSP seeks to improve the interoperability and supplemental uses of research data writ large by encouraging the use of established biomedical data standards and adherence to data management and data sharing best practices. Over time, better use of and refinement of data standards, buttressed by systematic scoring of plans, will optimize scientific data for continuous learning and discovery.
 - c. AMIA recommends the DMSP incentivize the deposition of scientific data and tools, software and/or code developed as part of NIH-supported projects into NIH-approved data repositories and knowledgebases. This will enable both large and small grantees to more easily comply with the DMSP.
- 2. The draft DMSP should improve institutional support and professional advancement for experts managing and sharing scientific data.**
 - a. We applaud NIH for suggesting that reasonable costs associated with data management and sharing could be requested under the budget for the proposed project. AMIA recommends that the DMSP establish a standard way to account for data management and sharing costs as both Direct costs and F&A costs.
 - b. The DMSP should facilitate implementation of the NIH Data Science Strategic Plan, especially the relevant aspects of the Strategic Plan that seek to credit experts who manage and share valuable data sets / software for their work. If data is seen as valuable, experts who enable FAIR data should also be valued. The NIH should support certifications for experts that manage and share scientific data. We also see a need for R&D on data management tools to facilitate compliance with the DMSP.
- 3. To operationalize the DMSP more specificity and clarity around concepts is needed.**
 - a. Data management is distinct from data sharing. The processes and activities that support data management and sharing are also different. AMIA recommends the

NIH develop a DMSP that specifies these distinctions through additional Plan Elements as described below.

- b. AMIA recommends that the DMSP expand the current list of definitions to include concepts for “Data Management,” “Covered Data,” “Covered Timeframe,” and refine definitions for “Metadata” and “Scientific Data.”
- c. While we support the scope of a pan-NIH DMSP that covers all grants, contracts, and/or other funding agreements, AMIA recommends the NIH convene stakeholders with individual ICs to operationalize the DMSP.

Finally, we offer AMIA and its members as resources during subsequent work on the DMSP. We strongly recommend the NIH develop a subsequent draft DMSP based on stakeholder feedback to the concepts in this RFI. Another comment period will provide NIH with valuable insights before issuing a final DMSP.

The enclosure includes detailed AMIA comments regarding the Proposed Provisions for a Draft NIH Data Management and Sharing Policy. Where possible, we provide both in-line edits and rationale for suggested edits.

- I. [Definitions](#)
 - a. [Data Management and Sharing Plan](#)
 - b. [Data Management](#)
 - c. [Data Sharing](#)
 - d. [Metadata](#)
 - e. [Scientific Data](#)
 - f. [AMIA Recommended New Definitions](#)
- II. [Purpose](#)
- III. [Scope and Requirements](#)
- IV. [Requirements for Data Management and Sharing Plans](#)
 - a. [Plan Review and Evaluation](#)
 - b. [Plan Elements](#)
 - i. [Data Type](#)
 - ii. [Related Tools, Software and/or Code](#)
 - iii. [Standards](#)
 - iv. [Data Preservation and Access](#)
 - v. [Data Preservation and Access Timeline](#)
 - vi. [Data Sharing Agreements, Licensing, and Intellectual Property](#)
 - vii. [Oversight of Data Management](#)
 - viii. [Other Considerations](#)
- V. [Compliance and Enforcement](#)

We hope our comments are helpful as you undertake this important work. Should you have questions about these comments or require additional information, please contact Jeffery Smith, Vice President of Public Policy at jsmith@amia.org or (301) 657-1291. We look forward to continued partnership and dialogue.

Sincerely,

A handwritten signature in black ink, appearing to read "Douglas B. Fridsma". The signature is fluid and cursive, with a long horizontal stroke extending to the right.

Douglas B. Fridsma, MD, PhD, FACP, FACMI
President and CEO
AMIA

Enclosed: Detailed AMLA comments regarding the Proposed Provisions for a Draft NIH Data Management and Sharing Policy.

I. Definitions

a. Data Management and Sharing Plan

AMIA Comments: The draft Data Management and Sharing Policy (DMSP) should differentiate between “data management” and “data sharing” as two distinct concepts and sets of activities with different, if overlapping, considerations and timeframes. For clarity, we refer to the Data Management and Sharing Plan as “Plan” and DMSP refers to the policy. While we are supportive of the focus on data sharing as part of data management, it is critical to acknowledge that upstream data collection and handling processes largely determine data quality necessary for research replicability, reproducibility, and traceability.¹

AMIA Comments: The draft DMSP should not distinguish between potential “others” who may use scientific data. The number and heterogeneity of “others,” even when confined to “researchers,” and “the broader public,” would needlessly complicate compliance with the DMSP. AMIA members note a major discrepancy between making scientific data available to another scientist in the same discipline and making it available to the general public. Further, we note that even within the scientific community, there will be wide gaps in knowledge across disciplines that requires extensive annotation and training to be understood.

AMIA Recommendation: AMIA recommends the following amendments to the Plan’s definitions to acknowledge differences in data management and data sharing. Further AMIA recommends the draft DMSP remove all references to “(e.g. researchers and the broader public)” when describing potential users of scientific data:

Data Management and Sharing Plan: A plan describing how scientific data will be **generated**, managed, **described, analyzed**, preserved, shared, and made accessible to others for **supplemental uses**, (e.g., ~~other researchers and the broader public~~), as appropriate. **This plan should include two distinct sections describing how scientific data will be managed across the life-cycle of the project and how scientific data will be shared at the project close, or at another appropriate interval(s).**

b. Data Management

AMIA Comments: As discussed above, the DMSP should explicitly describe what is necessary to manage data, not just share data, given that data management and data sharing are distinct. Data management is prerequisite for data sharing, ensuring that the data are accurate, complete, and maintained in a standardized manner. Without effective data management, you cannot have effective data sharing, thus we recommend the DMSP consider additional Plan Elements as described in that section of our comments.

AMIA Recommendation: Given this view, we recommend the draft DMSP include a new definition for data management as follows:

¹ Traceability of research data is the ability to reproduce the raw data from the analysis datasets and vice versa.

Data Management: The upstream management of scientific data that documents actions taken in making research observations, collecting research data, describing data (including relationships between datasets), processing data into intermediate forms as necessary for analysis, integrating distinct datasets, and creating metadata descriptions. Specifically, those actions that would likely have impact on the quality of data analyzed, published, or shared.²

c. Data Sharing

Data Sharing: ~~To make~~ **Making** scientific data accessible for use by others (~~e.g., other researchers and the broader public~~) in a manner that is consistent with the FAIR (Findable, Accessible, Interoperable, and Re-usable) data principles.

d. Metadata

AMIA Comments: We found the definition for metadata in need of refinement. Specifically, the phrase “additional information to make data more usable” implies that a data set could be usable at all without metadata, which is simply not the case. There is no data that can be correctly understood, much less re-used, without at least a data dictionary with field definitions and data types. Further, we view “Outcome measures” as actual data, not metadata. There may be metadata that defines how an outcome measure was derived, but the outcome data itself is not metadata.

AMIA Recommendation: Given this view, we recommend the draft DMSP amend the definition of metadata as follows:

Metadata: ~~Data that provide additional information to make data more usable (e.g., independent sample and variable description, outcome measures, and any intermediate, descriptive, or phenotypic observational variables).~~ Metadata is descriptive information about data, including variable/document definition/description, data type, and other characteristics. Areas discussed in metadata include, but are not limited to, instruments used to collect data; parameters or settings for such instruments; descriptors of physical samples from which data were collected; dates and times of data collection; any transformations applied to the data; relationships between datasets; provenance linking derived or modified datasets to original sources; phenotypic descriptors of data sources; and institutional/personal identifying information associated with the group or person(s) responsible for the data. Metadata also help establish (confidence in) the credibility of the data.

In survey data, “paradata” is used to describe confidence in the credibility of data. This may be an evaluation of the sincerity or seriousness of the respondent by the questioner (e.g. “Open/Frank” to “Uncomfortable/Evasive”, "Earnest" to "Flippant", etc.), or less subjectively, in an online survey, the time the respondent spent to complete the survey.

² Adapted from Williams, Bagwell and Zozus “Data management plans, the missing perspective,” Journal of Biomedical Informatics 71 (2017) 130–142

e. Scientific Data

AMIA Comments: We support the concept of “Scientific Data,” but do not support a definition of this concept through negation. The listing of what Scientific Data is not may serve better as part of ancillary materials published by the NIH, such as Frequently Asked Questions, rather than be included in a definition. Further, it is odd to place a command, “NIH expects...” into a definition.

AMIA Recommendation: Given this view, we recommend the draft DMSP include a new definition for Scientific Data as follows:

Scientific Data: ~~The recorded factual material commonly accepted in the scientific community as necessary to validate and replicate research findings including, but not limited to, data used to support scholarly publications. Scientific data do not include laboratory notebooks, preliminary analyses, completed case report forms, drafts of scientific papers, plans for future research, peer reviews, communications with colleagues, or physical objects, such as laboratory specimens. For the purposes of a possible Policy, scientific data may include certain individual level and summary or aggregate data, as well as metadata. NIH expects that reasonable efforts should be made to digitize all scientific data.~~ 1. Information that is gathered, derived or generated in the course of conducting research. It is the basis for reaching conclusions and inferences based on scientific principles and methodologies. Scientific data can be used to test existing hypotheses, to generate new hypotheses for future research, to validate or replicate prior research as well as for more exploratory purposes. Scientific data represent the foundation for both scientific theories and publications.

f. AMIA Recommended New Definitions

1. Scientific Software Artifacts

AMIA Comment: Increasingly, the NIH funds research that results in software tools, code, and analytic programs. These “software artifacts” are both explicitly funded as part of extramural research and developed as a means to conduct NIH-funded research. These software artifacts can be deposited in knowledgebases analogous to data into databases.

AMIA Recommendation: AMIA recommends the draft DMSP includes a definition for “Scientific Software Artifacts,” so that grantees clearly understand that both data and software tools created with NIH funds should be included as part of their data management and sharing plan. This definition would be limited to artifacts created with NIH funds, and omit proprietary software tools used to conduct research, such as a stat package. We recommend a definition such as:

Scientific Software Artifacts: Software, code, analytic programs, and other knowledge artifacts developed to conduct research or resulting from the conduct of research.

2. Covered Data

AMIA Comment: We see the need to define two additional terms so that the DMSP can address a number of questions that arise throughout our deliberations. Specifically, grantees need to have a

clear understanding of which Scientific Data are covered by the Policy and for what period of time those data are covered. These definitions do not need to establish a policy for these questions; rather, these concepts should facilitate conversations to answer those questions.

There is a distinction between data generated by and for research, and data that is used in research. We see a need to define what scientific data is covered under the DMSP and what data is not. For example, clinical trials routinely rely on data that has been generated during the course of clinical care and collected as part of research participants' electronic health record (EHRs). This data may be included in the study data set and used as part of an analysis. Such data was not specifically generated for the trial and the tests or other work involved in generating them were not paid for by the trial. Is such data covered by the policy or not?

As another example, we note that a number of large databases are currently used and made available for epidemiological research or data mining projects based entirely on real-world evidence. These data are generated and paid for in the course of routine clinical care and are maintained under private funding. If an NIH funded analytic project is based on the use of such data, can that data now be required to be made available more generally to the public? If so, this could represent a disincentive to the private organization to make such data available for research and might have the paradoxical effect of making less data available for research or making whole classes of data unavailable for research.

AMIA Recommendation: AMIA recommends the draft DMSP includes a definition for "Covered Data," so that grantees clearly understand which data must be included as part of their data management and sharing plan. We recommend a definition such as:

Covered Data: Those newly generated or derived Scientific Data used to conduct NIH-funded or -supported research and subject to this Policy. Such data may or may not be proprietary and subject to various access controls.

3. Covered Period

AMIA Comment: We also note a need to define the expected timeframe for which grantees must steward Scientific Data. While we have numerous questions, such as, does the transfer of data to an NIH-supported or endorsed repository complete the obligation of the grantee? Will there be funding available to grantees who steward their own Scientific Data associated with tracking and satisfying data use requests? Would there be some appeal process if the volume/complexity of requests exceeds what was anticipated or funded? Or could the grantee charge some reasonable administrative fee if total costs incurred exceed some threshold?

AMIA Recommendation: We recommend the NIH address these and other questions by incorporating a concept of "Covered Period." This term would facilitate greater understanding of the obligations of grantees

Covered Period: The period of time for which the Scientific Data is expected to be maintained by the grantee and for which it is to be made available to others.

II. Purpose

AMIA Comment: This section describes what the DMSP is, but only hints at why the NIH is proposing one and how it will interact with other NIH policies. This section should describe why a DMSP is necessary and what a DMSP will achieve.

AMIA Recommendation: The draft DMSP should bolster the Purpose section by adding language similar to the introductory language, beginning, “NIH has a longstanding commitment to making the results and accomplishments of the research that it funds and conducts available to the public. Increasing access to scientific data resulting from NIH funding or support offers many benefits and reflects NIH’s responsibility to maintain stewardship over taxpayer funds.” AMIA recommends the draft DMSP adds to this with the following:

Specifically, systematic management and sharing of scientific data and results enables researchers to more vigorously test the validity of research findings, strengthen analyses by combining data sets, access hard-to-generate data, and explore new frontiers. Data management and sharing also informs future research pathways, increases the return on investment of scientific research funding, and accelerates the translation of research results into knowledge, products, and procedures to improve health and prevent disease.

This Policy seeks to identify, adopt, and credit data management and sharing best practices, consistent with FAIR (Findable, Accessible, Interoperable, and Re-usable) data principles, so that the United States remains the leader in biomedical and life sciences research. This Policy establishes the requirements and responsibilities of researchers generating scientific data resulting from NIH-funded or -supported research and it will govern development and implementation of other NIH Policies related to the management and sharing of scientific data, such as the NIH Genomic Data Sharing Policy, the NIH Policy on the Dissemination of NIH-funded Clinical Trial Information, and the Intramural Research Program Human Data Sharing (HDS) Policy.

III. Scope and Requirements

AMIA Comment: We applaud the NIH for considering a comprehensive DMSP that would “apply to all intramural and extramural research, funding or supported in whole or in part by NIH, that results in scientific data, regardless of NIH funding or mechanism.” A pan-NIH DMSP will improve our national culture of data sharing, as well as facilitate the FAIR data principles.

We also support submission of a Data Management and Sharing Plan (Plan) as part of the funding/support application process and articulating if there are perceived barriers to sharing scientific data in this Plan. Finally, we greatly appreciate that these draft policy provisions state that “Reasonable costs associated with data management and sharing could be requested under the budget for the proposed project.”

AMIA Recommendation: We urge the NIH to proceed with the proposed DMSP scope, ensuring that the policy requirements are constructed in a way that both small and large awardees can comply. While we agree that it is important for all NIH research to be subject to this policy, regardless of funding or mechanism, the policy must maintain flexibility to accommodate individual ICs and individual project characteristics.

AMIA recommends the NIH draft this section as “**III. Scope**” and position the aspects of the current provisions related to “requirements” in the next section, “IV Requirements for Data Management and Sharing Plans.” The draft DMSP could expand on the rationale for its scope, similar to the Purpose section. We discuss issues related to IC-specific requirements and “reasonable costs,” for data management and sharing below.

IV. Requirements for Data Management and Sharing Plans

a. Plan Review and Evaluation

AMIA Comment: Establishing a flexible, yet consistent, and fair review and evaluation strategy will greatly improve the likelihood that this DMSP is successful. We note that the proposed policy envisions that review and evaluation would be the primary responsibility of the funding or supporting NIH IC, “which could be implemented in a variety of ways...” and that this section delineates how various funding mechanisms might differently approach the task of review and evaluation. We are generally supportive of this strategy as long as the DMSP provides direction for ICs to rationalize and harmonize their specific requirements.

As it relates to Extramural Grants, we are concerned that scoring in a binary way has contributed to our current shortcomings in quality data management and sharing. As stated previously, we view rigorous review and evaluation of Plans as a means to improve the FAIR-ness of data and encourage the NIH to treat these Plans as scorable elements of certain grant applications.

AMIA Recommendation: We recommend the DMSP establish parity between the rigor of Plan review/evaluation and amount of NIH funding support. We strongly recommend that the draft DMSP encourage ICs to factor the quality of the Plan into the overall impact score through the peer review process for those grants that are supported at high levels or support programmatic priorities. While we support negotiation, making Plans scorable will improve the use of best practices and the general management and sharing posture of applicants far more efficiently than an “acceptable or unacceptable,” evaluation schema. Rather than discouraging ICs from factoring Plan reviews/evaluations into the overall impact score, AMIA recommends ICs view quality Plans as essential to important research and design evaluation schemas to reflect this view.

Alternatively, the ICs could incentivize quality Plans by funding data management and sharing activities in an amount corresponding to the completeness of the Plan. For example, specific support of data managing and sharing activities might reflect the completeness of the plan, scored as "unsatisfactory" (0% of requested funds), "minimal" (25%), "adequate" (75%), "excellent" (100%). (Percentages for illustration only).

This recommendation notwithstanding, we do see value in considering a binary evaluation in limited circumstances, such as small grants to new investigators, or in cases where scientific data cannot be de-identified and shared.

b. Plan Elements

AMIA Comment and Recommendation: Given the extent of information expected as part of a Plan, we do not envision a 2-page limit will be sufficient in most circumstances. Rather than setting arbitrary page limits through the DMSP, we recommend the NIH leave length and depth of Plans to peer review and IC guidance.

We are generally supportive of the Plan Elements listed. However, we believe there is a need to include additional Elements so that applicants can describe their Data Management activities. We also recommend “Data Preservation and Access Timeline” be included as a sub-point of “Data Preservation and Access,” rather than a standalone Element. Below we offer comment and recommendation for each of the listed Elements.

i. Data Type

AMIA Comment and Recommendation: We recommend listing the find the term “rationale” in this section confusing. Given that the DMSP clearly articulates a rationale for scientific data preservation and sharing, we recommend this section simply state:

1. Data Type: Indicate the types and estimated amount of scientific data that will result from NIH-funded or -supported research and indicate ~~how the rationale for which~~ scientific data will be preserved and shared.

- 1.1. Amendments:** We recommend inserting “**expected**” following “scientific data” in 1.1 to reflect that the data actually collected may change slightly over time. The expectation should be that the Plan will be directionally correct and complete, but that it could be subject to amendment. Further, we recommend rewording the second sentence of 1.1 as follows:

Describe the data modality (e.g., imaging, genomic, mobile, **patient-reported**, and survey) and whether the scientific data will be individual, aggregated, or summarized, and **whether the data will be** ~~how raw or processed the data will be~~.

- 1.2. Amendments:** We recommend adding the word “**metadata**” to 1.2, and we encourage the NIH to reference this defined term as appropriate throughout the document.

Describe any other information that is anticipated to be shared along with the scientific data, such as relevant associated data, and any other information necessary to interpret the data (e.g., study protocols ~~and~~ data collection instruments, **and other metadata**).

ii. Related Tools, Software and/or Code

AMIA Comment and Recommendation: AMIA supports efforts to make tools, software and/or code available for use, if such artifacts were developed as the result of NIH funding. However, there is a fundamental difference between sharing data and sharing code or software, particularly if the code is considered proprietary, such as a purchased stat package. The intent of this policy should be twofold: (1) To improve replicability by ensuring transparency in how data were transformed and (2) encourage the sharing of related tools, software and/or code generated through NIH funding. The intent should not be to make researchers provide an analytic environment, open source or otherwise. The use of data and workflow diagrams, which graphically depicts at a high level the data sources, operations performed on the data, and the path taken by the data through information systems and operations may be useful.

While we support the use of alternative free or open source code, we do not view the DMSP as an appropriate vehicle to encourage such solutions. The effort to identify such tools could be significant and may require skills well beyond those of the investigator and requiring assistance from staff not included in any of the grant funding. We recommend the following changes to reflect these recommendations:

ii. Related Tools, Software and/or Code: Indicate what **tools**, software **and/or computer** code will be used to process or analyze the scientific data (~~the inclusion of scripts may be helpful~~), why the software/code was chosen, and whether it is free and open source. **Also indicate whether tools, software and/or code were developed to conduct NIH-supported research resulting in scientific data and if such artifacts are expected to be shared.** ~~If software/code that is not free and open source is needed to access or further analyze the scientific data, briefly describe why this particular software/code is needed. Describe whether there is an alternative free and open source software/code that may be used to further analyze the scientific data.~~ **The inclusion of scripts and the use of data and workflow diagrams, which graphically depicts at a high level the data sources, operations performed on the data, and the path taken by the data through information systems and operations may be useful.**

iii. Standards

AMIA Comment and Recommendation: AMIA appreciates the NIH pointing towards and encouraging use of established data standards, common data elements, and other publicly funded initiatives. We support leveraging this DMSP to encourage the use of existing data standards and common data elements (CDEs) to “facilitate broader and more effective use of scientific data and to advance research across studies.” We hope that, over time, researchers will coalesce around common standards when appropriate and that when common standards can be used they are used. This will only happen if Plans are critically peer reviewed by experts trained in the systematic collection, analysis, and application of data. We recommend the following changes to reflect these recommendations:

iii. Standards: Indicate what standards, if any, apply to the scientific data to be collected, including data formats, data identifiers, **data models**, definitions, **metadata** and other data documentation, including terms of use. NIH encourages the use of existing data standards, such as standards for

collecting and representing scientific data and information describing the scientific data. NIH encourages the use of common data elements (CDEs) to facilitate broader and more effective use of scientific data and to advance research across studies. For assistance in identifying NIH-supported CDEs, the NIH has established a Common Data Element Resource Portal. **For a list of established clinical data standards, please see the most recent Office of the National Coordinator for Health Information Technology Standards Advisory.**³ Where commonly accepted standards don't exist, the Plan should include description of these standards in this section.

iv. Data Preservation and Access

AMIA Comment and Recommendation: AMIA encourages the NIH to be more prescriptive in its expectations that Plans leverage NIH-supported data repositories.⁴ AMIA recommends the NIH incentivize the deposition of scientific data into NIH-supported data repositories by scoring or funding such Plans higher than Plans that do not use an NIH-supported or NIH-approved data repository (unless sufficient justification can be made) and by allowing Plans that leverage such repositories to forego this section of the DMSP. This may require the NIH to better understand the relative strengths and weaknesses of repositories currently/potentially supported by the NIH, but it will improve the likelihood of long-term data FAIRness. AMIA recommends the NIH develop a formal endorsement process to approve and list preferred repositories for scientific data and scientific software artifacts.

In addition, the NIH should use information gathered during the 2016 RFI on “Metrics to Assess Value of Biomedical Digital Repositories,” to inform policy development in this area. While AMIA acknowledged there “will be no ‘one-size fits all’ scorecard” in comments to this RFI, we provided several recommendations for the NIH to develop a rating schema for deposition repositories and knowledgebases.⁵

If Plans wish to rely on data repositories other than those supported or endorsed by NIH, we recommend the following aspects be articulated (we reference existing sub-element numbers below):

4.1 Amendments Data Deposition and Archiving: Indicate where scientific data will be archived to ensure its long-term preservation. If scientific data will be stored in an existing repository, provide the name and URL web address of the repository. If an existing repository will not be used, indicate why not and how scientific data preservation will be assured (e.g., in a newly created repository or by the investigator's organization).

4.2 Amendments Discoverability: Indicate how the scientific data will be made discoverable and whether a persistent unique identifier or other standard indexing tools will be used.

4.3 Amendments Security: Describe any provisions for maintaining the security and integrity of the scientific data (e.g., encryption and backups).

³ <https://www.healthit.gov/isa/>

⁴ https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html

⁵ AMIA Comments available at: <https://www.amia.org/sites/default/files/AMIA-Response-to-NIH-RFI-on-Metrics-to-Assess-Value-of-Biomedical-Digital-Repositories.pdf>

4.4 Amendments Plan Alternatives: Describe alternative plans for maintaining, preserving, and providing access to scientific data should the original Plan not be achieved.

4.5 Amendments Barriers: If perceived barriers to ~~sharing~~ **preserving and making accessible** scientific data exist (e.g., ~~sharing includes specific restrictions or sharing is not possible~~), ~~outline how scientific data will be managed and preserved and~~ include an explanation of the perceived barriers.

4.6 Amendments Other Considerations: Indicate whether additional considerations are needed to preserve and make accessible ~~implement~~ **the scientific data. Plan** (e.g., ~~prior permission to use a specific repository~~).

4.7 Amendments Biospecimens: Indicate whether scientific data generated from humans or human biospecimens will be available through unrestricted (made publicly available to anyone) or restricted access (made available after the requestor has received approval to use the requested scientific data for a particular project or projects). If the scientific data will be shared through a restricted access mechanism, describe the terms of **access** for the data.

New 4.8 Timeline: Provide information on the anticipated timeframes for scientific data storage and accessibility, and criteria for how decisions affecting scientific data storage and accessibility will be made throughout the course of the study.

New 4.9 Amendments: Secondary Use Timeline: Describe when the scientific data will be made available to secondary data users. This should be expressed in relation to some critical event, such as the publication of the major study findings, the end of data collection, or other similar activity.

v. Data Preservation and Access Timeline

AMIA Comment and Recommendation: AMIA recommends the DMSP merge Element 5 as subordinate points of Element 4 (see above Elements 4.8 and 4.9). We recommend that Element 5.2 be removed from the DMSP.

vi. Data Sharing Agreements, Licensing, and Intellectual Property

AMIA Comment and Recommendation: AMIA supports the expectation that scientific data will be broadly available, consistent with privacy, security, informed consent, and proprietary issues. We note that this information may be duplicative with information provided in prior Elements, such as barriers to preservation / access, and we encourage NIH to reduce sections that overlap in intent or required content.

6.1 Amendments Data Sharing Agreements: Describe any **existing** data sharing agreement(s), outlining the responsibilities of each party, as well as how scientific data can and cannot be used.

6.2 Amendments Licensing: Describe any ~~existing general~~ licensing terms, and any limitations on the scientific data use and reuse based on these terms. Describe whether the licensing is imposed by the applicant institution or whether it comes from any existing agreement(s).

6.3 Amendments Intellectual Property: If applicable, indicate how intellectual property, including invention or other proprietary rights, will be managed in a way to maximize sharing of scientific data. Include any information relevant to the intellectual property rights associated with the scientific data, such as whether the intellectual property stems from an existing agreement or is anticipated to arise from the proposed research project itself.

vii. Oversight of Data Management

AMIA Comment and Recommendation: AMIA recommends removal of this section, given that grantees already provide personnel information in other parts of the grant. If it remains in the draft DMSP, we recommend a focus on the role rather than the individual to describe data management oversight and execution of the Plan.

viii. Other Considerations

AMIA Comment and Recommendation: AMIA views the additional considerations as important context that could be used to

V. Compliance and Enforcement

AMIA Comment and Recommendation: AMIA generally supports the compliance section “During the Funding or Support Period,” and “Post-Funding or Support Period.” However, we note that data management is an ongoing process and that a management plan is updated, modified, and versioned. We anticipate that this part of the Plan could be part of the progress report statement. As for data sharing, we reiterate our recommendation that NIH develop a formal endorsement process of preferred databases and knowledgebases. These endorsed repositories would facilitate DMSP compliance and enforcement by having transparent terms and conditions and abide community consensus best practices. Researchers who use these NIH endorsed repositories would have a streamlined compliance process.

Submission #154**Date:** 12/10/2018**Name:** Juergen Klenk**Name of Organization:** Deloitte Consulting LLP**Type of Organization:** Other**Other Type of Organization:** Consulting Firm**Role:** Member of the Public**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

NIH Data Commons, Cancer Research, Bioinformatics and Data Science

I. The definition of Scientific Data

- The definition of scientific data should include references to FAIR (Wilkinson, Mark D., et al. "The FAIR Guiding Principles for scientific data management and stewardship." Scientific data 3 (2016)) and Open Science to promote a collaborative research environment open to researchers and the public

II. The requirements for Data Management and Sharing Plans

- Minimum requirements for FAIR principles and guidelines should be proposed by NIH for inclusion in every data management and sharing plan.
- Data management and sharing plans should include information on the rights to the data (i.e. data use restrictions or limitations as well as database access during data collections), including plans for short and long-term storage and preservation of the data.
- Preservation of data should ideally be in the cloud, but if not, a phased approach where silos of data are transitioned into the cloud should be implemented based on a reasonable timeline based on available technologies.
- Strongly encourage the use of standards including specific recommendations for standards.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible

phasing could relate to needed improvements in data infrastructure, resources, and standards

- NIH should provide guidance on Data Management and Sharing best practices (i.e. minimum requirements and technical approaches) before enforcing a policy. For example, NIH should provide a specific location and/or mechanism in which data can be stored if researchers are unable or unwilling to use their own local storage.
- FAIR guidelines and principles must be better defined to provide a clear representation of the expectations of researchers, including minimal requirements.
- Data sharing and reuse should be incentivized by providing a reward. Incentives can include appropriate credit via citations, co-authorships, or favored review of new grant applications. Publishers should not be able to withhold or delay the sharing of data.
- Penalties for misuse of shared data must be addressed by NIH and made enforceable. Penalties could include termination of grant, legal steps, etc. Penalties should be severe enough to deter intentional misuse of data.

Submission #155

Date: 12/10/2018

Name: Pamela A. Webb | Lisa Johnston

Name of Organization: University of Minnesota

Type of Organization: University

Role: Institutional Official

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

We are a comprehensive research university so all areas are important to us.

I. The definition of Scientific Data

Please see attached letter

II. The requirements for Data Management and Sharing Plans

Please see attached letter

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

Please see attached letter

Attachment:

December 10, 2018

Carrie D. Wolinetz, Ph.D.
Office of Science Policy
National Institutes of Health
6705 Rockledge Drive, Suite 750
Rockville, MD 20892

Subject: **Response to NOT-OD-19-014: Request for Information (RFI) on Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or supported Research**

Dr. Wolinetz:

The University of Minnesota writes in response to the Request for Information listed above, published October 10, 2018. As a public land grant institution, we strongly support federal agency policies ensuring public access to scientific research data. Concurrently, we support a thoughtful approach to balancing the need for data access with the administrative burden and cost associated with data management planning, curating, and storing data.

Any policy implemented by the NIH would have direct implications to the researchers we support. In fiscal year 2017, the University of Minnesota received 244 million dollars in funding from the NIH, accounting for 55.6% of our federal research funding.

Specifically, with input gathered from key research committees on campus as well as individual faculty, we would like to respond to the following proposed policy across several major themes:

I. The Definition of Scientific Data

- 1. We recommend that NIH adhere to the federal definition of data** in 2 CFR 200.315, which defines research data as:

{3} Research data means the recorded factual material commonly accepted in the scientific community as necessary to validate research findings, but not any of the following: preliminary analyses, drafts of scientific papers, plans for future research, peer reviews, or communications with colleagues. This "recorded" material excludes physical objects. Research data also do not include (i) Trade secrets, commercial information, materials necessary to be held confidential by a researcher until they are published, or similar information which is projected under law; and (ii)

Personnel and medical information and similar information the disclosure of which would constitute a clearly unwarranted invasion of personal privacy, such as information that could be used to identify a particular person in a research study."

By adhering to the national definition, NIH will promote commonality and consistency with other federal agencies, and will facilitate implementation and understanding by researchers, data repositories, journals, and the public alike.

In addition, the federal definition offers a guidepost to what should not be included in the definition, including most importantly data that is preliminary, incidental to research findings, or involving certain objects impractical to share, such as physical specimens or laboratory notebooks. The definition also sends a clear message that HIPAA-controlled data and other data that is subject to privacy laws would be excluded or handled in a way that protects confidentiality in accordance with existing laws. Such protections are a critical element to any data sharing strategy.

2. While the reproducibility issue is of great importance, we note that the scope of that issue goes well beyond data. **Any conversations about expanding the scope of research data to include "necessary to validate and replicate (emphasis added) research findings" should be taken only after discussion in context of all of the components necessary for successful study replication.** If the definition of research data needs to be amended, it should be considered through the change management process for alterations to the Uniform Guidance.
3. **We also recommend removing the expectations about digitizing scientific data from the definitions section** (e.g., *"NIH expects that reasonable efforts should be made to digitize all scientific data."*) Instead, asking a researcher to specify what research outcome data is expected to be suitable for deposit into a digital data repository may be a better approach. The cost and effort involved with digitizing an overly broad definition of scientific data may inadvertently subvert the purpose of having public and scientific access to meaningful research data. In addition, the oversight cost for some types of data (including patient data) to ensure that privacy obligations are fully met and that inadvertent disclosure does not occur may be very labor-intensive. Some data (including information in hard copy lab notebooks) may not be worth the cost, curation, and storage efforts involved in being digitized.
4. **Conversely, other more rare information (such as 3d scans of rare physical specimens) that are excluded from the definition may well be worthy of being digitized and funding should be identified for this purpose when access would be deemed valuable.** Alternative approaches, such as NHLBI's BioINCC initiative that makes available "recorded" material about physical objections in a sharable database or inventory of laboratory sample holdings, should also be explored as a cost-effective option.
5. **Commonly adopted and understood standards for metadata and other coding schemes**

inherent in data collection, organization and management are paramount to the long-term success of this effort. NIH should continue to invest in development, evaluation and adoption of broadly understood and accepted metadata schemas (such as the NIH common data elements: <https://www.nlm.nih.gov/cde/>) as a preface to requiring long-term data storage.

II. Requirements for Data Management and Sharing Plans

1. **The proposed requirements for the data management plan in the proposal are extensive. We are concerned about the level of detail requested at time of proposal submission.** We note that approximately 18% of NIH proposals are funded, and believe it is wasteful of researcher time to require detailed information to be embedded in the proposal when most proposals are not funded. As you know, based on the most recent national FDP Faculty Workload Survey, researchers are expending 44% of their available research time on administrative tasks - and that is before the addition of this new significant requirement. **We recommend that the data management plan be required at a later point in the process.** If possible we recommend that the DMP be furnished at Just-in-Time or, if NIH feels that the plan is essential in order for reviewers to have confidence in the investigator's commitment and ability to share data, then the Plan should be given additional importance in the review process (see item #2 below). If a Plan is required at time of original proposal, we recommend that requirements be bifurcated into only those elements deemed critical for the proposal review (included in the proposal and the proposal budget) and then supplemented either at JIT or at time of RPPR to be expanded/refined. It is critical to mutually develop a mechanism that honors the importance of data sharing while minimizing its burden footprint. Faculty should be directly involved in the solutions developed.
2. Within the proposed plan review and evaluation criteria, the data management plan is currently proposed as an *Additional Review Consideration* for extramural grants. As an *Additional Review Consideration*, the DMP would not be individually scored nor would it influence the overall score, although there is an expectation that compliance with the plan "*would be integrated into terms and conditions as appropriate*" and that NIH staff would engage with potential awardees to modify the plan as appropriate prior to award. Given the extent of the proposed data management plan requirements and enforcement expectations, it would appear that the effort and implications of the data management plan are not aligned in their value within the review process. **While we strongly prefer the option listed in Item 1 above, if NIH continues to believe that the DMP is essential at time of full proposal review, we recommend positioning the data management plan as an *Additional Review Criteria*. This would allow the Plan to not be scored individually but still be considered in the overall impact score.**
3. We feel that restricted access should not be the responsibility of the individual researcher. Not only is this administratively burdensome, but it also introduces a dependence on the researcher (and their current contact information) that undermines the goal of long-term data accessibility..

NIH should recommend restricted access repositories that provide this level of control for sharing their data.

4. The "Compliance" section sends a strong message. However, it is our experience that data management and sharing is difficult to fully anticipate in detail. **Our researchers require flexibility to update and change their DMP as the project progresses.** This is especially relevant as data management plans embedded in proposals may not be used (or fully used) until several years later, and repositories and data standards can reasonably be expected to evolve in the meantime. Complications may arise and a DMP should not simply be followed in order to meet requirements NIH mandates to measure compliance. **We recommend that the DMP could be revised annually or as needed (with explanation). Such revisions could be included in the project's annual report.** We note that this could be an excellent opportunity to increase dialogue between researchers and program officials, and to allow the natural sharing of advancements of data sharing mechanisms in a given field.

5. We appreciate the "Oversight of Data Management" section as an explicit component of the data management plan. This highlights the fact that the **management of research data is an active process which requires the long-term investment of resources. To the extent that these resources must be paid by grant budgets, NIH should help develop and widely publicize allowable, reasonable and allocable charging guidelines, and should incorporate the need for such costs into its own budgeting and planning.** Specifically, we note two challenges: (1) many researchers already feel that funding is very limited to support their science, and new "draws" on existing pots of available funding would have the inevitable consequence of reducing the amount available for the direct costs of science; and (2) some of the costs associated with data storage and sharing cannot reasonably be incurred with the period of the grant (or its closeout period). **A new charging mechanism is therefore needed to allow for costs to be either separately funded or, at minimum, set aside to cover these costs. This is a less significant issue if federally supported, non-profit repositories that do not require deposit fees can be made available.** We note that making available non-profit repositories will also facilitate implementation of national data deposit standards and record retention expectations, as well as public access to data. In addition, sustainable national data repositories can be expected to reduce confusion associated with data deposits from multi-site projects, and reduces issues associated with local storage failures (lost or corrupted computer, servers, or other storage facilities) or an inability of a university or other grantee to continue to support its data storage environment.

6. **Of particular importance is the degree of variation in the myriad types and sizes of data that require curation, storage and access.** We have attached a document from Jakub Tolar, Vice President for Clinical Affairs/Dean of the University of Minnesota Medical School that does an excellent job of articulating some of the specific issues associated with the different populations of data such as derived from clinical studies, basic sciences, informatics and "dry lab" data analysis, and large proprietary datasets. We ask that any data management strategy be

significantly flexible to address the issues raised by Dean Tolar.

III. Timing for NIH to implement

1. **The opinions and needs of Researchers and Library professionals are critical to this issue, and must be sought and understood at a detailed level to properly plan for not only how data can best be accessed, but also appraised, selected, retained, and deaccessioned.** We recommend that NIH proceed with sufficient lead time and engagement with the consumers of the data being produced in order to "get it right". Imposition of additional administrative and cost burden on NIH and on grantees should not be undertaken unless there is a high confidence that benefit can be demonstrated to exceed the cost.

IV. Additional Comments

1. In the "Purpose" section, the RFI states that "scientific data... should be managed, preserved, and made accessible in a *timely* manner for *appropriate* use by the research community and the broader public" [italics added]. **We recommend defining the words "timely" and "appropriate" more specifically and including examples of what is (and what is not) timely or appropriate.**
 - a. In terms of timeliness, NIH could recommend sharing data X years after the close of the grant, x months after the publication of a paper that uses the data, or x weeks after the data sharing deadline stated in the DMP.
 - b. If principal investigators or their delegates are responsible for determining what constitutes appropriate use, this should be clearly stated in the Licenses and Terms of Use section. For example, would PIs be able to place commercial-limiting licenses on their data? Also note that depending on the repository selected, the terms of use may not always be flexible (e.g., blanket repository-wide statement). The researcher should consult with the repository to help them consider what terms, licences are appropriate/available for their data.
2. **In the Preservation section, please link to clear preservation guidelines.** There do not appear to be well-established protocols listed in the NIH Strategic Plan for Data Science. Instead consider more explicit best practices captured in a dynamic way that will stay up to date with the evolving community of practice. One suggestion is the LOC Digital Preservation web site.
3. Preservation is not an inherent attribute of a repository. Specific actions must be taken to preserve digital files and not all repositories do so. Researchers should include a link to the preservation policy of a repository to provide evidence of such actions. A CoreTrustSeal, TRAC, or other trusted digital repository certification could also be evidence of this.
4. Finally, we noted a number of strong elements of the proposed plan and believe these should be highlighted for adoption by other funders (please see Appendix A) .

Finally, we thank you for giving us this opportunity to provide input on this critical topic to the NIH. We look forward to continuing the dialogue on this topic.

Sincerely,



Pamela A. Webb
Associate Vice President for Research



Lisa Johnston
Director, Data Repository for the University of
Minnesota (DRUM), University Libraries

Enclosure: Memorandum dated November 30, 2018 from Dean Jakub Tolar, MD, Ph.D

cc with enclosure: Chris Cramer,
Wendy Lougee
Jakub Tolar

Appendix A: Elements in the proposed NIH Data Management and Sharing Policy We Liked

- a. S4.4 - We like the suggestion that an alternative preservation plan be included should the original Plan not be achieved.
- b. S1.2 - We agree that study protocols and data collection instruments should be considered metadata and shared along with the data.
- c. S2.0 - We like that suggestion for justification when choosing a non open source software and highlighting where open source alternatives are available.
- d. S4.2 - We like that a "persistent unique identifier" is used, rather than DOI specifically (brand name).
- e. 57 - Oversight. We appreciate this section being included and other federal agency policies should follow your lead.

[https://datascience.nih.gov/sites/default/files/NIH Strategic Plan for Data Science Final 508.pdf](https://datascience.nih.gov/sites/default/files/NIH%20Strategic%20Plan%20for%20Data%20Science%20Final%20508.pdf)

See section Objective 3-3 | Improve Discovery and Cataloging Resources (page 19 of the PDF)

UNIVERSITY OF MINNESOTA

Twill Cities C(1111p11s

*Office of the De""
Medical School*

*C'6/i7 A!C1yo. t.iMC 2JJ
4J0 Velmvarc* .Ntrceer \ I-:
M1111eclpl1)i, MN 55455
Off1Cil. 612-(,264949
Nn· 6/2-6.20 4911*

November 30, 2018

Christopher Cramer, Vice President for Research
Office of the Vice President for Research
420 Johnston Hall 101 Pleasant St SE
Minneapolis, MN 55455

Pamela Webb, Associate Vice President
Office of the Vice President for Research
420 Johnston Hall 101 Pleasant St SE
Minneapolis, MN 55455

RE: Medical School Comments on "RFI on Proposed Provisions of a Draft Data Sharing and Management Policy"

Dear Chris and Pamela:

On behalf of the Medical School, please see below our comments on proposed changes to the Data Sharing policy of the NIH. We ask that these be included in the University response to this policy.

We asked several senior faculty for feedback. In general, our faculty responded that the NIH needs to be more pro-active in recommending best practices for data sharing, should provide details on how to fund this activity and information about how this will work for multi-site projects or projects with multiple PI's. There is concern that this request could distract from our central core mission and that the effort that will be required to achieve what is being mandated will take already scarce resources away from the actual research.

The questions to be raised are, how will this be resourced; how often does this occur where there is a request for data sharing that is not covered by the current policy; how long is data storage required; how will data storage be subsidized given the greater demands and length of storage time; and how will requests be paid for once the grant is closed and there are no longer funds to support effort on that proposal as this happens often at our institution; if the computer, server, or other storage facilities are lost, corrupted, or otherwise results in irretrievable data, how will such affect publications or grants, including those already under review, in press or published. There should be standardized templates and procedures established for each type of data that might be requested to be shared. What federally funded program will be responsible for oversight? Will it be the CTSA organizations at each site (and what about sites and organizations that do not have a CTSA)? Will it be the responsibility of the institution to oversee this policy? How will sharing without restriction be protected to avoid precluding meaningful research including follow-up

studies or revelation of novel insights, secondary or derivative analyses of the primary studies and future comparisons with other approaches?

In regards to Clinical Studies: how will this affect the consent and the consenting process; how much time will this add to the process and affect the length of the consent; how will this change procedures for consenting vulnerable or unique populations where there might be cultural or legal constraints on what can be done with the data (e.g., prisoners); how will clinical data be de-identified, who will do this and will there be a standardized process and format.

In regards to Basic Sciences: how will this policy be applied to images, captured and or manipulated by core resources, where the files might be exceedingly large and the processing techniques to render the final image are proprietary; will this pertain to only the final data set used for analyses presented in a manuscript or a grant application; will this pertain to data presented at scientific or other meetings, especially data presented that might not necessarily be published (negative data or multiple rejections and the decision is made not to pursue publication). Flow cytometry data is often stored on servers maintained by the PI. The level of organization and life time of the information varies from lab to lab. Is it the intent of this policy that data related to publications be somehow partitioned into publicly accessible sectors on these servers, or some common server for the institution? Will there be an expectation that uniform file naming system tied to a directory will exist that someone else could use to know which experimental condition applies to each file? Who will be responsible for oversight of that process and who will pay for it?

In regards to Informatics and Data Analysis, we recommend excluding all references to methods and software for analysis of data from any data management plan. Analyses and research design are not data per se and will have been addressed in the research plan and software sharing section of proposals. There should be recognition that a significant portion of the software used for informatics analyses are proprietary and cannot be shared. We would recommend that data management software be appropriately referenced and investigators seeking access to it can contact the person/organization holding rights to the software. Attention should be given to the fact that many labs use an alternative to a traditional lab notebook. Dry labs often use code books + code documentation + analysis design revisions + intermediate analysis reports. Much of this is proprietary. Does this policy cover "intermediate or preliminary" results? This is particularly important as intermediate analyses are often used to establish IP claims and to refute challenges to integrity of the research conducted. Intermediate and preliminary analyses should be excluded from this policy. There should be a hard deadline for any commitment to share data. For example x years after close of a grant, x months after publication of a paper that uses the data, or x weeks after data sharing deadline stated in the DMP.

Sharing large proprietary data is often not possible and should be excluded. Examples of this include: learning health systems studies where broad data mining of the EHR is used to generate hypotheses or model the totality of such data or large portions of it. In the US such data are in the majority of cases the property of the health providers and any data sharing provision will create future resistance *in* obtaining the data for research; learning health systems data where proprietary economic or other data is used; proprietary datasets in data sharing networks where participation in the network (and thus creation of network effect) is predicated on sharing data by

participants with the network; proprietary datasets provided by commercial aggregators using proprietary collection and encoding methods, etc. If this policy is implemented it would make sense to have a unique research set identifier. Exceptions or special provision will be required when the data set to be shared is quite large. For example, consider a study comparing, evaluating and benchmarking *50* algorithms times 100 dataset-tasks on Big Datasets, times 1000s of model selection setups and protocols times 100s of bootstrap or cross validation repeats, and its algorithmic model fit generating 1000s of intermediate data representations. This type of study is rare but has happened and it is currently impossible to store and share such data with ordinary resources and sharing mechanisms. As we move to highly protocolized and rigorous Big Data analytic BPs we will be seeing more of those.

Thank you for including our perspective in this response. This is an important issue for our faculty.

Sincerely,

To

Jakub Tolar, MD, PhD
Dean, Medical School
Vice President for Clinical Affairs

Submission #156**Date:** 12/10/2018**Name:** Heidi Rehm, PhD, FACMG**Name of Organization:** Clinical Genome Resource**Type of Organization:** Nonprofit Research Organization**Role:** Scientific Researcher**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Clinical Genomics

I. The definition of Scientific Data

I am submitting feedback on the “Request for Information on Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research” on behalf of the Clinical Genome Resource (ClinGen) Steering Committee. ClinGen is an NHGRI-funded initiative dedicated to building a central resource defining the clinical relevance of genes and variants for use in precision medicine and research. ClinGen is increasing the community’s knowledge of the relationship between genomics and human health by supporting the deposition of genomic and health knowledge and data into the public domain, specifically the ClinVar database hosted by the National Center for Biotechnology Information (NCBI) and on our website www.clinicalgenome.org. As data sharing is a central aim of ClinGen, we are very interested in the proposed provisions for a Data Management and Sharing Policy for NIH Funded or Supported Research.

We encourage you to specifically expand the definition of Scientific Data to include interpreted data, when appropriate. In the field of human genetics, an investigator’s interpretation of a particular genomic variant, and the supporting evidence used to arrive at that interpretation, is a key aspect of data analysis. For example, in a genomics research study in which interpretation is an aim of the funded work, in addition to sharing a list of genomic variants identified in the research subjects, an investigator would be expected to also share structured data annotations for each variant, such as the clinical significance classification and condition, inheritance pattern and evidence on which the classification was based (e.g. published evidence, summary of one or more case observations, information on additional variants observed in individual patients as it relates to the assertion and/or experimental evidence). In this context, interpreted data would meet the Scientific Data description set forth in the provision “necessary to validate and

replicate research findings.” In addition, ClinGen has developed and published processes for structured assessment of dosage sensitivity of genes and genomic regions (Riggs et al., 2012 PMC5008023), clinical actionability of genetic disorders (Hunter et al., 2016 PMCID: PMC5085884) and clinical validity gene-disease associations (Strande et al., 2017 PMC5473734). The results of these assessments are another example of interpreted data that should be included as Scientific Data. Thus, we strongly encourage you to consider interpreted data and structured annotations of data as part of your definition of Scientific Data and to state this in any future policy. While we are sharing our perspective from the field of genomic data sharing, we assume this would be applicable to other fields as well.

II. The requirements for Data Management and Sharing Plans

Expanding the definition of Scientific Data to include interpreted data would lead to revising the Data Sharing plans to include appropriate repositories for interpreted data. In the context of genomic data, the NIH Genomic Data Sharing Policy currently requires all applicable studies to register and submit data to the controlled access NCBI database of Genotypes and Phenotypes (dbGaP, <https://www.ncbi.nlm.nih.gov/gap>). By including interpreted data in the Scientific Data Sharing Plan, a genomics research study in which interpretation is an aim of the funded work should also be expected to submit interpreted variant data and supporting evidence to NCBI’s ClinVar database, a publicly available submission-driven archive of genomic variation and its relationship to human health. Highlighting the importance of publicly available interpreted data, the FDA has formally recognized ClinGen’s clinical significance assertions (which are accessible to the community via the ClinVar database) as a source of valid scientific evidence that can be used to support clinical validity for genetic and genomic-based in vitro diagnostics (<https://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm627555.htm>).

Attachment:

Heidi L. Rehm, PhD, FACMG

The Broad Institute of MIT and Harvard
105 Broadway
Cambridge, MA 02142
Tel: 617-714-7939 (Tue/Thu/Fri)
hrehm@broadinstitute.org

Heidi L. Rehm, PhD, FACMG

Center for Genomic Medicine
Massachusetts General Hospital
Simches Research Bldg, CPZN-5-821A
185 Cambridge Street
Boston, MA 02114
Tel: 617-643-3217 (Mon/Wed)
hrehm@mgh.harvard.edu

December 10, 2018

To Whom It May Concern:

I am submitting feedback on the “Request for Information on Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research” on behalf of the Clinical Genome Resource (ClinGen) Steering Committee. ClinGen is an NHGRI-funded initiative dedicated to building a central resource defining the clinical relevance of genes and variants for use in precision medicine and research. ClinGen is increasing the community’s knowledge of the relationship between genomics and human health by supporting the deposition of genomic and health knowledge and data into the public domain, specifically the ClinVar database hosted by the National Center for Biotechnology Information (NCBI) and on our website www.clinicalgenome.org. As data sharing is a central aim of ClinGen, we are very interested in the proposed provisions for a Data Management and Sharing Policy for NIH Funded or Supported Research.

We encourage you to specifically expand the definition of Scientific Data to include interpreted data, when appropriate. In the field of human genetics, an investigator’s interpretation of a particular genomic variant, and the supporting evidence used to arrive at that interpretation, is a key aspect of data analysis. For example, in a genomics research study in which interpretation is an aim of the funded work, in addition to sharing a list of genomic variants identified in the research subjects, an investigator would be expected to also share structured data annotations for each variant, such as the clinical significance classification and condition, inheritance pattern and evidence on which the classification was based (e.g. published evidence, summary of one or more case observations, information on additional variants observed in individual patients as it relates to the assertion and/or experimental evidence). In this context, interpreted data would meet the Scientific Data description set forth in the provision “necessary to validate and replicate research findings.” In addition, ClinGen has developed and published processes for structured assessment of dosage sensitivity of genes and genomic regions (Riggs et al., 2012 PMC5008023), clinical actionability of genetic disorders (Hunter et al., 2016 PMCID: PMC5085884) and clinical validity gene-disease associations (Strande et al., 2017 PMC5473734). The results of these assessments are another example of interpreted data that should be included as Scientific Data. Thus, we strongly encourage you to consider interpreted data and structured annotations of data as part of your definition of Scientific Data and to state this in any future policy. While we are sharing our perspective from the field of genomic data sharing, we assume this would be applicable to other fields as well.

Expanding the definition of Scientific Data to include interpreted data would lead to revising the Data Sharing plans to include appropriate repositories for interpreted data. In the context of genomic data, the NIH Genomic Data Sharing Policy currently requires all applicable studies to register and submit data to the controlled access NCBI database of Genotypes and Phenotypes (dbGaP, <https://www.ncbi.nlm.nih.gov/gap>). By including interpreted data in the Scientific Data Sharing Plan, a

genomics research study in which interpretation is an aim of the funded work should also be expected to submit interpreted variant data and supporting evidence to NCBI's ClinVar database, a publicly available submission-driven archive of genomic variation and its relationship to human health. Highlighting the importance of publicly available interpreted data, the FDA has formally recognized ClinGen's clinical significance assertions (which are accessible to the community via the ClinVar database) as a source of valid scientific evidence that can be used to support clinical validity for genetic and genomic-based in vitro diagnostics (<https://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm627555.htm>).

Thank you for your consideration.

Sincerely,

A handwritten signature in cursive script that reads "Heidi L. Rehm".

Heidi L. Rehm, Ph.D., FACMG
Institute Member and Genomics Platform Medical Director, The Broad Institute of MIT and Harvard
Chief Genomics Officer, Department of Medicine, Massachusetts General Hospital
Professor of Pathology, Harvard Medical School

On behalf of the ClinGen Steering Committee
<https://www.clinicalgenome.org/about/clingen-leadership/>

Submission #157**Date:** 12/10/2018**Name:** James Reecy**Name of Organization:** Iowa State University**Type of Organization:** University**Role:** Institutional Official**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Iowa State University has a growing NIH funded research portfolio across a broad range of species, e.g. Arabidopsis, dogs, human, nematodes, swine, mice and zebrafish. Our research portfolio has focal areas of interest in: anti-microbial resistance; elucidation of basic molecular mechanisms; gene editing; mechanisms of infectious disease; neuroscience; social behavior; vaccines; and virology.

I. The definition of Scientific Data

One potential improvement could be to add a “machine-readable and digitally accessible” statement like that include in the Department of Energy’s Data Management Plan requirements (<https://science.energy.gov/funding-opportunities/digital-data-management/#Requirements>). This may directly address some of the “Scientific data do not include...” items in the proposed scientific data definition. As is correctly stated in the proposed definition, there is often confusion over if a chart or summary table is “scientific data” under the current definition.

A second area to potentially address would be to include “computer code/algorithms” developed during the conduction of research as potential “factual material” to validate and replicate research findings. Disclosure/sharing of this information is critical for the evaluation of the rigor and reproducibility of the research. Furthermore, it would allow the research community to address new questions in a more-timely manner as well as promote a more inclusive dissemination of knowledge from funded research projects.

II. The requirements for Data Management and Sharing Plans

The proposed “Requirements for Data Management and Sharing Plans” are very extensive and well laid out. The National Institute for Health should be commended for its efforts in this area. However, limiting this section of grant applications to only two pages may be problematic. It

may be very tough for researchers with complex proposals to appropriately address the extensive nature of the requested information.

Ideally Data Management Plans should be updated regularly over the entire life cycle of a project. We suggest adding this point to the proposed requirements. Furthermore, please consider reporting on data sharing progress in annual and final reports. The proposed requirements can be ambiguously interpreted in this regard.

The phrasing of “meets community-based standards” is open to interpretation. Many researchers do not have the required expertise to know what “community-based standards” are. It would be helpful if NIH provides minimal data repository requirements similar to the Department of Transportation (<https://ntl.bts.gov/public-access/guidelines-evaluating-repositories>). Guidance like this may help to alleviate potential confusion. Furthermore, NIH may want to consider including a statement about the need to share Metadata, which includes study design, methods and code, in addition to the sharing of “scientific data”. This could potentially be added to the Data Preservation and Access section. While some communities have developed robust minimal standards, see MIBBI guidelines publication for additional information (Taylor et al., 2008), many communities have not yet developed standards. Furthermore, new technologies will continue to be developed, which will result in the need for new standards.

Taylor et al., 2008. Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nature Biotechnology* 26:889-896.
<https://www.nature.com/articles/nbt.1411>

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

Some NIH research communities, e.g. genomics and functional genomics, are well positioned to be in full compliance with these proposed guidelines today. Those communities have had a long tradition of data management and data sharing. However, other communities, e.g. economics or sociology, are not well positioned to share data. It is important that the entire NIH research portfolio can meet the proposed guidelines.

The recent AAU APLU Public Access Working Group Report (<https://www.aau.edu/key-issues/aau-aplu-public-access-working-group-report-and-recommendations>) clearly articulated the need to share “scientific data”. However, Universities in general are not well positioned to advance public access to data in a viable and sustainable way. The recent APLU/AAU Workshop on Accelerating Public Access to Research Data (<https://www.aplu.org/projects-and->

initiatives/research-science-and-technology/public-access/index.html) brought representatives from 30 Universities together to discuss the following goals:

- Accelerate the progress with which research institutions are developing and implementing institutional plans to provide public access to data;
- Foster cross-institutional collaboration that yields alternative models to publicly sharing data, reduces total effort of developing public access to data across the system of research universities, and builds consensus on key system elements that foster effective storage and sharing of data in ways that are findable, accessible, interoperable and reusable (FAIR); and
- Foster discussions among various data access stakeholders from universities and elsewhere to facilitate common, streamlined, and efficient approaches that help facilitate, support, and encourage data access.

While Universities are making great strides in efforts to provide public access to “scientific data”, there is clearly room for improving the system. Therefore, we would recommend that NIH work directly with the its research communities to develop plans/time lines for phasing in these proposed guidelines to help insure that community can be within compliance with them without causing undo stress on the research community.

Submission #158

Date: 12/10/2018

Name: Everett R. Rhoades MD, FACP (ret.)

Name of Organization: Private citizen Member Kiowa Tribe of Oklahoma

Type of Organization: Not Applicable

Role: Member of the Public

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

I don't know what you want with these credentialing questions. I'm just a person retired from several professional careers (Director of IHS, Professor of Medicine and Microbiology, Senior Consultant to the Center for American Indian Health Research of the University of Oklahoma College of Public Health, Member Kiowa Business Committee (the Tribe's governing body), former Adjunct Professor of International Health at Hopkins School of Public Health, President of the Board of Directors of the Oklahoma City Indian Clinic, many terms as a member of Councils of two or three NIH Institutes, etc. etc. etc. I don't know what it is you are looking for. I now have no special official posts that is likely to impress you. Let me just say no one comes close to the depth, breadth, and knowledge of American Indian and Alaska Native Health that I have. You can look it up.

I am not a robot and I resent being asked if I am. Are you?

Attachment:

RESPONSES TO PROPOSED NEW RULE MAKING RE: DATA SHARING

Everett R. Rhoades MD, FACP (ret.)
1808 Dorchester Drive
Nichols Hills, OK 73120-4706
December 10, 2018

Francis Collins, MD, PhD.
Director, National Institutes of Health
9000 Rockville Pike
Bethesda, Maryland 20892

Dear Dr. Collins,

I am responding to the latest proposed rule of the National Institutes of Health, dated October 10, 2018, regarding the dissemination of scientific data obtained from human individuals. The National Institutes of Health (NIH) has been engaged for more than a decade in trying to assure the widest possible dissemination of such data while protecting certain rights reserved to individuals who are participants in scientific and other investigations. The present proposed rulemaking appears to be an effort by the NIH to deal with conflicts arising from “Big Data” and human privacy, dignity, and humanness.

Privacy protections are of special concern for certain subpopulations considered to be vulnerable to certain harms. One such subpopulation is American Indians and Alaska Natives. This response is primarily concerned with the special situations in which these subpopulations find themselves. While NIH is to be applauded for its efforts regarding protection of human research participants and has made some progress in dealing with the sovereign Indian Nations, efforts have not yet resulted in a satisfactory outcome. The present response is intended to be of service to NIH and to the various American Indian populations.

DEFINITIONS

For purposes of the present discussion, the terms, “Sovereign Indian Nations”, “Sovereign American Indian Tribes” and “Tribes” are synonymous and are used interchangeably. For emphasis the term “tribal” is often capitalized as “Tribal”. “Big Data” as a concept is sufficiently embedded in general consciousness that it needs no further definition herein. However the term continues to evolve and attention to what the term “Big Data” refers to is necessary in all proposed rulemaking.

THE DOCTRINE OF PRIVACY

- The *right to privacy* is the foundation for protection of human research participants.
- *Privacy is absolute*. One either has privacy or one does not.
- *Privacy is an inherent right*, it is not delegated by any agency.

- Privacy rights reside upon a foundation of *informed consent* by the human participant.
- Albeit unintended, the operation of Big Data is *an assault on privacy*.

PRIVACY, DEHUMANIZATION and INFORMED CONSENT

- Abridgement of the doctrine of privacy is *dehumanizing*.
- Participant consents are obsolete in the era of *Big Data*. In fact the situation is backward. *In the Big Data era, the party yielding consent should not be the participant; rather it should be the investigator and the funding agency that consent to participant requirements*. There is no way that yielding consent can do anything but put the individual in a subordinate position. In an enlightened society, this is no longer an acceptable paradigm.
- The effort of NIH to obtain the *broadest possible consent* is disturbing. *Any consent for some unknown future investigator pursuing some unspecified end, cannot possibly be informed*. Therefore, such a consent is *unethical on its face*.
- In the age of Big Data, tensions are inherent between leaning toward dehumanization and leaning toward protecting human freedom, dignity, and privacy. This conflict affects certain populations as well as individuals.
- Errors must always be made in the tilt toward human privacy, dignity, and humanness rather than in the direction of dehumanization.

SOVEREIGN INDIAN NATIONS

The above considerations and principles are greatly accentuated in the case of Sovereign Indian Nations. In addition to being vulnerable to exploitation and injury, sovereign Indian Nations possess certain authorities that have established a government to government relationship between the Tribal Nations and the federal government. *Tribal privacy is as critical as individual privacy* and is far more difficult to achieve, particularly today.

- In many instances, the respective sovereign Indian Nations must be treated *as if they were individuals*.
- It is often impossible to protect the privacy of any given Indian Tribe. Indeed data belonging to Indian Nations are usually sought precisely because they derive from identifiable populations. Adding “anonymous” American Indian data to other pools is simply increasing the size of the *n*, often useful of course, but one is entitled to question just how important it may be to add the relatively small numbers of anonymized American Indian data to “Big Data”.
- De-identification of individuals and Tribes is a very insecure guard against privacy and must not be depended upon.
- Specific provision must be made for waiving the sharing requirement of certain data proscribed by an involved Tribe.
- Privacy and associated protections within Sovereign Indian Nations are determined by the respective Indian Nations themselves, not by an agency of the

US government. Any such federal authority resides in the United States Congress (Article I, Section 8 of the US Constitution).

- NIH rules regarding acquisition of Big Data are directed primarily at investigators. However, many of the questions relating to data sharing involve Tribes and the NIH must *negotiate* with the affected Tribes rather than issue rules to associated investigators.
- Negotiations between the funding agency and the Tribes are to be conducted in a government to government atmosphere. Involvement of the associated investigators in this process is inappropriate and unacceptable.
- Any Indian law, regulation, rule or other determination is what the respective Tribe determines it to be, not what a government official deems it to be. The NIH has no authority to rule upon the validity of any given rule established by the respective Tribal governments.
- Tribal ways of thinking, perceiving, believing, and acting remain obscure to the uninformed but they enter into almost all American Indian deliberations in one form or another. They must be given credence even if they are not understood.
- Tribes are not opposed to research nor the sharing of data. In fact many Tribes have been generously sharing data for decades, *always within the boundaries imposed by the Tribes*.
- There are very likely many Tribes that have no concerns about turning their data over to other investigators.

RECOMMENDATIONS

- Tribal requirements for privacy protection must never be used to deny funding for a worthy research project.
- Utilization, handling, and disposition of bio-specimens and data are subject to Tribal requirements expressed in laws, rules, regulations and policies.
- The proposed rule will emphasize that whenever an investigator detects a barrier to dissemination of data, the investigator must submit an extensive documentation describing not only the barrier but any plans to overcome the real or perceived barrier. This is backwards. Rather, it is the obligation of the funding agency to provide the respective Tribal governments a detailed description of the precise injury that is done to science by withholding certain Tribal data, including identification of the specific loss of knowledge occasioned by that withholding.
- In Tribally based NIH funded research, a statement that data sharing in any given instance is not possible is to be sufficient to provide an NIH waiver to data sharing in that instance. This was provided for in the 2003 NIH data sharing policy. It is imperative that it be included in the proposed policy.
- Further, a very important consideration is stated in the rule for sharing genome research: *Following, as appropriate, all applicable national, tribal, and state laws and regulations, as well as relevant institutional policies and procedures for handling genomic data* (FINAL GENOMIC DATA SHARING POLICY; Notice Number: NOT-OD-14-124). This rule is one of the most important of all NIH

rules and it must be included in the new proposed rule, at least as it applies to sovereign Indian Nations.

- Depending upon “de-identification” is not a sufficient safeguard for sovereign Indian Nations.

The present state of “Big Data” and the “sharing of data” particularly in regard to American Indian sovereign Nations is chaotic. During recent years several policies have been established, including this latest effort. It is necessary to have one NIH common policy regarding data sharing. It is unreasonable, and somewhat suspect, that the Tribes have to contend with several such policies. Even when the policies may be essentially the same, an inordinate amount of tribal time and resources is necessary to properly protect the interests of the Tribes and their respective members.

Tribes contend with an inordinate amount of time and effort in responding to data sharing requests and NIH rules. Many Tribes simply have no resources with which to respond to the lengthy considerations related to data sharing. It is now imperative that research proposals involving American Indian Tribes be properly budgeted in order to provide resources necessitated by the extensive responses to the many NIH demands.

Most respectfully,

Everett R. Rhoades MD FACP (ret.)

Submission #159

Date: 12/10/2018

Name: Julie Stoner

Name of Organization: University of Oklahoma Health Sciences Center

Type of Organization: University

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

Chronic disease epidemiology; Clinical research; Health-related research in partnership with American Indian tribes and tribal communities

II. The requirements for Data Management and Sharing Plans

For NIH-funded or -supported research projects that involve scientific data from American Indian and Alaska Native tribes, the Data Management and Sharing Plans must comply with tribal requirements and preferences regarding data management and sharing.

Submission #160**Date:** 12/10/2018**Name:** Heather Stevens**Name of Organization:** Accenture Federal Services LLC**Type of Organization:** Other**Other Type of Organization:** Professional Services**Role:** Member of the Public**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Our passion is ensuring that research data is shared effectively among the research community so that the knowledge gained can be converted into effective therapies faster.

I. The definition of Scientific Data

The current definition focuses on what Scientific Data is not. The definition should be expanded and made more descriptive to describe what Scientific Data is.

The definition should also make clear that data in isolation is of limited value. Meeting the spirit of the requirement that data be available for sharing and reuse means that data sets must be formatted and described in a standard way. Practically speaking, the associated descriptive information and metadata is as important as the data itself and must be included for a data set to be usable. This should be made clear in the definition.

The definition of scientific data must acknowledge that it is derived from primary observation. The definition should accompany examples of scientific data that can be used to help inform the definition, e.g.:

- Protein Structures (X-Ray Crystallography – numerical and images)
- Chemical Structures
- Protein Sequences
- Genomic data
- Proteomic Data
- Other -Omics data

- Cell-line Genealogy, Identifiers
- Gene Expression Data – as distinct from other genomic data
- Analytical Data (NMR, MS, IR, etc.)
- Purity Data – for reagents, compounds, etc.
- Assay Data – Numerical Results, Raw Images, etc.
- Coordinate Data – Plate Maps, etc.
- Animal Data – Species, Biology, Histo-Path (labs, images)
- Anatomical Image Data (CT, MRI, PET, etc.) – Human and Animal
- Physiological Measurements – Human and Animal
- Human Data (clinical data, demographic data, EHR, Claims, Social Media) – highly regulated
- Entity Registration Identifiers
- Scientist Notes (ELN, Paper Notebooks, Audio, Video, Files, --- highly variable)
- Literature References
- Patent Data – clinical, pharmacovigilance (PV), regulatory
- Biomarker data – gene

Further, the definition may also include examples of Operational Data that help frame or decipher scientific data, e.g.:

- User data – demographics, access, privileges, security
- Entity request and tracking – transactional data, location data, fulfillment data
- Inventory data
- Location data – sample stores, stockrooms, etc. – can range from barcodes, RFID, NFC, etc. to Video, Images, etc.
- Safety Data – MSDS, Specialized Safety Data, etc. – safety for all staff, vendors, janitorial staff, patients, etc.
- Corporate publications (annual reports, pipeline data, investor presentations, etc.)
- Regulatory filings, regulatory guidelines
- SOPs, Operating Manuals

II. The requirements for Data Management and Sharing Plans

The Data Management and Sharing Plan should be prescriptive. The scale and complexity of modern scientific data means that researchers must have data management skills, but it is important to remember that they are not professional data managers. A simple mandate requiring researchers to share their data will not succeed. The NIH needs to provide support in the form of guidance, standards, tools, and education for researchers to meet the requirements.

It is useful to use the analogy to an experimental section or materials and methods of a scientific paper when considering the requirements for a data management plan. That is, data must be described with sufficient detail to support understanding, evaluation, and reproduction by outside experts.

Recommendations for areas where explicit guidance from the Data Management and Sharing Plan is needed are discussed below.

Supported Disciplines

The current policy does not define specific knowledge domains but applies to all data generated from NIH-funded research. A one-size-fits-all data management policy will necessarily be very high-level and will not provide the specific guidance that researchers need.

The disciplinary scope of the data covered by the plan should be clearly defined. Supported disciplines should be identified and prioritized based on the number of researchers, number of grants, quantity of data, and expected contribution to high-level NIH goals. Specific discipline-focused data management guidance should then be released in a phased manner, based on the priority list.

Data Descriptors and Standards

Data sets must be accompanied by sufficient descriptive information to allow an external expert to understand and evaluate how the primary data was collected; how it was transformed to create the submitted data set; and to understand, evaluate, and reuse the submitted data.

As a practical matter, this means that appropriate data formats and standardized terminology, both for data annotation and data values, as well as guidelines for description and capture of experimental data must be available to researchers who submit data. Examples of each include the BAM data format, the Units of Measurement Ontology, and the Minimum Information About a Microarray Experiment (MIAME) standard, respectively.

A list of required data formats, standard vocabularies, and experimental guidelines should be created, maintained, and published. If an appropriate format, vocabulary or guideline does not

exist or does not meet NIH requirements, one should be developed by the NIH, in consultation with subject matter experts and recognized thought leaders.

Repositories

A list of approved repositories for data deposition should be created, maintained and published. Requirements for an approved repository should be explicitly and transparently enumerated.

If a publicly-accessible, discipline-specific data repository that meets NIH requirements does not exist, NIH should consider creating and maintaining one.

Data Lifecycle

Policies should be developed to describe the full data lifecycle. Important questions to answer include:

- How long is data maintained?
- What are the responsibilities of the researcher after data has been deposited in a repository?
- What is the process for updates and/or corrections to published data sets?
- What are the policies for data access, reuse, transformation, and citation by third parties?
- What is the process for data transformation or migration in response to new formats, standards, or technology?
- What is the process for data retirement or long-term archiving?

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

Accenture has worked with clients such as the Office of the National Coordinator for Health IT (ONC) in convening the health IT community on priority initiatives related to health information interoperability and in identifying relevant standards to address data sharing and access challenges in the clinical workflow. That work has informed our opinion on scientific data management.

The NIH maintains many domain-specific data repositories for use by the scientific community to deposit and share data. However, the repositories exist as information silos, with independent submission guidelines, formats and standards. This means that the user community must spend considerable time and effort to uncover the details of data submission.

In addition, not all domains have clearly defined standards. Finally, there is no mechanism to measure or enforce compliance in terms of data sharing requirements, use of appropriate data standards, or data quality within many domains or across the full set of data generated from NIH-supported research.

We propose a two-fold solution. NIH should implement a unified data loading solution in combination with a concerted effort to define data sharing standards for all data domains across NIH.

A modern smart data loading solution would consist of a single website where all types of data generated by NIH-funded research can be uploaded. The user would select the type of data that is of interest, download a template for description of the data, and then upload the filled-out template containing the data set. Data processing based on machine learning will check the uploaded data for consistency and compliance with the relevant standards prior to publishing the data in the appropriate repository. Upon completion of the data processing steps, the user would be notified of a successful upload or provided with details of exceptions where there are issues with data consistency or lack of compliance with standards. In case of exceptions, the user would be asked to take suggested corrective action and re-submit their data set for processing.

We also propose that data standards be defined for all types of scientific domains that are supported by the NIH with a view to automation of the data loading, processing and publishing steps. As a phased approach to developing the smart data loading solution, we would recommend prioritizing data domains based on the maturity of the current data sharing standards published by the NIH. As a parallel phased approach, an initiative to define data sharing standards for the new and/or less mature data domains should be started. These can be prioritized by the user community representatives tasked with working on the initiative.

Attachment:

Section I: The Definition of Scientific Data - Comments

The current definition focuses on what Scientific Data is not. The definition should be expanded and made more descriptive to describe what Scientific Data *is*.

The definition should also make clear that data in isolation is of limited value. Meeting the spirit of the requirement that data be available for sharing and reuse means that data sets must be formatted and described in a standard way. Practically speaking, the associated descriptive information and metadata is as important as the data itself and must be included for a data set to be usable. This should be made clear in the definition.

The definition of scientific data must acknowledge that it is derived from primary observation. The definition should accompany examples of scientific data that can be used to help inform the definition, e.g.:

- Protein Structures (X-Ray Crystallography – numerical and images)
- Chemical Structures
- Protein Sequences
- Genomic data
- Proteomic Data
- Other -Omics data
- Cell-line Genealogy, Identifiers
- Gene Expression Data – as distinct from other genomic data
- Analytical Data (NMR, MS, IR, etc.)
- Purity Data – for reagents, compounds, etc.
- Assay Data – Numerical Results, Raw Images, etc.
- Coordinate Data – Plate Maps, etc.
- Animal Data – Species, Biology, Histo-Path (labs, images)
- Anatomical Image Data (CT, MRI, PET, etc.) – Human and Animal
- Physiological Measurements – Human and Animal
- Human Data (clinical data, demographic data, EHR, Claims, Social Media) – highly regulated
- Entity Registration Identifiers
- Scientist Notes (ELN, Paper Notebooks, Audio, Video, Files, --- highly variable)
- Literature References
- Patent Data – clinical, pharmacovigilance (PV), regulatory

- Biomarker data – gene

Further, the definition may also include examples of Operational Data that help frame or decipher scientific data, e.g.:

- User data – demographics, access, privileges, security
- Entity request and tracking – transactional data, location data, fulfillment data
- Inventory data
- Location data – sample stores, stockrooms, etc. – can range from barcodes, RFID, NFC, etc. to Video, Images, etc.
- Safety Data – MSDS, Specialized Safety Data, etc. – safety for all staff, vendors, janitorial staff, patients, etc.
- Corporate publications (annual reports, pipeline data, investor presentations, etc.)
- Regulatory filings, regulatory guidelines
- SOPs, Operating Manuals

Section II: The Requirements for Data Management and Sharing Plans - Comments

The Data Management and Sharing Plan should be prescriptive. The scale and complexity of modern scientific data means that researchers must have data management skills, but it is important to remember that they are not professional data managers. A simple mandate requiring researchers to share their data will not succeed. The NIH needs to provide support in the form of guidance, standards, tools, and education for researchers to meet the requirements.

It is useful to use the analogy to an experimental section or materials and methods of a scientific paper when considering the requirements for a data management plan. That is, data must be described with sufficient detail to support understanding, evaluation, and reproduction by outside experts.

Recommendations for areas where explicit guidance from the Data Management and Sharing Plan is needed are discussed below.

Supported Disciplines

The current policy does not define specific knowledge domains but applies to all data generated from NIH-funded research. A one-size-fits-all data management policy will necessarily be very high-level and will not provide the specific guidance that researchers need.

The disciplinary scope of the data covered by the plan should be clearly defined. Supported disciplines should be identified and prioritized based on the number of researchers, number of grants, quantity of data, and expected contribution to high-level NIH goals. Specific discipline-focused data management guidance should then be released in a phased manner, based on the priority list.

Data Descriptors and Standards

Data sets must be accompanied by sufficient descriptive information to allow an external expert to understand and evaluate how the primary data was collected; how it was transformed to create the submitted data set; and to understand, evaluate, and reuse the submitted data.

As a practical matter, this means that appropriate data formats and standardized terminology, both for data annotation and data values, as well as guidelines for description and capture of experimental data must be available to researchers who submit data. Examples of each include the BAM data format, the Units of Measurement Ontology, and the Minimum Information About a Microarray Experiment (MIAME) standard, respectively.

A list of required data formats, standard vocabularies, and experimental guidelines should be created, maintained, and published. If an appropriate format, vocabulary or guideline does not exist or does not meet NIH requirements, one should be developed by the NIH, in consultation with subject matter experts and recognized thought leaders.

Repositories

A list of approved repositories for data deposition should be created, maintained and published. Requirements for an approved repository should be explicitly and transparently enumerated.

If a publicly-accessible, discipline-specific data repository that meets NIH requirements does not exist, NIH should consider creating and maintaining one.

Data Lifecycle

Policies should be developed to describe the full data lifecycle. Important questions to answer include:

- How long is data maintained?
- What are the responsibilities of the researcher after data has been deposited in a repository?
- What is the process for updates and/or corrections to published data sets?
- What are the policies for data access, reuse, transformation, and citation by third parties?
- What is the process for data transformation or migration in response to new formats, standards, or technology?
- What is the process for data retirement or long-term archiving?

Section III: The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards - Comments

Accenture has worked with clients such as the Office of the National Coordinator for Health IT (ONC) in convening the health IT community on priority initiatives related to health information interoperability and in identifying relevant standards to address data sharing and access challenges in the clinical workflow. That work has informed our opinion on scientific data management.

The NIH maintains many domain-specific data repositories for use by the scientific community to deposit and share data. However, the repositories exist as information silos, with independent submission guidelines, formats and standards. This means that the user community must spend considerable time and effort to uncover the details of data submission. In addition, not all domains have clearly defined standards. Finally, there is no mechanism to measure or enforce compliance in terms of data sharing requirements, use of appropriate data standards, or data quality within many domains or across the full set of data generated from NIH-supported research.

We propose a two-fold solution. NIH should implement a unified data loading solution in combination with a concerted effort to define data sharing standards for all data domains across NIH.

A modern smart data loading solution would consist of a single website where all types of data generated by NIH-funded research can be uploaded. The user would select the type of data that is of interest, download a template for description of the data, and then upload the filled-out template containing the data set. Data processing based on machine learning will check the uploaded data for consistency and compliance with the relevant standards prior to publishing the data in the appropriate repository. Upon completion of the data processing steps, the user would be notified of a successful upload or provided with details of exceptions where there are issues with data consistency or lack of compliance with standards. In case of exceptions, the user would be asked to take suggested corrective action and re-submit their data set for processing.

We also propose that data standards be defined for all types of scientific domains that are supported by the NIH with a view to automation of the data loading, processing and publishing steps. As a phased approach to developing the smart data loading solution, we would recommend prioritizing data domains based on the maturity of the current data sharing standards published by the NIH. As a parallel phased approach, an initiative to define data sharing standards for the new and/or less mature data domains should be started. These can be prioritized by the user community representatives tasked with working on the initiative.

Submission #161

Date: 12/10/2018

Name: Heidi Imker

Name of Organization: University of Illinois at Urbana Champaign

Type of Organization: University

Role: Institutional Official

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

Our university has strengths across all of the sciences, including basic life sciences and biomedical research, as well as in the arts, humanities, and social sciences. Our faculty are also highly interdisciplinary, hence the importance in ensuring these policies are easily harmonized across funding agencies and research domains.

I. The definition of Scientific Data

This definition is consistent with previous definitions used by OSTP and other agencies, which is helpful.

II. The requirements for Data Management and Sharing Plans

The University of Illinois at Urbana-Champaign has several comments on the proposed provisions for a draft Data Management and Sharing Policy. Please see the attached letter which outlines our feedback and is co-signed by the Vice Chancellor for Research and the Associate Dean for Research at Illinois' University Library.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

The policy should be implemented after integration of feedback received from the RFI and after a program has been developed to train NIH reviewers on how to evaluate the newly required data management and sharing plans. Without a structure for thoughtful review, these policies will be carried out unevenly, which will frustrate everyone and minimize the potential for positive impact.

Attachment:



OFFICE OF THE VICE CHANCELLOR FOR RESEARCH

Dr. Susan A. Martinis
Fourth Floor, Swanlund Administration Building, MC-304
601 E. John St.
Champaign, IL 61820-5711

December 07, 2018

Carrie D. Wolinetz, Ph.D.
Associate Director
NIH Office of Science Policy
National Institutes of Health
6705 Rockledge Drive #750
Bethesda, Maryland 20817

Dear Dr. Wolinetz:

Thank you for the opportunity to comment on proposed provisions for a draft Data Management and Sharing Policy for NIH-funded or supported research, as outlined in NOT-OD-19-014. The University of Illinois at Urbana-Champaign is committed to stewarding the data resulting from our federally funded research, and making this data as available as possible while safeguarding the privacy of research participants and protecting confidential or proprietary data. While we applaud NIH's initiative in putting forward the proposed provisions, we recommend that they be strengthened in a number of ways.

We encourage adherence to several critical recommendations from the [APLU-AAU Public Access Working Group Report and Recommendations](#), specifically:

- "Agencies should provide clear information on expectations regarding what data do and do not need to be shared" (and additional highly pragmatic recommendations therein) ... see also comment below about use of the phrase "all data" in the draft
- "Agency expectations for data access after the funding period has ended should be specific and finite in duration..." – the draft includes vague language like "as long as it is useful to the scientific community." This makes sense given uncertain value of some data but is, in effect, extremely unhelpful. Instead a minimum should be offered as a window to assess value – recommend calling out the expected 3 year retention period in OMB Circular A-110 as a *minimum* with a reference to HHS's RCR site:- https://ori.hhs.gov/education/products/rcradmin/topics/data/tutorial_11.shtml
- "Agencies should provide clear information on how compliance with data sharing requirements will be monitored, evaluated, and enforced ..." Without clarity on the first two bullets above, this piece, and, in particular, "evaluation" is not possible.

Attention should be focused on the terminology used in the policy. NIH should take this opportunity to align its policy with other agencies, and NSF in particular. Specifically, we urge NIH to use the same names that are used by NSF for similar elements of the two policies.

In the definition of **Scientific Data**, rather than saying, "NIH expects that reasonable efforts should be made to digitize all scientific data," please consider a more realistic standard, for example, "From this point forward, NIH expects that reasonable efforts should be made to digitize scientific data that is of value to the scientific community."

We noted that the terms storage and preservation are used interchangeably in the document. It is often not clear when the Plan Elements are referring to shared data only or any management of data

In the **Requirements for Data Management and Sharing Plans**, the draft includes a number of statements that may create confusion, and possibly lead to overreach or adoption of bad practices unless clarified. These include

- Any use of the phrase “all data”
- “indicate how intellectual property, including invention or other proprietary rights, will be managed in a way to maximize sharing of scientific data” – implying what options specifically? geographic exclusivity similar to that done for vaccines? Additional clarification could be offered if the policy called out Creative Commons licenses; for example, CC-BY-NC-SA 4.0 (<https://creativecommons.org/licenses/by-nc-sa/4.0/legalcode>) or even the data-specific licenses that have been developed (<https://opendatacommons.org/licenses/>).
- “Data preservation must be consistent with the NIH Strategic Plan for Data Science” – this plan has many parts, several of which are aspirational, and it would be best to be specific about standards for consistency.

In section IV-2, Related Tools, Software and/or Code must also specify a version and OS.

The draft calls out NIH-supported repositories in Section IV - 5, specifically those on the BMIC website, but these repositories will not accommodate all data types and several listed have uncertain funding futures (e.g. the MODs specifically). In addition to calling out repositories on the BMIC website, the Scientific Data Archiving section should be included in section IV - 5 and not part of “other considerations.”

When a federal repository is not available, the draft suggests appropriate repositories must “meet community-based standards”; however, such standards are often not codified, especially for new and/or emerging data types. Instead, appropriate repositories could be required to adhere, at minimum, to FAIR principals and, preferably, be Core Trust Seal certified. This provides some bounds for quality but leaves no one stranded without an “appropriate” repository.

We are very appreciative of NIH’s consultative approach in developing this policy. Please let us know if we can be of further assistance as it is finalized.

Sincerely yours,



Susan A. Martinis
Stephen G. Sligar Endowed Professorship in the School of Molecular and Cellular Biology
Interim Vice Chancellor for Research



Heidi Imker
Associate Dean for Research
University Library

Submission #162**Date:** 12/11/2018**Name:** Francis P. Crawley**Name of Organization:**

Good Clinical Practice Alliance - Europe (GCPA) & Strategic Initiative for Developing Capacity in Ethical Review (SIDCER)

Type of Organization: Nonprofit Research Organization**Role:** Bioethicist/Social Science Researcher**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

clinical, genomics, ethics, data protection, data sharing, ethical review (IEC/IRB)

I. The definition of Scientific Data

The GCPA and SIDCER thank the NIH Office of Science Policy for the opportunity to respond to this 'Request for Information on Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research'.

We consider the definition of 'scientific data' to be far too narrowly construed as essentially 'data used to support scientific publications'. Indeed, a much wider definition would be appropriate. We suggest: 'any data generated according to a scientific protocol or data used to support a scientific protocol, scientific analyses, and/or scientific reporting (including, but not limited to, publication'.

We further suggest that the distinction between 'data' and 'specimens / human tissue samples' is not tenable. Data and specimens should be handled in a similar way regarding data/specimens protection and data/specimens sharing.

The distinction between 'data' and 'metadata' is perhaps also not helpful. This distinction is based on context and use. There is no in se distinction.

II. The requirements for Data Management and Sharing Plans

Data management and sharing plans should be based on shared criteria and formats that promotes both data protection for identifiable persons and communities, while also ensuring the transferability and utility of the data to be shared. Criteria should be set down for the role of data managers (data controllers and data processors) for the collection, storage, processing,

and destruction of scientific data. Criteria should also be established for the de-identification and redaction of data (and this so as to ensure the resulting data's greatest utility). The roles and responsibilities of scientists, sponsors, funders, IRBs/IECs, and journal editors and publishers should be clearly delineated in the policy. There should be strong standards of, and enforcement of, data integrity. There should also be a clear indication of the data providers (data subjects) rights and the limitation of those rights with regard to data management and data sharing.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

The NIH should identify the specific goals it wants to achieve through adopting a policy on data management and data sharing. These goals should be articulated in terms of scientific needs, scientific outcomes, public health needs, and public health outcomes. The NIH should perhaps first phase in criteria and protocols for data management and data sharing, followed by registries and data bases, followed by a period of testing. The NIH should also consider to establish policy on who is entitled to receive scientific data, on what bases, and for what purposes.

Submission #163

Date: 12/10/2018

Name: YooRi Kim

Name of Organization: Gilbert Family Foundation

Type of Organization: Nonprofit Research Organization

Role: Institutional Official

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

clinical, genomics, neuroscience, gene therapy, cell therapy

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

How much training will the peer reviewers need to evaluate the Data Sharing Plans? It will be important to have very clear evaluation criteria or guidelines to assist the reviewers. This may take experience and a couple iterations of the criteria to achieve so I think the Plans being evaluated as "Additional Review Consideration" initially makes sense. However, as the NIH review system gets refined and the overall data sharing infrastructure gets more sophisticated (decreased barriers to participate in data sharing), one might consider moving to the "Additional Review Criteria" category because the actual impact of a research project does depend on the data and knowledge generated from the project becoming widely available to the research, medical, and patient communities, and on a timely basis.

Submission #164

Date: 12/10/2018

Name: Jennifer Hall

Name of Organization: American Heart Association

Type of Organization: Nonprofit Research Organization

Role: Institutional Official

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

Heart and brain health, genomics, precision health, precision medicine, epidemiology, clinical care and outcomes.

Attachment:

AHA response to proposed NIH Data Management and Sharing Policy

Background: On October 10, 2018, the NIH released a notice in its *Guide to Grants and Contracts* to solicit public input on proposed key provisions that could serve as the foundation for a future NIH policy for data management and sharing. Comments will be accepted until December 10, 2018. The feedback we obtain will help to inform the development of a draft NIH policy for data management and sharing, which is expected to be released for an additional public comment period upon its development.

Respondents are free to address any or all of the topics listed below, or any other relevant topic for NIH to consider. Respondents should not feel compelled to address all items.

Section I	The definition of Scientific Data
Section II	The requirements for Data Management and Sharing Plans
Section III	The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

NIH will consider all public comments before taking any next steps. No proprietary, classified, confidential, or sensitive information should be included in your response. Comments received, including any personal information, will be posted without change to [here](#). Comments may also be mailed to: Office of Science Policy, National Institutes of Health, 6705 Rockledge Drive, Suite 750, Bethesda, MD 20892, 301-496-9838.

To ensure consideration, responses must be submitted by: December 10, 2018 11:59:59 PM EDT.

AHA proposed comments

Section I The definition of Scientific Data

- The definition includes all the factual data necessary to replicate the research, but in the Scope and Requirement section (p. 2 of proposed provisions document) it only references sharing

scientific data that results from the research. This does not address the fact that in many cases a research question is assessed through further analysis of a previously generated data set (i.e., starting data) obtained from another investigator. It may or may not be the case that the starting data would have been generated in a manner such that it would have been subject to the proposed policy. To that end, we suggest the policy should encourage awardees, when applicable, to seek approval to share all data.

Section II The requirements for Data Management and Sharing Plans

- Item 4. Data Preservation and Access
 - Acceptable data repositories – A number of data repositories that are not NIH-supported may be appropriate sites for data storage. We suggest that NIH work with outside groups to publish a best-practices guide for data repositories that include minimal standards and accepted processes for data storage, data access, technology, security, harmonization, etc.
 - Additionally, we request it be clear that, whereas journals can include language about data sharing, they not be required to police data repositories.
 - Data Format -we request that the data format be preserved in a way that is interactable by researchers for software and computer coding.
 - While we acknowledge NIH’s efforts around data standards, we recommend a move towards international data standards.
 - We suggest a data preservation and data sharing policy that is more accessible by patients and consumers.
 - We suggest a data preservation and sharing policy that is modifiable for all types of data including but not limited to wearable device data, online application data, and social determinants of health data.
 - We suggest the NIH work with the electronic health record vendors on making this data more accessible to all researchers.
 - We suggest that NIH build in a timetable around GDPR.
- Item 5. Data Preservation and Access Timeline
 - Timeline for data deposits – The draft policy does not require awardees to deposit their data in a specific timeline. We recommend a maximum of 12 months after the award ends. An open data policy is designed to provide the researcher with prolonged – but not indefinite - first use of the data.
- Item 6. Data Sharing Agreements, Licensing, and Intellectual Property
 - Section 6.3. One could infer from the policy as written that the existence of or potential for intellectual property may be an appropriate reason for precluding data sharing. While investigators/institutions should be afforded reasonable time to protect intellectual property, the policy should more clearly convey that data sharing shall still occur when IP is present or anticipated.

Additional Comments

- re: Data Management and Sharing Plans

- Whereas it is always desired that publication(s) should result from funded research, we encourage NIH to make clear that even if research does not result in publications, that research is not exempt from the data sharing policy.
- Potential subject identification issues – Our experience has been that some awardees seek exemption from data sharing because of concern that study participants could potentially be identified. When/if this is suggested by an investigator, we encourage NIH to both design and enforce a process and policy that require the investigator to provide the levels of justification for exemption that may require the investigator’s IRB approval.
- re: Compliance and Enforcement/During the Funding or Support Period
 - Addressing how data plans can be modified throughout the life of the project – As the research, data, and/or available repositories may change throughout the course of a funded project, we encourage NIH to consider approaches that allow researchers to modify their data plan, i.e., treat it as a living document.

Section III The optimal Timing...

- We do not take a position on a specific timeline for implementation, but do encourage it to occur in the earliest timeline that is feasible for investigators and institutions.

Submission #165

Date: 12/10/2018

Name: Mark Cullen

Name of Organization: Stanford University School of Medicine

Type of Organization: University

Role: Institutional Official

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

Stanford University engages in a wide array of research ranging from physics to education. As policies set by the NIH are likely to act as a benchmark for other funding agencies, we expect that all schools will watch the outcome of these policies with interest. The Stanford School of Medicine engages in all manner of health research from dry lab research (biomedical data science and population health to name two), bench research, clinical trials and behavioral interventions. This policy will impact virtually all research activity at the School of Medicine.

I. The definition of Scientific Data

Please see attached letter.

II. The requirements for Data Management and Sharing Plans

Please see attached letter.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

Please see attached letter.

Attachment:



MARK R. CULLEN, M.D.
Senior Associate Vice Provost for Research
Senior Associate Dean for Research, SoM
Director, Center for Population Health Sciences
Professor of Medicine and Biomedical Data Science

Response to Proposed Provisions for a draft NIH Data Management and Sharing Policy
December 10, 2018

We appreciate the opportunity to comment on the NIH Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research (NOT-OD-19-014) released on October 10th, 2018.

Stanford University is among the top recipients of support from the National Institutes of Health, receiving a total of \$504,394,113 across 1040 awards in the 2018 fiscal year . Therefore, the proposed provisions for a draft data management and sharing policy would be highly significant for stakeholders across our organization

In the following sections, we respond specifically to three aspects of the proposed provisions:

1. The definition of scientific data
2. The requirements for data management and sharing plans
3. The optimal timing for NIH to consider in implementing various parts of a new data management and sharing policy.

In general, our responses emphasize the need for specificity in the definition of key terms and requirements as well as the development of guidance related to how researchers, reviewers, and other stakeholders should respond to the proposed requirements. Though it is implicit in many of our responses, we would further emphasize that the success of any policy related to data management and sharing will require empowering not only the recipients on NIH grants but also librarians, information technologists, and other parties involved in facilitating the management and sharing of research data.

At Stanford, data management and sharing-related resources are available through a variety of sources. Supporting the proposed policy will require the coordination and advancement of existing research support personnel, services, and infrastructure. Reasonable costs associated with data management and sharing can already be requested as part of the budgets for individual NIH funded projects but providing support for these activities at scale will require a substantial institutional investment in both technological infrastructure and human resources. We hope that, as the proposed policy is developed and implemented, it will be accompanied by mechanisms for supporting these investments.

Yours,

A handwritten signature in black ink, appearing to read "Mark Cullen".

Mark Cullen, MD
Senior Associate Dean for Research, Stanford University School of Medicine
Senior Associate Vice Provost for Research, Stanford University

I. The Definition of Scientific Data

Scientific Data: The recorded factual material commonly accepted in the scientific community as necessary to validate and replicate research findings including, but not limited to, data used to support scholarly publications. Scientific data do not include laboratory notebooks, preliminary analyses, completed case report forms, drafts of scientific

papers, plans for future research, peer reviews, communications with colleagues, or physical objects, such as laboratory specimens. For the purposes of a possible Policy, scientific data may include certain individual 2 level and summary or aggregate data, as well as metadata. NIH expects that reasonable efforts should be made to digitize all scientific data.

Given the breadth of research supported by the NIH, a bottom-up approach to defining a term as complex as “scientific data” is reasonable. By deferring to how the term is used within the research community, rather than imposing a top-down interpretation that would likely not be uniformly applicable across different domains, the definition outlined in the draft proposal is as potentially inclusive as it is succinct. However, to further clarify the definition of “scientific data” in the proposed data management and sharing policy, we make the following recommendations:

1. Data sharing is already being advocated within many research communities as a way to validate and replicate research findings. However, this does not necessarily mean that there is consensus on how data should actually be shared. For example, best practice documents written by and for researchers working with brain imaging data have promoted data sharing as a way to address concerns related to the validity and reproducibility of research results (e.g. Nichols et al., 2017) and standards have been developed to ensure the usability of shared data (e.g. Gorgolewski et al., 2016). In practice, such data is currently made available in a wide variety of formats (e.g. “raw” data, preprocessed data, statistical maps) which may or may not be accompanied by documentation, research related source code, and other essential materials. For this reason, we recommend that the definition of “scientific data” in the proposed policy be expanded to include guidance or operational definitions related to the meaning of terms such as “validate” and “replicate”.

Addressing whether or not the intent of the policy is for data to be shared in a form that would allow others to trace research results all the way back to the raw data (including a definition of what data is actually considered “raw” and details on how analytical pipelines and related materials should also be made available) may help research communities reach consensus- where none currently exists- about how data should be curated and the form in which it should ultimately be shared.

2. We also recommend that the definition of “scientific data” be expanded to, in plain language, state that the material necessary to make use of a particular set of observations (e.g. documentation, data dictionaries, notes about procedures or processes applied, etc) also counts as “scientific data”. Though this is addressed somewhat with the inclusion of “metadata” in the current definition, that term can easily be misunderstood as applying particular standards or schemas rather than simply additional information that makes data more usable.

3. Just as the definition of what constitutes a clinical trial according to the NIH is accompanied by a list of examples, we recommend that the definition of “scientific data” in the proposed policy include illustrative examples and guidance drawn from a range of methodological techniques and research areas. The current definition includes a number of terms that may be used in different ways by individuals in different contexts (e.g. “laboratory notebook”, “metadata”). Specific examples and guidance would help provide clear, operationalizable, criteria about how to comply with the proposed policy.

We applaud the NIH for working to ensure that scientific data are “managed, preserved, and made accessible in a timely manner for appropriate use by the research community and the public.” In line with our above recommendations pertaining to the definition of “scientific data”, we recommend clarification, written in plain language and accompanied by examples, about how terms including “timely”, “accessible”, and “appropriate” will be defined under data management and sharing policy.

II. The Requirements for Data Management and Sharing Plans

Applications and proposals for NIH-funded or -supported research projects that result in scientific data would be required to include a Plan. If perceived barriers to sharing scientific data exist (e.g., sharing includes specific restrictions or sharing is not possible), the Plan would be required to outline how scientific data will be managed and preserved and include an explanation of the perceived barriers. The Plan would also need to identify strategies or approaches to ensure adequate data security and compliance with privacy protections.

In developing its own policy, the NIH has an opportunity to draw upon a body of work examining the efficacy of data management plan (DMP)-related policies instituted by other funding agencies. In general, such work highlights the necessity of providing a clear explanation of requirements (including definitions of key terms and phrases) as well as specific instructions for managing compliance. For example, analyses of DMPs associated with grant proposals submitted to the National Science Foundation (NSF) generally demonstrate a wide variation in terms of how well researchers describe both their data and practices related to its organization, documentation, preservation, and dissemination (e.g. Bishoff and Johnson, 2015; Parham et al., 2016; Van Loon et al., 2017). Given this, it is perhaps not surprising that requiring a DMP, at least on its own, appears to have little bearing on if data is ultimately shared in a manner that enables its (re)use (Van Tuyl and Whitmire, 2016).

The above summary is not intended as a critique of researchers or the policies of a specific agency, but rather to illustrate the difference between instituting a policy and motivating a change in behavior. To help ensure that the requirements for the proposed data management and sharing plans are effective in facilitating the management, preservation, and accessibility of scientific data, we make the following recommendations:

1. The draft proposal outlines a number of elements to be included in the proposed data management and sharing plans. However, despite the fact that researchers receive relatively little formal education on related topics (Tenopir et al., 2016), these elements are generally described using language and terminology that assumes a high degree of familiarity with data management, preservation, and sharing. Drawing from approaches developed by the scholarly communications community (e.g. Borghi et al., 2018; Kafel et al., 2014), we recommend that proposed requirements be accompanied by guidance that includes jargon free descriptions of these concepts and specific examples of how researchers can apply them within the context of their work.

2. Surveys of data management practices in specific research communities (e.g. Borghi and Van Gulick, 2018) indicate that researchers' data management practices are primarily motivated and limited by practical concerns such as a desire not to lose data and a lack of time and discipline-specific best practices. For this reason, we recommend that any guidance related to the elements of the proposed data management and sharing plans also describe how such elements are meaningful in the context of a researcher's day-to-day work with data. For example, emphasizing that establishing robust data management practices during the early stages of a project can help prevent the loss of data and ultimately make the sharing of data at the conclusion of a project more efficient may help to reinforce the importance of such practices beyond simply satisfying the requirements of a funding agency.

3. We recommend that, when defining the requirements for data management and sharing plans, the NIH be as transparent as possible regarding how the content of such plans will be evaluated, both independently and in the context of evaluations of associated proposals. Such transparency would be useful for researchers as they are completing their plans but would also significantly aid in the development of guidance and best practices by other stakeholders (e.g. libraries, research offices).

The proposition mentioned in the proposed policy, that plans may be evaluated as an "additional review consideration", is not unreasonable, as it would allow NIH staff to work with potential awardees to address reviewer concerns related to their plans. However, we urge the NIH to develop guidance to educate reviewers as well as potential awardees about what constitutes an "acceptable" data management plan.

- III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

Perhaps the best model for assessing how the proposed data management and sharing policy could be implemented is the NIH public access policy, which requires scientists to submit final peer-reviewed journal manuscripts that arise from NIH funds to PubMed Central (PMC) immediately upon acceptance for publication. This policy boasts a high degree of compliance when compared to those of other funders (Larivière and Sugimoto, 2018), likely due to an alignment of requirements with what is currently possible as well as the availability of technical infrastructure,

guidance and resources, and clearly defined consequences for non-compliance. Though data poses a number of unique challenges, we would urge the NIH to consider similar factors when implementing the various parts of the new data management and sharing policy.

Alignment of Expectations with What is Currently Possible

An essential first step in implementing the new data management and sharing policy is to ensure that researchers are actually capable of complying. Just as the public access policy includes provisions that allow researchers to comply while also fulfilling requirements from other stakeholders (e.g. copyright requirements from publishers) by embargoing their papers and/or depositing author manuscripts in PMC, the proposed data policy must be adopted in such a way that researchers are actually able to comply.

In practice this means that the phasing of the proposed data management and sharing policy must take into account an evaluation of the related policies, resources, and infrastructure of other data stakeholders (e.g. publishers, other funding agencies, institutions). If there are gaps, such as a lack of policy frameworks, standards (including metadata standards), or technical infrastructure necessary for managing and sharing particular types of data (e.g. data containing protected health information, data that is subject to copyright or other intellectual property concerns, etc), that must be accounted for as the policy is implemented. Then, as such frameworks, standards, and infrastructure develop, so too should the policy.

Availability of Infrastructure

Key to the NIH public access policy's high rate of compliance is PubMed Central (PMC)- a single repository where articles are deposited and subsequently made available with minimal effort and no financial cost to the researcher. We hope that the proposed data management and sharing policy will leverage the efforts to modernize and support data-related infrastructure outlined in the NIH strategic plan for data science but add the caveat that fostering data management and sharing requires more than developing new and improved technology.

In addition to PMC, the success of the NIH public access policy can be traced to the compliance monitor, the online tool that allows an individual at a given institution to monitor compliance with the policy. Using the compliance monitor, a librarian or other professional can identify articles that are currently out of compliance and subsequently contact the responsible researcher to offer guidance and resources. For this reason, we recommend that the phasing of the NIH data management and sharing policy also be tied to the development of tools that allow librarians, compliance managers, and other professionals to track the data associated with a given award at their institution (e.g. access DMPs, identify when datasets have been deposited into an appropriate repository) and subsequently dispatch guidance and resources as needed to ensure compliance.

Availability of Guidance and Resources

Though the FAIR data principles were developed to address the machine usability of data, ensuring it is actually findable, accessible, interoperable, and reusable largely depends on the data management practices of individual researchers during their day-to-day work with data.

In a recent survey, training in data management was cited as among the greatest unmet needs of researchers funded by the biological sciences directorate of the National Science Foundation (Barone et al., 2017). To ensure its effectiveness, we therefore recommend that the phasing of the data management and sharing policy be tied not just to the development of infrastructure, but also to the development of research programs designed to characterize current practices within the research community and data-driven educational resources (e.g. online guides, workshops, etc) designed to advance researchers' understanding of how to apply concepts related to data management, sharing, and preservation within their own work.

Well Defined Consequences for Non-Compliance

Compliance with the NIH public access policy is mandatory, and non-compliance may result in the suspension or withholding of funding. As with the public access policy, we recommend that the phasing of the data management and sharing policy be tied to well defined consequences for non-compliance. At the start of the policy, the consequences for non-compliance should be relatively minimal and seen as an opportunity to identify gaps in the education and infrastructure needed to effectively manage and share data. As such gaps are closed, the consequences for non-compliance should be gradually increased.

In making this recommendation we acknowledge the complexity of assessing compliance with any proposed data policy. Identifying whether or not a dataset has been deposited into an appropriate repository is relatively simple compared to determining whether or not it has been deposited in a form that is actually useful (i.e. with appropriate documentation and metadata, stored in accessible file formats, etc). We therefore reiterate our previous recommendation for transparency- this time in how compliance will be measured during and following the support period- so that data stakeholders (researchers, librarians, etc) can develop appropriate services, guidance, and best practice recommendations.

Work Cited

Barone, L., Williams, J., & Micklos, D. (2017). Unmet needs for analyzing biological big data: A survey of 704 NSF principal investigators. *PLOS Computational Biology*, 13(10), e1005755. <https://doi.org/10.1371/journal.pcbi.1005755>

Bishoff, C., & Johnston, L. (2015). Approaches to data sharing: An analysis of NSF data management plans from a large research university. *Journal of Librarianship and Scholarly Communication*, 3(2), eP1231. <https://doi.org/10.7710/2162-3309.1231>

Borghi, J. A., & Van Gulick, A. E. (2018). Data management and sharing in neuroimaging: Practices and perceptions of MRI researchers. *PLOS ONE*, 13(7), e0200562. <https://doi.org/10.1371/journal.pone.0200562>

Borghi, J., Abrams, S., Lowenberg, D., Simms, S., & Chodacki, J. (2018). Support Your Data: A research data management guide for researchers. *Research Ideas and Outcomes*, 4, e26439. <https://doi.org/10.3897/rio.4.e26439>

Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., ... Poldrack, R. A. (2016). The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific Data*, 3, 160044. <https://doi.org/10.1038/sdata.2016.44>

Kafel, D., Creamer, A., & Martin, E. (2014). Building the New England Collaborative Data Management Curriculum. *Journal of EScience Librarianship*, 3(1). e1066. <https://doi.org/10.7191/jeslib.2014.1066>

Larivière, V., & Sugimoto, C. R. (2018). Do authors comply when funders enforce open access to research? *Nature*, 562(7728), 483-486. <https://doi.org/10.1038/d41586-018-07101-w>

Nichols, T. E., Das, S., Eickhoff, S. B., Evans, A. C., Glatard, T., Hanke, M., ... Yeo, B. T. T. (2017). Best practices in data analysis and sharing in neuroimaging using MRI. *Nature Neuroscience*, 20(3), 299–303. <https://doi.org/10.1038/nn.4500>

Parham, S. W., Carleson, J., Hswe, P., Westra, B., & Whitmire, A. (2016). Using data management plans to explore variability in research data management practices across domains, *International Journal of Digital Curation*, 11(1). 53-67. <https://doi.org/10.2218/ijdc.v11i1.423>

Tenopir, C., Allard, S., Sinha, P., Pollock, D., Newman, J., Dalton, E., ... Baird, L. (2016). Data management education from the perspective of science educators. *International Journal of Digital Curation*, 11(1), 232–251. <https://doi.org/10.2218/ijdc.v11i1.389>

Van Loon, J. E., Akers, K. G., Hudson, C., & Sarkozy, A. (2017). Quality evaluation of data management plans at a research university. *IFLA Journal*, 43(1), 98–104. <https://doi.org/10.1177/0340035216682041>

Van Tuyl, S. V., & Whitmire, A. L. (2016). Water, water, everywhere: Defining and assessing data sharing in academia. *PLOS ONE*, 11(2), e0147942. <https://doi.org/10.1371/journal.pone.0147942>

Submission #166**Date:** 12/10/2018**Name:** Laura Platero**Name of Organization:** Northwest Portland Area Indian Health Board**Type of Organization:** Other**Other Type of Organization:** Area Indian Health Board**Role:** Patient Advocate**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

On behalf of the Northwest Portland Area Indian Health Board (NPAIHB), I submit the following comments on the National Institutes of Health (NIH) Request for Information (RFI) on Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research, dated October 10, 2018. Established in 1972, the NPAIHB is a non-profit, Tribal organization under the Indian Self-Determination and Education Assistance Act (ISDEAA), P.L. 93-638, representing the 43 federally-recognized Indian Tribes in Idaho, Oregon, and Washington on health care issues. In the Portland Area, 75% of the total IHS funding is compacted or contracted and includes 6 federally operated service units, 16 Title I Tribes, 26 Title V Tribes, 3 urban facilities, and 3 treatment centers. NPAIHB operates a variety of important health programs on behalf of our member tribes, including the Northwest Tribal Epidemiology Center, and works closely with the IHS Portland Area Office. NPAIHB would like to offer the following comments with regard to the aforementioned request for information.

The Tribes of the Portland Area have issued a tribal resolution (Attachment A) indicating that the tribes retain ownership of data. To this end the Northwest Tribal Epidemiology Center has created a HIPAA compliant data repository to house data from tribal research, and research and surveillance efforts of our own Epidemiology Center and others in our region who may request this service.

We believe that, as part of the U.S. federal government's trust obligation to federally recognized Tribal Nations, NIH has a duty to honor, protect and uphold Tribal Nation sovereignty in its efforts to 'seek fundamental health and scientific knowledge and the application of that knowledge for health enhancement. Therefore, it is NPAIHB's recommendation that all submitted data management and data sharing plans require evidence of Tribal Nation consent for data sharing.

No Tribal Nation data should be included in any level of access without explicit Tribal Nation consent. The consent mechanism varies from Tribal Nation to Tribal Nation and may take the form of Tribal Nation Council resolutions, signed MOU's with the designated Tribal Nation leader, etc. In addition to documented Tribal Nation consent, the plan must address additional considerations between the researcher and the Tribal Nation such as:

- Data ownership and sovereignty, pursuant to NPAIHB Tribal Resolution 05-04-04
- Publication requirements and Tribal Nation consent procedures, including the use of relevant Institutional Review Boards, as tribally designated.
- Tribal preference and consideration for specimen use, storage, and destruction policy, including possible return to the tribal nation and or individual
- Work product ownership, stewardship and sovereignty agreements.

We strongly believe that any funded research in Indian Country must have the appropriate tribal approvals in place prior to funding being released. We further believe that tribes must be involved in the conception of the work to ensure that it is congruent with the beliefs and wishes of the tribe(s) involved as sovereign nations.

Conclusion

NPAIHB commends NIH for requesting comments on these issues and we thank you for this opportunity to provide comments and recommendations on behalf of NPAIHB and our member tribes. If you have any questions about the information discussed above, please contact Laura Platero, Government Affairs/Policy Director at (503) 407-4082 or by email to lplatero@npaihb.org. and Victoria Warren-Mears, PhD. Northwest Tribal Epidemiology Center Director at (503) 228-4185 or by email to vwarrenmears@npaihb.org.

I. The definition of Scientific Data

For the purpose of the NIH Data Management and Sharing Policy, in Section I, Definitions, the definition of "Scientific Data" should be modified to include....."data used to support scholarly publications and scholarly presentations." Often scientists conduct research that may not be submitted or accepted for publication and use scholarly presentations as a method to disseminate their data to advance scientific knowledge. Scholarly presentations share data to help advance the field of knowledge, which is fundamental to the definition of research. This definition should also note that American Indian and Alaska Native tribes, as sovereign nations, have the right to determine their own definition of data in their own data sharing and management plans and tribal research codes. The NIH Common Rule affirms that researchers with federal funding must follow tribal research codes, which can be more restrictive than the Common Rule. Nothing in this data management and data sharing policy should conflict with the ability of a tribe to establish its own definitions in its own research codes and data

management and sharing policies for researchers it chooses to partner with on research. NIH should consult with American Indian and Alaska Native tribes on this draft NIH Data Management and Sharing Policy before a final policy is implemented and report back on how their input was incorporated into the final version.

II. The requirements for Data Management and Sharing Plans

The Tribes of the Portland Area have issued a tribal resolution (Attachment A) indicating that the tribes retain ownership of data. To this end the Northwest Tribal Epidemiology Center has created a HIPAA compliant data repository to house data from tribal research, and research and surveillance efforts of our own Epidemiology Center and others in our region who may request this service.

We believe that, as part of the U.S. federal government's trust obligation to federally recognized Tribal Nations, NIH has a duty to honor, protect and uphold Tribal Nation sovereignty in its efforts to 'seek fundamental health and scientific knowledge and the application of that knowledge for health enhancement. Therefore, it is NPAIHB's recommendation that all submitted data management and data sharing plans require evidence of Tribal Nation consent for data sharing.

No Tribal Nation data should be included in any level of access without explicit Tribal Nation consent. The consent mechanism varies from Tribal Nation to Tribal Nation and may take the form of Tribal Nation Council resolutions, signed MOU's with the designated Tribal Nation leader, etc. In addition to documented Tribal Nation consent, the plan must address additional considerations between the researcher and the Tribal Nation such as:

- Data ownership and sovereignty, pursuant to NPAIHB Tribal Resolution 05-04-04
- Publication requirements and Tribal Nation consent procedures, including the use of relevant Institutional Review Boards, as tribally designated.
- Tribal preference and consideration for specimen use, storage, and destruction policy, including possible return to the tribal nation and or individual
- Work product ownership, stewardship and sovereignty agreements.

We strongly believe that any funded research in Indian Country must have the appropriate tribal approvals in place prior to funding being released. We further believe that tribes must be involved in the conception of the work to ensure that it is congruent with the beliefs and wishes of the tribe(s) involved as sovereign nations.

For the purpose of the NIH Data Management and Sharing Policy, in Section IV, "Requirements for Data Management and Sharing Plans," at the end of the first paragraph, the following should be added: "All Plans involving research and Scientific Data with American Indian and Alaska Native tribe(s) should include specific information on how the Plan complies with their tribal research codes, documentation of official tribal approval(s) for the Plan, and should

describe in detail how the Plan implements tribal requirements and preferences on data management and sharing to ensure that tribal nation(s) and their citizens, lands, and resources are protected, along with how the Plan will implement any tribal restrictions to data sharing.”

Under “Plan Review and Evaluation”, add a new bullet at the end of the list that states the following: “For all Extramural Grants, Contracts, NIH Intramural Research Projects, and Other funding/support agreements, the Plan should be determined to be “unacceptable” by reviewers or NIH staff if the Plan involves data from American Indian and Alaska Native tribes and does not include specific information on how the Plan complies with tribal research codes, documentation of official tribal approval(s) for the Plan, and detailed descriptions of how the Plan implements tribal requirements and preferences on data management and sharing to ensure that tribal nation(s) and their citizens, land and resources are protected, along with how the Plan will implement any tribal restrictions to data sharing.”

Under “Plan Elements”, add a new number 4 that states the following: “All Plans should indicate if their data includes data and information from American Indian and Alaska Native tribes or individuals, and if so, the Plan should include a one page addendum that describes in detail how the data was obtained, whether there is/was documentation of American Indian/Alaska Native tribal approval to obtain the data, and documentation of approval(s) from American Indian and/or Alaska Native tribes for the Plan for data sharing. The Plan should describe for each element listed above how the Plan will implement any tribal requirements or restrictions relevant to each element of the Plan. The Plan should have documentation that the tribe(s) affirmatively approve each element of the Plan.” Examples of how the tribe may approve each element of the Plan include the following: the tribal must approval all types of data that may be collected and shared, and reserves the right to not approve certain or any types of data to be shared; the tribe must approve all other information, including relevant associated data, that may be shared; the tribe must approve the methods of how the data will be processed or analyzed; the tribe must approve any standards used in data collection and sharing. Even though some tribes may value and encourage data sharing, any Plan should clearly include information that affirms any tribal approvals, requirements, restrictions, or denials of any or all elements of the Plan. NIH should consult with American Indian and Alaska Native tribes on this draft NIH Data Management and Sharing Policy before a final policy is implemented and report back on how their input was incorporated into the final version.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

NIH must consult as soon as possible with American Indian and Alaska Native tribes on any data management and sharing policy before implementation. According to the Department of Health and Human Services Tribal Consultation Policy, which applies to all Divisions in the Department including NIH, “Before any action is taken that will significantly affect Indian Tribes

it is the HHS policy that, to the extent practicable and permitted by law, consultation with Indian Tribes will occur. Such actions refer to policies that... have tribal implications and have substantial direct effects on one or more Indian Tribes..." NIH's draft Data Sharing and Management Policy meets this definition of an action that will significantly affect Indian tribes and thus requires tribal consultation before it is implemented.

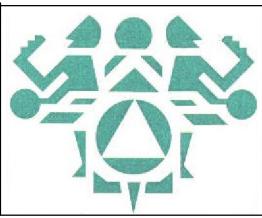
To our knowledge, NIH has received a large amount of input from American Indian and Alaska Native tribes and other individuals with related expertise on the importance of developing its Data Sharing and Management plan in partnership with tribes and the importance of incorporating the input and recommendations of tribes to protect their sovereign rights to govern research and data that involves their citizens, lands, and resources. As of the date of this submission, American Indian and Alaska Native tribes and others have expressed in public forums their dissatisfaction with NIH's efforts to date to incorporate tribal recommendations into this draft policy.

We recommend that NIH immediately consider this draft NIH data sharing and management policy to represent a Critical Event, affirm that it impacts all 573 American Indian and Alaska Native tribes, and immediately initiate tribal consultation through a process that begins with a letter to all tribes with the current draft of the policy attached. NIH should understand that a Request for Information released to the public is not a form of tribal consultation. NIH should initiate a tribal consultation on this draft policy as soon as possible and should not wait for the draft NIH tribal consultation protocol under development by the NIH Tribal Research Office to be finalized. NIH should allow for in-person tribal consultation, should carefully review tribal input, and should communicate with all tribes how their recommendations will be reflected in the final draft of the NIH Data Sharing and Management Policy and give tribes one more opportunity at that point to consult that the final draft. All input received by tribes to date should be summarized by NIH and shared with tribes at the onset of this tribal consultation. NIH should understand that tribal consultation is a policy of the Department of Health and Human Services and that any policies developed without tribal consultation represent a litigation risk to the agency, the Department, and individual investigators. The topic of data sharing and management is a priority topic for American Indian and Alaska Native tribes who must have the opportunity to exercise their sovereignty over data management and data sharing, both of which have potential significant risks to their citizens, lands and resources.

Given that American Indian and Alaska Native tribes have experienced the harmful effects of inappropriate research and data sharing, and in some cases continue to experience those harmful effects, ensuring that a meaningful tribal consultation on the NIH data management and sharing policy is of utmost importance and urgency. Tribes understand the importance of research and data to eliminating disparities and can be helpful in developing recommendations and solutions to data management and sharing policies that respect their sovereignty, the government to government relationship between tribes and the federal government, and the

rights of tribes to enter into respectful partnerships with researchers on data management and data sharing.

Attachment:



**NORTHWEST
PORTLAND
AREA
INDIAN
HEALTH
BOARD**

- Burns-Paiute Tribe
- Chehalis Tribe
- Coeur d'Alene Tribe
- Colville Tribe
- Coos, Suislaw & Lower Umpqua Tribe
- Coquille Tribe
- Cow Creek Tribe
- Cowlitz Tribe
- Grand Ronde Tribe
- Hoh Tribe
- Jamestown S'Klallam Tribe
- Kalispel Tribe
- Klamath Tribe
- Kootenai Tribe
- Lower Elwha Tribe
- Lummi Tribe
- Makah Tribe
- Muckleshoot Tribe
- Nez Perce Tribe
- Nisqually Tribe
- Nooksack Tribe
- NW Band of Shoshoni Tribe
- Port Gamble S'Klallam Tribe
- Puyallup Tribe
- Quileute Tribe
- Quinaltult Tribe
- Sami, h Indian Nation
- Sauk-Suiattle Tribe
- Shoalwater Bay Tribe
- Shoshone-Bannock Tribe
- Siletz Tribe
- Skokomish Tribe
- Snoqualmie Tribe
- Spokane Tribe
- Squaxi11 bland Tribe
- Stillaguamish Tribe
- Suquamish Tribe
- Swinomish Tribe
- Tulalip Tribe
- Umatilla Tribe
- Upper Skagit Tribe
- Wam1 Springs Tribe
- Yakama Nation

527 SW Hall
Suite 300
Portland, OR 97201
-n- (503) 228-4185
FAX (503) 228-8182
www.npaihb.org

RESOLUTION# 05-(0)1...-}0 Y

Tribal Ownership of Health-Related Data

WHEREAS, the Northwest Portland Area Indian Health Board (NPAIHB) is a tribal organization under P.L. 93-638 that represents forty-three federally-recognized Indian tribes in Oregon, Washington and Idaho and is dedicated to assisting and promoting the health needs and concerns of Indian people in the Northwest, and

WHEREAS, the Northwest Portland Area Indian Health Board is dedicated to assisting and promoting the health needs and concerns of Indian people, and

WHEREAS, the primary goal of the Northwest Portland Area Indian Health Board is to improve the health and quality of life of its member tribes, and

WHEREAS, Northwest Tribes have the right to self-determination, and in exercising that right must be recognized as the exclusive owner of indigenous knowledge, cultural and biogenetic resources, and intellectual property: and

WHEREAS, these elements have been, and continue to be, damaged, destroyed, stolen, and misappropriated, as Tribal members have been the subjects of research for decades, with virtually no benefits returning back to the community from the research: and

WHEREAS, members of the NPAIHB recognize that one way to help safeguard the best interests of Northwest tribal communities is to utilize the Portland Area Indian Health Service Institutional Review Board (PAIHS IRB) to review proposed research protocols and in so doing help prevent research-related abuses of individuals and tribal communities, protect human subjects and traditional knowledge and properties, and to identify research-related benefits and risks to their Tribal communities; and

WHEREAS, members of the NPAIHB recognize that it must: (1) protect the people, culture, and natural resources of the NPAIHB from unauthorized scientific research; (2) reduce the adverse effects of research on Tribal communities; (3) ensure that researchers recognize Tribal control of research activities and Tribal ownership of all data and information generated or produced by such research, and; (4) Establish and provide a statutory basis to review and govern any research, collection, database, or publication undertaken on their Reservations; and

WHEREAS, any tribe that participates in health-related research must be given possession of the primary data (with the necessary protections taken to protect the rights and privacy and confidentiality of individuals).

NOW THEREFORE BE IT RESOLVED, that the Northwest Portland Area Indian Health Board hereby recommends that all health-related research undergo review and approval by the PAIHS-IRB prior to data collection and associated publication of reports; and

BE IT FURTHER RESOLVED that the tribe (and the PAIRS IRB, acting as an agent of the interests of American Indian and Alaska Native people, though not speaking for any individual tribe) have the opportunity to review and give input on publications (and presentations to the extent possible) while they are in draft form (NOT after already submitted to a journal or conference).

BE IT FINALLY RESOLVED, that there will be a formal process by which tribes and tribal organizations will give input as how data concerning their community is presented, and the following principles are adhered to in research projects concerning Northwest Tribal communities:

- I. that investigators will not transfer the data to any other party without formal agreement from the tribe (and oversight by the PAIRS IRB, if involved), and
2. that no secondary analyses are performed on the data that are different than those proposed in the original research protocol without a formal request to the affected tribe (and PAIRS IRB, if involved), and
3. that there are measures taken to meaningfully inform the community of the results of research, and
4. that the tribe has the opportunity to benefit from gains that come out of the research (whether that means monetary profits or benefits in terms of better health), and
5. that the tribe has control over how and when data is disposed of (meaning that the storage of data is explicitly laid out, as are the plans for where and when and how it will be destroyed when no longer needed).

CERTIFICATION

NO. 05 -OY-0:i

The foregoing resolution was duly adopted at the regular session of the Northwest Portland Area Indian Health Board. A quorum being established; 1 q for, 0 against, (2f abstain on -Suly lgn: , 200s. :

Dean Capatman Palley
Chairman

:S..1t\ 9\ 1).tfJ5
Date

Stella a (. . / u?
Secretary

Submission #167**Date:** 12/10/2018**Name:** Megan Potterbusch, Hiromi Sanders, Nina Hamburg, Anne Linton**Name of Organization:** George Washington University**Type of Organization:** University**Role:** Institutional Official**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Clinical, genomics, epidemiology, oncology, public health, geodata

I. The definition of Scientific Data

At our institution, we consider “Scientific Data” to include:

- Facts or information that serve as the basis for decision making, discussion and reasoning, or calculation
- Information on individuals, aggregated records, and summary data
- Intangible data: measurements, observations, calculations, interpretations, and conclusions
- Tangible data: materials including cells, tissues, biological specimens, gels, photographs, and micrographs, and other physical manifestations.

“Scientific Data” should also include:

- Information on how the data were compiled, processed, and stored
- Data files embedded with data dictionaries -- For data to be re-used/reusable, data files alone will not suffice.
- Supplementary information beyond metadata that describes the data

Other Considerations re Scientific Data:

- More requirements about software selection for processing the data
- More support for selecting and applying descriptive elements
- Data must be archived appropriately – with backups and persistent unique identifiers so that they can be included in repositories

II. The requirements for Data Management and Sharing Plans

Benefits from the new requirements:

- If researchers are required to outline how they will overcome the challenges to sharing their data as a part of their data management plans, they may be more likely to seek support/consultation services sooner in the process. This will hopefully result in more actionable plans, earlier contact with data repositories, improved IRB documentation, and appropriate language about ethical data sharing in participant consent forms.

Observations:

- The guidance provided in the draft outline is clear and helpful.
- Researchers may need additional help selecting software and tools that fit the requirements.
- It looks like the NIH will expect some of the costs for this work to be built into the grant proposal, but researchers will need help determining the costs, and the university will need to adjust how we manage these earmarked funds for technology and consultation infrastructure. A phased approach would greatly assist in the local adaptation necessary.
- The Glossary for Common Data Element Resource Portal and the support for data citations provided in PubMed Central will help researchers to comply with the FAIR Data Principles, but there are still some link issues in the Common Data Element Resource Portal, which will need to be resolved for researchers to trust and understand that resource.
- The draft seems to suggest that researchers use an established digital repository instead of “archiving” the files on a local server whenever possible. If this is the case, it would be helpful if the final language is even clearer about this expectation, and if the data management and sharing plan submitted goes through peer review, clear feedback and recommendations to researchers explicitly stating the importance of using an identifier-issuing data repository for archiving their data would be helpful.

Anticipated Challenges:

- Researchers often are not thinking through all the steps necessary to action their data management and sharing plans from the beginning. The more that the NIH can do to get the message of these changes out now, the better for researchers to begin adjusting their behavior and budgets for data sharing.
- Having a maximum of two pages assumes that there will be straightforward answers to local data management infrastructure. For example, our researchers may need additional space to fully “describe plans for protecting privacy and confidentiality, e.g., through de-identification or data aggregation prior to sharing.” Fully outlining data types to be collected and how they will

be described, managed, and shared may also require more space than would be allotted in a two page plan. An optional addendum section, for researchers who need it, might resolve this issue.

- If the university is to add compliance structure around projects, plans should allow for audits of the parties involved or some level of oversight to ensure that plans are being followed throughout the research.

- The infrastructure and resources provided by the NIH will greatly support researcher in complying with the FAIR Data Principles, but there is a knowledge gap to overcome in terms of all of these aspects (Findable, Accessible, Interoperable, and Re-Usable).

- Interoperability and Re-Usability will probably be the largest challenges in terms of compliance with FAIR Data Principles. These will be greatly supported by NIH repositories with support-staff available for consultations.

Questions:

- Is there a vision from the NIH regarding universities' efforts in data management regarding compliance or a required/recommended level of support from grantee institutions?

- What expectations should be managed by the institution regarding curating and archiving the data in the event that a researcher leaves in the middle of a project? How about after a project has completed, but before the data have been deposited in an appropriate repository?

Suggestions:

- NIH should be mindful of other agencies' policy on data management and sharing and should consider working with them early while drafting the policy so to harmonize policies across agencies to not increase administrative burden.

- Perhaps NIH could work with the FDP-DTUA sub-committee. The organization works to streamline processes so to reduce administrative burden.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

Potential Benefits:

- Phased adoption would be very helpful for universities to roll out services and augment infrastructure modularly.

- Enable enhanced infrastructure for locating data and improved reporting of summary results in ClinicalTrials.gov. Continued improvements for locating data sets in PubMed and PMC and PMC Europe.

- Promotion of FAIR principles so that data are findable, accessible, interoperable, and reusable across various subject repositories.
- Identification of data with long-term value for inclusion in an NIH super-repository. (Based on the NASEM report and other expert opinions.)

Observations and Challenges:

- Two of the main challenges to data sharing are time and funding.
- Timeframe: Implementing a phased adoption of this data management policy would enable adequate time to evaluate budget and resource constraints associated with preparing, curating, archiving and sharing data. This evaluation could lead to improved standard practices based on grant type, volume of data, length of funding, etc. and more accurate resource requests for data management overall.
- Funding: the draft policy references that “reasonable costs associated with data management and sharing could be requested.” To this end, general guidelines for conventional costs for labor and time expected to prepare data, metadata and documentation for sharing would be beneficial.
- Many investigators could leverage existing federal and academic data sharing platforms (most at no cost), but there may be a need to utilize other archiving environments, and a clear expectation of the time commitment for data availability (5 years? 10 years?) would better inform budget and resource decisions up front, during the grant and contract proposal stage.

Suggestions:

- It would be helpful for NIH to specify the length of time that access to data is expected, and to provide funding for that timeframe (which will occur beyond the life cycle of the grant) if the data cannot be deposited into an external repository.
- Specifying an expected timeframe for data sharing could also have potential impact on terms in data use agreements and consent forms, as well as IRB support.
- Allowing a budget line item for data archiving, including the personnel hours necessary for data curation and the preparation of metadata, would enhance the ability of organizations to comply with the new policy regardless of the size of particular research projects.

Submission #168

Date: 12/10/2018

Name: Chris Shaffer & John Chodacki

Name of Organization: University of California San Francisco & California Digital Library

Type of Organization: University

Role: Institutional Official

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

Graduate and professional health sciences

I. The definition of Scientific Data

See attached letter

II. The requirements for Data Management and Sharing Plans

See attached letter

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

See attached letter

Attachment:



December 10, 2018

Carrie D. Wolinetz, Ph.D.
Office of Science Policy
National Institutes of Health
6705 Rockledge Drive, Suite 750
Rockville, MD 20892

RE: NOT-OD-19-014: Request for Information on Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research

Dear Dr. Wolinetz,

The California Digital Library and University of California, San Francisco Library applaud the effort by the NIH to encourage FAIR (Findable, Accessible, Interoperable, and Re-usable) data principles, and we agree with the general statement in the **Data Management and Sharing Policy**: “Increasing access to scientific data resulting from NIH funding or support offers many benefits, and reflects NIH’s responsibility to maintain stewardship over taxpayer funds.” Our specific comments on the proposed revisions are below.

Section I. The definition of Scientific Data

Definitions

FAIR data principles should not stand alone as a definition of **data sharing**, though it is important to highlight FAIR principles to be consistent with the European Commission's Horizon2020 policy and other open data policies. To ensure that researchers can understand and comply, the definition should avoid jargon and strive to use language that is meaningful to those affected by the policy (i.e., researchers). The definition of data sharing should also give a clear sense of what actions a researcher should take (e.g., make data available in approved preservation repositories with citable DOIs, apply Creative Commons licenses or mediate access where appropriate, etc.). In addition, **data management** should be defined to clarify whether the NIH considers this to include acquisition practices, processing, validation, storage and security, etc.

The **Metadata** definition should be more descriptive and should indicate which types of metadata are to be included (descriptive, structural, and/or administrative?), and what types of scientific data should include metadata. The example given seems to be referring to data dictionaries, but DOIs and other examples could be given to illustrate the range of metadata types as well.

The definition of **Scientific Data** seems to partially follow the OMB Uniform Guidance § 200.315 (e)(3) definition, but this proposal adds the exclusion of laboratory notebooks, the option to include individual, summary, and/or metadata, and the expectation that “reasonable” efforts are made to digitize all data. These additions to the OMB Uniform Guidance are vague and may cause confusion to the researcher about what stages and level of information they are required to share. In addition, to validate and replicate research findings it is typically necessary to trace results from raw data through each step of analysis. Laboratory notebooks, which are increasingly digital, often contain all of this information, including scripts that automate data processing and analysis. In our opinion, the definition should not exclude laboratory notebooks since these may contain key factual recorded information that is essential to promote validation and reproducibility.

In general, the plan should provide more specific guidance about how data should be made accessible and usable (such as by including data dictionaries, documentation about data processing, protocols, code used for analysis, etc.). It should also provide specific examples to accompany key terms (e.g. “aggregate data”) that span a range of research areas to further clarify what the policy applies to and how researchers can comply.

Purpose

Timing requirements for data sharing should be more specific. For example, what does it mean to make data “accessible in a timely manner”, or to evaluate data plans “at appropriate intervals”?

This section also mentions the relationship between this policy and other NIH policies related to data sharing (e.g. the NIH GDS policy). It would be worth further explaining how these various NIH policies align with this one for various stakeholders, researchers and those supporting the research enterprise.

Scope and Requirements

The scope is quite broad and requirements are vague. Since the data management plan requirement will apply to all research that is funded in full or in part by the NIH, more guidance will be needed to assist researchers in preparing their data management and sharing plans. The NIH should indicate how they will be providing this guidance to researchers, especially those who have limited access to professional grant officers or other institutional support systems.

Section II. Requirements for Data Management and Sharing Plans

Plan Review and Evaluation:

The plan should describe how it will establish review criteria and provide training for NIH staff and reviewers to uphold these criteria. This very important consideration is missing from other federal policies as well, and constitutes a major obstacle for successful implementation.

Plan Elements:

Data Types:

Overall, this section overlaps with the definition of scientific data. This should be reconciled with the definition, with terms clearly defined (e.g. “raw data”, “processed data”, “data modality”) and with more examples given.

- 1.1 In addition to asking what types of data will be collected, the plan should describe which types of data the researcher intends to share.
- 1.2 It would be helpful to have more examples of relevant associated data, such as code, data dictionaries, etc. This could be part of a separate section (e.g. the next section).

Related Tools, Software and/or Code:

Data collection tools should be included in this section as well. This section could be renamed “Related Products: Software, Code, and Tools”. It should also be made clear that any such products needed to validate and replicate research findings are "required" instead of "may be helpful".

Standards:

This section may be vague and confusing to researchers not familiar with these concepts and distinction between data formats, identifiers, definitions and common data elements. These concepts should also be defined in the Data Types section, and standards should be emphasized throughout the proposal.

In addition to the NIH CDE Resource Portal, there should be a more training and resources to help researchers fulfill these requirements.

Data Preservation and Access:

Some of these sections could be combined (4.1 and 4.6, 4.5 and 4.7) or even removed (4.4).

4.1 Again, a more specifically defined timeline or benchmark (e.g. research article publication) for making data accessible should be specified, and “long-term preservation” should also be defined. It should also be mentioned that there are field-specific NIH repositories that researchers should target, with some key examples. In addition, more specific information about new repositories should be provided, e.g. do individual researcher’s laboratory websites qualify? Section 4.6 can be combined with this one.

4.2 Terms such as “persistent identifier”, “indexing tools” should be defined, and examples should be given.

4.3 It should be clarified whether this refers to active data storage practices or specifically to long-term preservation.

4.4 This should be removed as a requirement, since it would be hard for researchers to anticipate at the planning stage, may discourage commitment to the original plan, and will make enforcement more difficult.

4.5 This section also seems to conflate active data storage/management and preservation. It should also be clearly distinguished or combined with 4.7 using broader language around ethics and legal compliance.

4.6 This section should be part of 4.1.

4.7 See comments on 4.5. There should be more information about how the NIH will support researchers that must provide restricted access to assist them with compliance.

Data Preservation and Access Timeline:

5.1 and 5.2 These sections could be combined together, with more specific details on *when* secondary data should be provided, and also on the *time period* (e.g. does this refer to the duration of time that it will be made available?). Terms previously defined should be used here for consistency, including distinctions between data management and preservation. In general, the policy should provide more precise language and guidance rather than vague references to what researchers need to describe.

Data Sharing Agreements, Licensing, and Intellectual Property:

There should be more guidance and specificity in what the criteria should be. For example, data that will be made available to the public should be provided under the Creative Commons “No Rights Reserved” license (CC0). Proposed repositories that do not comply should be required to apply for a license or waiver.

Oversight of Data Management:

This section should be renamed "Roles and Responsibilities" or "Responsibilities and Resources" to be consistent with other policies specifying the same requirements. The list of components should also include storage and backup, and resourcing/ budgeting responsibilities.

Compliance and Enforcement

It is not clear how the NIH will monitor compliance on an annual basis. Will they be by looking for links to shared data, the same way they currently do with publications? It would also be helpful to know what the implications of non-compliance will be. The plan should mandate that all proposals include a data management and sharing plan according to FAIR principles in a trusted repository, especially if it contains Protected Health Information. It should also provide specific timelines for sharing and preservation of the data (e.g. within one year of publications supported by the grant, provided that the data is de-identified or shared in compliance with HIPPA and ethical guidelines for PHI).

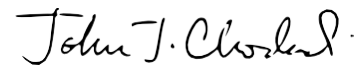
Section III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

Implementation of the plan should only occur after the NIH has evaluated and incorporated all input from the scientific community. All requirements and expectations need to be made clear for each phase, with clear assessment plans for advancement to the next phase. In addition, training for all stakeholders should occur throughout the process. Program officers and researchers need training to understand and support the requirements. A phased implementation could be appropriate, perhaps by first implementing the plan and training programs for a subset of proposals that are fully supported by the NIH with costs exceeding a certain amount.

Sincerely,

A handwritten signature in black ink, appearing to read "Chris Shaffer". The signature is fluid and cursive, with the first name "Chris" and last name "Shaffer" clearly distinguishable.

Chris Shaffer,
University Librarian & Assistant Vice Chancellor
UCSF Library

A handwritten signature in black ink, appearing to read "John J. Chodacki". The signature is written in a cursive style with a period at the end.

John Chodacki,
Director,
University of California Curation Center
California Digital Library

Submission #169**Date:** 12/10/2018**Name:** Rajni Samavedam**Name of Organization:** Booz Allen Hamilton Inc.**Type of Organization:** Other**Other Type of Organization:** Consulting firm**Role:** Scientific Researcher**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Genomics, bioinformatics, and data science.

I. The definition of Scientific Data

We support the general definition of ‘scientific data’ as given in the Proposed Provisions. This definition should be considered to include both raw and processed data as well as metadata describing not only the data, but the methods used to analyze the data, experimental samples, experimental methods, software/code, and statistical findings. We recommend that the Proposed Provisions also include additional context to the explicit exclusions listed in the definition for scientific data: “laboratory notebooks, preliminary analyses, completed case report forms, drafts of scientific papers, plans for future research, peer reviews, communications with colleagues, or physical objects, such as laboratory specimens.” Objects such as these may hold valuable historical, legal, or other information that may be used for research on the biomedical research enterprise as well as NIH’s organization, operations, and policies. In addition, documents that reflect activities of the NIH or interactions of the NIH with the public are Federal records (44 U.S.C. 3301; 36 CFR 1222.10). We recommend including text indicating that Federal records of NIH’s activities (e.g., laboratory notebooks, meeting minutes), while not deemed ‘scientific data’ per the Proposed Provisions, are nevertheless valuable and retained for a period of time in accordance with the NIH’s Extramural Research Records Schedule, Intramural Research Records Schedule, or by determination of the National Archives & Records Administration (NARA). Additionally, statistical data/statistical features (e.g., probability distributions, regression analysis) associated with the scientific data should also be included in the proposed definition of “metadata” and shared with the broader community.

II. The requirements for Data Management and Sharing Plans

Data Attribution: We also agree with the underlying premise of the Proposed Provisions – and of NIH’s data sharing policies in general – that the data produced by NIH-funded or -supported researchers is valuable to scientific community at large. The value of these data should also be recognized for the individual scientists or research teams who created them. However, scientific credit currently accrues to the authors of peer-reviewed journal articles, not necessarily to the providers of previously-generated data, experimental samples, software algorithms, or other scientific and analytical contributions. Ensuring traceable data and resource authorship will promote sharing of the full spectrum of scientific data. In addition, retaining details on the origin of digital sequence information (DSI) may be needed by institutions as they seek to comply with the 2010 Nagoya Protocol on Access to Genetic Resources & the Fair & Equitable Sharing of Benefits Arising from Their Utilization. Thus, the proposed data sharing and management plans should include details of data and resource authorship and/or provenance. One example of a flexible and interoperable data annotation scheme is NICHD’s Data and Specimen Hub (DASH). DASH stores all data as “digital objects” and each is annotated with structured, standardized metadata attributes suited to the type of object (e.g., datasets, documents, images). These annotations (metadata values) are derived from established standardized resources (e.g., Clinical Data Interchange Standards Consortium, NCI Thesaurus) whenever possible and presented to DASH users as a code list to select during annotation process.

Data Type: We agree that descriptions of data type should include information on the data modality (e.g., imaging, genomic, mobile, survey); level of aggregation (e.g. individual, aggregated, summarized); degree of data processing that has occurred (i.e., how raw or processed the data will be); the rationale for sharing or preserving data; and, for human-derived data, a description of the measures planned (e.g., de-identification, data aggregation). Because the storage of scientific data requires spaces and resources, we suggest that NIH consider that estimated amounts of data should also be expressed in terms of needed disk space in addition to estimates of the expected gene variants, cases, or study participants.

Related Tools, Software and/or Code: When referencing the software/computer code used to process and/or analyze the data, it is important to include both the code and the values of any variable parameters used when running that code, as well as the details of what platform on which the processing or analysis was done. Both platform and parameter settings can impact the reproducibility of the results. In addition, because open source tools, imputation pipelines, and algorithms that parse data or related software and code are often research outputs and “necessary to validate and replicated research findings,” we believe that they fall within the definition of scientific data and should therefore be deposited into a repository such as GitHub, with accompanying SOPs, user guide or “how-to” files, to ensure proper archiving, access and use by the scientific community.

Data Preservation, Access, and Discoverability: Current data-resource ecosystems tend to be “siloes” and are not optimally integrated or interconnected. An ecosystem of repositories is

only possible through the use of common data elements (CDEs) that serve as a Rosetta Stone to existing data management resources and tools and facilitate development of new ones. Unfortunately, CDEs are not yet fully adopted. NIH should continue to encourage and actively promote the use of CDEs as well as assess the feasibility of making CDE use mandatory. The National Library of Medicine (NLM) can serve as a central organizational hub that implements CDEs while transparently regulated, third-party run, secure, and interoperable enterprise data solutions may further support access to data to authorized parties in machine readable formats, such as through APIs. However, UUIDs (Universally Unique Identifiers) should also be used to enhance discoverability of the data as a persistent unique identifier.

Compliance and Enforcement: To encourage compliance, NIH should consider holding back a fraction (~5%) of funds and releasing them once the approved data sharing plan is fully implemented and data is completely shared. In order to determine if institutions are in compliance with their data sharing obligations, we suggest the plans also include additional details and guidance to assist in the determination of whether “the data sharing repository meets community-based standards at the time of deposition,” such as: use of CDEs; efforts to ensure data in the repository are secure and interoperable; the provision of APIs and software tools that facilitate storing, managing, standardizing and publishing data; linking of data to articles and software tools; inclusion of a data dictionary defining all metadata and data fields associated with the datasets to be stored; and contingency plans in the event that a repository becomes unavailable. Plans should be evaluated for compliance at least annually and must be a requirement for funded projects and contracts. The funded academic and industrial stakeholders should provide annual reports on the status of data sharing and management as outlined in plan.

Data Preservation and Access Timeline: NIH will need to set reasonable standards for when data is to be made available to secondary data users and how long it should be preserved. While the data generators should have rights of first analysis, it is unreasonable to keep the data from being made public for a year or more after deposition. Deposition of the data must also be made in a timely fashion after completion of the research that generates the data. It will be difficult to estimate how long the data should be preserved, due to factors such as how well the data ages in terms of the technology used to produce it, the cost of generating the data, how difficult recreating similar data would be, the expected usage of the data over time, and of course, the cost of preserving the data and accessing it over time.

Scientific Data Archiving: We are heartened to see that the NLM is working with the National Academies of Sciences, Engineering, and Medicine to conduct a study on forecasting the long-term costs for preserving, archiving, and promoting access to biomedical data. We do ask that NIH provide examples of “repositories that make scientific data available at no cost for extended periods of use.” We believe that the best way to maintain the data will be in Federally-owned repositories, perhaps at the NLM, to enforce consistency in the storage of the data. Leaving this to individual research groups or institutions invites too much variability, and

too much chance that the data will not be adequately maintained. Growing costs of data management are a leading challenge for implementing data sharing and management plans. Cloud-based data solutions will be effective in maintaining and sharing large datasets for long periods of time. Mechanisms to determine the usability of data and how long to store it should be explicitly outlined. Additionally, cost associated with maintaining data should also be considered as part of the broader plan.

Attachment:

In Response to Solicitation No.:
NOT OD 19 014
NIH Office of the Director

**Draft Data Management
and Sharing Policy for
NIH Funded or
Supported Research**

Request for Information

December 10, 2018

SUBMITTED TO
Office of Science Policy (OSP)
National Institutes of Health
6705 Rockledge Drive, Suite 750
Bethesda, MD 20892

SUBMITTED BY
Booz Allen Hamilton
One Preserve Parkway
Rockville, MD 20852

Company Business Size: 23,000
Company Business Type: Large
DUNS: 006928857

This proposal includes data that shall not be disclosed outside the Government and shall not be duplicated, used, or disclosed in whole or in part for any purpose other than to evaluate this proposal. If, however, a contract is awarded to this offeror as a result of or in connection with the submission of this data, the Government shall have the right to duplicate, use, or disclose the data to the extent provided in the resulting contract. This restriction does not limit the Government's right to use information contained in this data if it is obtained from another source without restriction. The data subject to this restriction are contained on all sheets.

Booz | Allen | Hamilton

Booz Allen Hamilton Inc.
One Preserve Parkway
Rockville, MD 20852
Tel 1-301-838-3600
Fax 1-301-838-3606

December 10, 2018

Carrie D. Wolinetz, Ph.D.
Associate Director for Science Policy
Office of Science Policy (OSP),
National Institutes of Health,
6705 Rockledge Drive, Bethesda, MD 20892

Dear Dr. Wolinetz,

Booz Allen Hamilton (Booz Allen) is pleased to submit our response to Request for Information for The Draft Data Management and Sharing Policy for NIH Funded or Supported Research, Notice Number: NOT-OD-19-014.

Please feel free to reach to me at 301-838-3647 or samavedam_rajni@bah.com with any questions.

Sincerely,

Rajni Samavedam

Rajni Samavedam
Principal

BOOZ ALLEN HAMILTON

Introduction

The availability of large, well-curated shared scientific, biological, clinical, and healthcare datasets permits secondary usage of these data to further inform the field, optimize healthcare systems, and customize individual patient care. Open data can shed light on the etiology of diseases, improve diagnosis, facilitate both discovery and validation science, and accelerate bench-to-bedside translation, thereby increasing return on investment (ROI). Because these datasets exist in many different formats that are often not easily shared, findable, or interoperable, accessing these data can be difficult. Additional barriers include concerns over privacy, lack of transparency and confidentiality, and securing control of PHI/PII data. These concerns are important when implementing strategies to link, harmonize, retain, retrieve, and share individuals' healthcare data. The *Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research (Proposed Provisions)* will facilitate data sharing and access through a balanced agenda that safeguards personal information while enabling access and use of data for healthcare research purposes. Per NIH solicitation# NOT-OD-19-014, Booz Allen provides the following feedback on proposed key policy provisions and elements for data managements and sharing plans.

Definition of Scientific Data

We support the general definition of 'scientific data' as given in the *Proposed Provisions*. This definition should be considered to include both raw and processed data as well as metadata describing not only the data, but the methods used to analyze the data, experimental samples, experimental methods, software/code, and statistical findings. We recommend that the *Proposed Provisions* also include additional context to the explicit exclusions listed in the definition for scientific data: "laboratory notebooks, preliminary analyses, completed case report forms, drafts of scientific papers, plans for future research, peer reviews, communications with colleagues, or physical objects, such as laboratory specimens." Objects such as these may hold valuable historical, legal, or other information that may be used for research on the biomedical research enterprise as well as NIH's organization, operations, and policies. In addition, documents that reflect activities of the NIH or interactions of the NIH with the public are Federal records (44 U.S.C. 3301; 36 CFR 1222.10). We recommend including text indicating that Federal records of NIH's activities (e.g., laboratory notebooks, meeting minutes), while not deemed 'scientific data' per the *Proposed Provisions*, are nevertheless valuable and retained for a period of time in accordance with the NIH's Extramural Research Records Schedule, Intramural Research Records Schedule, or by determination of the National Archives & Records Administration (NARA). Additionally, statistical data/statistical features (e.g., probability distributions, regression analysis) associated with the scientific data should also be included in the proposed definition of "metadata" and shared with the broader community.

Requirements for Data Management and Sharing Plans

Data Attribution: We also agree with the underlying premise of the *Proposed Provisions* – and of NIH's data sharing policies in general – that the data produced by NIH-funded or -supported researchers is valuable to scientific community at large. The value of these data should also be recognized for the individual scientists or research teams who created them. However, scientific credit currently accrues to the authors of peer-reviewed journal articles, not necessarily to the providers of previously-generated data, experimental samples, software algorithms, or other scientific and analytical contributions. Ensuring traceable data and resource authorship will promote sharing of the full spectrum of scientific data. In addition, retaining details on the origin of digital sequence information (DSI) may be needed by institutions as they seek to comply with the *2010 Nagoya Protocol on Access to Genetic Resources & the Fair & Equitable Sharing of Benefits Arising from Their Utilization*. Thus, the proposed data sharing and management plans should include details of data and resource authorship and/or provenance. One example of a flexible and interoperable data annotation scheme is NICHD's Data and Specimen Hub (DASH). DASH stores all data as "digital objects" and each is annotated with structured, standardized metadata attributes suited to the type of object (e.g., datasets, documents, images). These annotations (metadata values) are derived from established standardized resources (e.g., Clinical Data Interchange Standards Consortium, NCI Thesaurus) whenever possible and presented to DASH users as a code list to select during annotation process.

Data Type: We agree that descriptions of data type should include information on the data modality (e.g., imaging, genomic, mobile, survey); level of aggregation (e.g. individual, aggregated, summarized); degree of data processing that has occurred (i.e., how raw or processed the data will be); the rationale for sharing or preserving data; and, for human-derived data, a description of the measures planned (e.g., de-identification, data aggregation). Because the storage of scientific data requires spaces and resources, we suggest that NIH consider that estimated amounts of data should also be expressed in terms of needed disk space in addition to estimates of the expected gene variants, cases, or study participants.

Related Tools, Software and/or Code: When referencing the software/computer code used to process and/or analyze the data, it is important to include both the code and the values of any variable parameters used when running that code, as well as the details of what platform on which the processing or analysis was done. Both platform and parameter settings can impact the reproducibility of the results. In addition, because open source tools, imputation pipelines, and algorithms that parse data or related software and code are often research outputs and “necessary to validate and replicated research findings,” we believe that they fall within the definition of scientific data and should therefore be deposited into a repository such as GitHub, with accompanying SOPs, user guide or “how-to” files, to ensure proper archiving, access and use by the scientific community.

Data Preservation, Access, and Discoverability: Current data-resource ecosystems tend to be “siloeed” and are not optimally integrated or interconnected. An ecosystem of repositories is only possible through the use of common data elements (CDEs) that serve as a Rosetta Stone to existing data management resources and tools and facilitate development of new ones. Unfortunately, CDEs are not yet fully adopted. NIH should continue to encourage and actively promote the use of CDEs as well as assess the feasibility of making CDE use mandatory. The National Library of Medicine (NLM) can serve as a central organizational hub that implements CDEs while transparently regulated, third-party run, secure, and interoperable enterprise data solutions may further support access to data to authorized parties in machine readable formats, such as through APIs. However, UUIDs (Universally Unique Identifiers) should also be used to enhance discoverability of the data as a persistent unique identifier.

Compliance and Enforcement: To encourage compliance, NIH should consider holding back a fraction (~5%) of funds and releasing them once the approved data sharing plan is fully implemented and data is completely shared. In order to determine if institutions are in compliance with their data sharing obligations, we suggest the plans also include additional details and guidance to assist in the determination of whether “the data sharing repository meets community-based standards at the time of deposition,” such as: use of CDEs; efforts to ensure data in the repository are secure and interoperable; the provision of APIs and software tools that facilitate storing, managing, standardizing and publishing data; linking of data to articles and software tools; inclusion of a data dictionary defining all metadata and data fields associated with the datasets to be stored; and contingency plans in the event that a repository becomes unavailable. Plans should be evaluated for compliance at least annually and must be a requirement for funded projects and contracts. The funded academic and industrial stakeholders should provide annual reports on the status of data sharing and management as outlined in plan.

Data Preservation and Access Timeline: NIH will need to set reasonable standards for when data is to be made available to secondary data users and how long it should be preserved. While the data generators should have rights of first analysis, it is unreasonable to keep the data from being made public for a year or more after deposition. Deposition of the data must also be made in a timely fashion after completion of the research that generates the data. It will be difficult to estimate how long the data should be preserved, due to factors such as how well the data ages in terms of the technology used to produce it, the cost of generating the data, how difficult recreating similar data would be, the expected usage of the data over time, and of course, the cost of preserving the data and accessing it over time.

Scientific Data Archiving: We are heartened to see that the NLM is working with the National Academies of Sciences, Engineering, and Medicine to conduct a study on forecasting the long-term costs for preserving, archiving, and promoting access to biomedical data. We do ask that NIH provide examples of “repositories that

make scientific data available at no cost for extended periods of use.” We believe that the best way to maintain the data will be in Federally-owned repositories, perhaps at the NLM, to enforce consistency in the storage of the data. Leaving this to individual research groups or institutions invites too much variability, and too much chance that the data will not be adequately maintained. Growing costs of data management are a leading challenge for implementing data sharing and management plans. Cloud-based data solutions will be effective in maintaining and sharing large datasets for long periods of time. Mechanisms to determine the usability of data and how long to store it should be explicitly outlined. Additionally, cost associated with maintaining data should also be considered as part of the broader plan.

Submission #170

Date: 12/10/2018

Name: James Luther

Name of Organization: Duke University

Type of Organization: University

Role: Institutional Official

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

Clinical and Basic Science

Attachment:

December 10, 2018

Office of Science Policy
National Institutes of Health
6705 Rockledge Drive, Suite 750
Bethesda, MD 20892

Subject: NIH Request for Information: Proposed Provisions for a Draft Data Management and Sharing Policy

Comments in response to notice number NOT-OD-19-014

On behalf of Duke University, I am pleased to provide comments on the National Institutes of Health "*Request for Information: Proposed Provisions for a Draft Data Management and Sharing Policy*." As one of the leading research institutions in the United States, Duke welcomes this opportunity to comment on concerns regarding the potential financial and administrative burdens of policies requiring data management plans and long-term storage.

We thank the NIH for their continued interest and willingness to accept input from universities and the associations and organizations that represent us. In particular, we would like to take this opportunity to endorse and support the December 10, 2018 memo submitted by the Association of American Universities (AAU), the Association of Public and Land-grant Universities (APLU), and the Council on Governmental Relations (COGR). We would also like to affirm the comments submitted by other members of the Duke research community, including the Duke University Libraries, Research Integrity Office, and the Office of Information Technology which are being submitted under separate cover. The intent of this submission is to provide additional context for costing and administrative items of critical importance.

As an institution, Duke has a demonstrated commitment to internal resources for data management, as well as consistent engagement in the national conversation surrounding these issues. As the submission from Duke University Libraries demonstrates, Duke commits significant financial and human resources to data stewardship and administrative costs incurred therein. Furthermore, Duke has joined fellow institutions in developing and establishing data depositories and other networks aimed at encouraging Investigators to make their data publically accessible in a secure manner. As a representative of the University, I have served on working groups, panel discussions, and engaged directly with federal partners.

Therefore, while we support the overall mission of data management and sharing, we must reinforce the importance of addressing costing and administrative burden issues that such a policy may create if implemented without thorough consideration of the potential for unintended consequences. These concerns can be grouped into three areas of need: harmonization of approaches across sponsoring agencies, a determination on the allowability of costs related to

data management as direct or indirect charges, and clarification within the award documents of length of availability and access concerns.

The NIH should seek to harmonize their policies, documentation and compliance approach across NIH institutes and with other federal agencies to the greatest extent possible to minimize the administrative requirements and costs for both agencies and funding recipients.

As you know, there is a rapidly evolving landscape for data management as sponsors, funding agencies, institutions, and publishing entities actively modify and shift their policies. This uncertainty causes challenges for planning and allocation of resources on the part of the institutions. Therefore, we recommend that any proposed policy on data management should recognize the various policies that already exist and work with the established frameworks rather than create contradictory or conflicting policies.

In order to achieve this, the NIH should welcome further coordination and conversation with fellow federal agencies and institutions of higher education. As an example, the National Science Foundation sponsored an October 2018 workshop to engage substantively on these issues with members of the research community. The NIH should hold a similar workshop to hear, in person, the challenges faced by research institutions in this area and support full collaboration with universities and faculty.

Costing aspects of data management plans will have significant budgetary implications. Therefore, the NIH should allow these costs to be directly charged to the award if data management plans are to be included in award proposals.

In January of 2017, Duke created and filled four full-time dedicated Research Data Specialists and Repository Ingest Specialists to support data management planning, compliance, public access and retention requirements. Furthermore, the institution made a commitment of funding to support baseline, minimum levels of computing and digital storage for research projects. This financial and personnel investment demonstrates the critical importance of increasing public access to scientifically generated data in a secure, appropriate manner. However, the NIH should be aware that the effort required to build, maintain, and coordinate this storage represents added administrative burden and increased cost for the investigators and their institutions.

We therefore feel it is of critical importance that the regulations provide institutions the ability to recover costs associated with these expanding requirements.

Award documents should specify requirements for the length of availability and access concerns.

By providing these parameters within the award documents, the NIH would be providing enough certainty for researchers to appropriately budget for costs and address embargoes or restrictions placed on the data. These requirements should reflect the specifications of the science being conducted and allow for adaption based on any privacy or embargo concerns. Furthermore, including these parameters will allow for compliance and oversight as the NIH will have a concrete marker by which to evaluate management plans that provides structure for the recipient.

On behalf of Duke University, I wish to express again our most sincere appreciation for the opportunity to provide thoughtful comments regarding this essential initiative. Please feel free to contact me at any time for further discussion.

I would be willing to meet at any time to continue discussions relating to costing considerations. Thank you for your efforts on behalf of the research community.

Sincerely,

A handwritten signature in cursive script that reads "James D. Luther".

James D. Luther
Associate Vice President
Research Costing Compliance Officer

Cc Dr. Lawrence Carin, Vice Provost for Research
 Tracy Futhey, Vice President Information Technology and Chief Information Officer
 Tim McGeary, Associate University Librarian for Information Technology Services
 Dr. Geeta Swamy, Vice Dean and Assistant Vice Provost for Scientific Integrity
 Dr. Raphael Valdivia, Vice Dean for Research
 Tim Walsh, Vice President Finance

Submission #171**Date:** 12/10/2018**Name:** Mary Piorun, Ph.D., Director, Lamar Soutter Library**Name of Organization:** Lamar Soutter Library, University of Massachusetts Medical School**Type of Organization:** University**Role:** Institutional Official**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

RNA interference, gene therapy, gene function and expression, systems biology, infectious disease, neurotherapeutics, clinical, and translational science and research.

I. The definition of Scientific Data

We agree with the definition of scientific data as proposed, with the exception that we feel for the purposes of a Policy, scientific data includes individual level and summary data, as well as metadata. As written, the definition states that it may include such. Without this additional data, including metadata, the scientific data shared will fail at meeting the guidelines of the FAIR principle (findable, accessible, interoperable and reusable) for scientific data management.

II. The requirements for Data Management and Sharing Plans

The proposed policy statement lacks any mention of (1) how NIH will know if the data sharing requirements have been met and, related, (2) data citation. Will a system that generates something similar to a PMCID be created? Will an identifier, e.g. a digital object identifier, be assigned to data sets so that others can both locate the data and attribute the creators of it accordingly? We feel these are significant gaps in the Policy at this point.

Regarding the archiving of data, we agree with encouraging researchers to deposit in no-cost repositories, however there is limited detail in how the situation will be addressed if no such repository exists. This is likely a question that researchers will have and thus is important to state from the beginning.

The creation of a checklist for preparing a data set is helpful. Providing details as to the minimal requirements for data will make it easier for researchers and their staff to prepare the data throughout their research process. We suggest the NIH's policy include a strong working connection with institution's Offices of Sponsored Research and libraries, in order to assure compliance with the proposed Data Management and Sharing Plan's policy and procedures.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

We view issues related to the release and implementation of the Policy through the historical lens of the NIH Public Access Mandate, meaning lessons learned from instituting the Mandate can well-inform the same for the Policy. A phased approach, while giving researchers and administrators time to understand the issues and create the necessary new steps within their research process to share their data, also lacks the incentives necessary for adoption. As experienced in the Public Access Mandate, this type of approach resulted in confusion and delays in researchers' compliance. Expectations of researchers should be clearly stated and enforced from the beginning, rather than incremental steps along the way.

As proposed, making NIH-funded research data available "in a timely manner" is vague and unhelpful for anyone seeking to comply with the Policy. Be specific regarding the time allowed between generation of data, publication, and the release of the data for public use. Embargos are certainly fair, but they need to be defined clearly. Similarly, acceptable places for data deposit need to be stated, in fact, we feel it would be helpful to researchers, administrators, and those assisting with the process, if the Policy offered a list of archival entities available. Some are surely already aware of discipline-specific repositories where data can be deposited, but not everyone will.

Attachment:

Research area most important...

RNA interference, gene therapy, gene function and expression, systems biology, infectious disease, neurotherapeutics, and translational science and research

The definition of scientific data:

We agree with the definition of scientific data as proposed, with the exception that we feel for the purposes of a Policy, scientific data *includes* individual level and summary data, as well as metadata. As written, the definition states that it *may* include such. Without this additional data, including metadata, the scientific data shared will fail at meeting the guidelines of the FAIR principle (findable, accessible, interoperable and reusable) for scientific data management.

The requirements for Data Management and Sharing Plans:

The proposed policy statement lacks any mention of (1) how NIH will know if the data sharing requirements have been met and, related, (2) data citation. Will a system that generates something similar to a PMCID be created? Will an identifier, e.g. a digital object identifier, be assigned to data sets so that others can both locate the data and attribute the creators of it accordingly? We feel these are significant gaps in the Policy at this point.

Regarding the archiving of data, we agree with encouraging researchers to deposit in no-cost repositories, however there is limited detail in how the situation will be addressed if no such repository exists. This is likely a question that researchers will have and thus is important to state from the beginning.

The creation of a checklist for preparing a data set is helpful. Providing details as to the minimal requirements for data will make it easier for researchers and their staff to prepare the data throughout their research process. We suggest the NIH's policy include a strong working connection with the Office of Sponsored Research and institution's library, in order to assure compliance with the proposed Data Management and Sharing Plan's policy and procedures.

The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards.

We view issues related to the release and implementation of the Policy through the historical lens of the NIH Public Access Mandate, meaning lessons learned from instituting the Mandate can well-inform the same for the Policy. A phased approach, while giving researchers and administrators time to understand the issues and create the necessary new steps within their research process to share their data, also lacks the incentives necessary for adoption. As experienced in the Public Access Mandate, this type of approach resulted in confusion and delays in researchers' compliance. Expectations of researchers should be clearly stated and enforced from the beginning, rather than incremental steps along the way.

As proposed, making NIH-funded research data available “in a timely manner” is vague and unhelpful for anyone seeking to comply with the Policy. Be specific regarding the time allowed between generation of data, publication, and the release of the data for public use. Embargos are certainly fair, but they need to be defined clearly. Similarly, acceptable places for data deposit need to be stated, in fact, we feel it would be helpful to researchers, administrators, and those assisting with the process, if the Policy offered a list of archival entities available. Some are surely already aware of discipline-specific repositories where data can be deposited, but not everyone will.

Submission #172

Date: 12/10/2018

Name: Melissa Haendel

Name of Organization: Oregon State University

Type of Organization: University

Role: Scientific Researcher

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

Rare disease, informatics, data science, semantic data integration, open science, reproducibility, team science, disease classification, data standards, clinical data management, clinical informatics, data quality, genomic medicine

Attachment:

On behalf of the Monarch Initiative and the Center for Data to Health, and as experts in the integration and sharing of public datasets and clinical data management, we provide the following response to the Request for Information on Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research Notice Number: NOT-OD-19-014.

Melissa Haendel, Center for Data to Health, Monarch Initiative

Anne E Thessen, Monarch Initiative

Julie McMurry, Monarch Initiative

Kristi Holmes, Center for Data to Health

Harold Solbrig, Center for Data to Health

Bill Hersh, Center for Data to Health

Monica Munoz-Torres, Monarch Initiative

Response:

I. The definition of Scientific Data

The draft definition of “Scientific Data” is too narrow in scope because it refers only to data collected for the primary purpose of science. It is important to recognize the increasing amount of data used in scientific research that originates from sources outside experiments and observations. Examples include real-world measurements of health and disease, which include everything from data routinely collected by mobile or wearable devices to electronic health records (EHRs). Numerous other examples exist, such as social network data, research metrics, community health data, and crowdsourced knowledge generated knowledge (such as in WikiData). Of course, special considerations are needed for personal and private information, such as raw EHR data, genomic data, or even our Google search histories - included here to reinforce that they too are “scientific data.”

Nevertheless while we believe that the distinctions between “scientific vs. non-scientific” and “data vs knowledge” are not the most meaningful, we also understand that the guidelines should be scoped appropriately. Therefore we propose the term “publicly-funded data resources”. This designation would be both clearer at face value and more comprehensive--including not only raw data, but also derived data, knowledge, and tools such as software and algorithms--whether or not these resources originate from professional scientists. Any data or data resources generated with support from public funds should be freely accessible and repurposable by the public. Our proposed designation is likely to be more easily enforced and will help awardees make good decisions about management and dissemination.

II. The requirements for Data Management and Sharing Plans

The basic components of a DMP. Resources are finite; perfunctory sharing of poorly-documented data does not achieve the objectives of the spirit of FAIR. However, we also recognize that not all data can be held to the same standard. Wherever practicable, data should follow good data practice; for example, data should be published together with methods and factors known to impact reliability. The full “chain-of-custody” should be documented from the generation of data

through its safeguarding and analysis. Not only does this provenance help the users of data assess its veracity--but it also helps ensure attribution, whether to scientists or to members of the public.

Data management and sharing plans should describe a reliable, consistent, and well understood mechanism for recording data and its associated metadata. Supporting “knowledge services” should be described that provide the appropriate shared vocabularies, ontologies and reference information for documenting data. Knowledge services should include information about units, representational forms, scientific methods, domain knowledge, organizations, researchers and ontologies that define the methods and knowledge of the domain itself. This service must be made freely available, reliable and subject to independent verifiability and community correction. A description of the search and retrieval services should also be required, to allow researchers to query, discover and utilize data. The retrieval service must include both the data and its associated pedigree (see attribution, provenance, and reproducibility below) and must provide a mechanism that allows the provenance of derived and enhanced data to be re-entered into the data management plan faithfully and traceably.

Assessment of quality.

Data Management and Sharing plans are required for any grant proposals over 500K direct costs/year; however, **making them required and scored for all grants is one of the single most impactful changes that the NIH could make.**

Moreover, we believe that this section should be scored as part of the review criteria. Because most people on NIH review panels are expert in the given specific area of science and not in data management or open science, it would be beneficial for NIH to seek out this expertise for the review panels. It should be noted that often the professional profile of open science experts expands outside the usual background of more typical investigators with R01 funding, and as such, a different approach and criteria for selecting these experts should be created. This may be true for other areas of the review panel as well, but our RFI response is focused exclusively on the Data Management and Sharing plans.

Execution of the data management plans. Throughout the life of a funding award, the data provider/investigator should be evaluated for adherence to the data management and sharing plans by external expert reviewers. Data management and sharing plans should be versioned and updated as the science, community technology, and standards evolve. To encourage investigators to outline effective data management and sharing plans, funders should incentivize grant applicants to include funding allocations within their grant applications to establish partnerships with teams developing standards or data management systems. This also requires an increase in the overall award to reflect the added burden of data management. Otherwise the data management and sharing plan becomes an unfunded mandate. As well, there should be consequences for project leaders who do not abide by the spirit of their proposed plans.

Considerations regarding data reusability. We urge NIH to consider attributes of data management and sharing that most limit the actual reusability of data. While the FAIR principles have helped to promote awareness, these don't provide guidance on how to actually make data

more reusable and therefore useful. We have written extensively about factors that maximize data reusability, specifically for data integrators and third party users. For example, in response to RFI NOT-OD-16-133 (Metrics to Assess Value of Biomedical Digital Repositories), we wrote about FAIR-TLC, where the T is Traceability, L is licensure, and C is connectedness. <https://doi.org/10.5281/zenodo.203295>. There are numerous specific recommendations in this RFI that are relevant here, but we therefore refer the reader to that other document. Finally, data should be openly available for use, in both human and machine readable formats.

Data licensing and data use agreements. We recommend that the NIH mandate that all publicly funded sources of data, knowledge, or tools be documented with a clearly defined and preferably standard licence. [The \(Re\)usable Data Project](#) has evaluated the licensing of 56 NIH publicly funded data resources (including some NIH sources), and has illuminated the fundamental barriers to data science due to a lack of ability to mash-up and redistribute these data. A preprint describing the findings is here: [A Measure of Open Data: A Metric and Analysis of Reusable Data Practices in Biomedical Data Resources](#). This inability to redistribute seemingly “open” data sources limits their potential impact, not only for discovery, but also in for patient care. This problem is so significant, the community wrote a letter to Francis Collins, “[Request for Community partnership in data resource licensing planning](#)”.

A change in the Data management and sharing plan seems ideal timing for coordinated improvements to licensing across NIH funded resources. **We believe that licensing is one of the most important and overlooked requirements needed to make data reuse a reality. A license MUST be required in all DMPs.**

Identifiers. All genes, proteins, variants, phenotypes, diseases, chemicals, species, biosamples etc. should be referenced using appropriate vocabularies/ontologies/reference data. The scientists that generate and publish data are those best positioned to properly annotate it; they should do so using community best practices. Provisioning and management of identifiers should be detailed in the data management and sharing plan; a three-year community coordination effort led to some agreed-upon best practices published in PloS Biology: “[Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data.](#)” We would highly recommend any data creator or publisher be directed to these best practices or contact the authors for assistance. Persons should be referenced by an identifier, with ORCID being the most prominent and most integrated person-level identifier in use. Moreover, organizations should also be referred to by a persistent identifier, with efforts currently in progress to create an open [Research Organization Registry \(ROR\)](#). Resolvers (examples are N2T.net and identifiers.org) should be used to persistently redirect identifiers where they are made public, and details to avoid link-rot should be included in the management plan. Documentation of identifier provisioning, schema, and examples are often lacking, and we would recommend that this be a required component.

Attribution, provenance, and reproducibility. Reproducible science depends in part on knowing what has been specifically performed and by whom. Not only does this hamper reproducibility and evidence for scientific conclusions, it also disincentivizes diverse types of contributions. A data management plan should also document how it plans to include attribution, provenance of

the data, and to address any reproducibility aspects (including identification of primary resources, see [“On the reproducibility of science: unique identification of research resources in the biomedical literature”](#)). The goals should be to (1) to give credit where credit is due for everyone - regardless of their discipline, title, or contribution, regardless of whether it fits into narrowly defined construct of success; (2) to enable linking of all research outputs created during a study together, enabling access to data and results (even negative results), tools, and ideas which often remain invisible because they do not appear in a published manuscript, as access helps drive discovery; (3) facilitate re-use of the information in a variety of ways by all stakeholders (individuals, publishers, scholarly organizations, data repositories, and funding agencies). For example, relationships between people and their products/activities can be used to track research trends; to understand and leverage influences or projects; to promote collaboration and team formation; as recommender systems for scholarly products or methodologies; and to present a complete record of results and research outputs. Fundamentally, the data about the contributions that scholars make should be as open as the data and resources themselves if we really aim to incentivize sharing and open science. As part of the NCATS Center for Data to Health (CD2H), the CD2H has completed a number of activities to aid in supporting improved attribution, including the development of a data model and implementation into existing research systems, both with community and stakeholder input and coordination.

Sustainability. While not every DMP need include a business plan, there should be a requirement to address persistence of the data and access to it, as well as the financial support required to do so. DMPs for data about people also need to include indigenous rights as well as the confidentiality of the person, and the sustainability and maintenance for governance rights.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards.

The [European Commission Expert Group on FAIR Data have developed a series of recommendations](#) to define, implement, and sustain an appropriate ecosystem to support many of the concepts outlined in our response. This phased plan takes an opportunity to define broadly the concept of “digital object” to include data, software, and other research resources, much as we have taken an inclusive approach to our own definition of data (above). In the EC plan, digital objects are stored in standard formats, accompanied by persistent identifiers, metadata, and documentation. For maximum benefits to be realized, an ecosystem which leverages strong policies, data management plans, persistent identifiers, standards, repositories, and culture is needed. The technical and cultural recommendations in the EC report present an excellent discussion on timing and phases, both technical and cultural, for consideration.

Submission #173

Date: 12/10/2018

Name: Laura Quilter

Name of Organization: University of Massachusetts Amherst, Libraries

Type of Organization: University

Role: Institutional Official

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

data management and preservation; open access

I. The definition of Scientific Data

In our attached letter, we recommend approaches to handling negatives results and metadata, flag ambiguity in the recommendation to "digitize all scientific data", and recommend further investigation into the question of preservation of laboratory notebooks.

II. The requirements for Data Management and Sharing Plans

In our attached letter, we make recommendations about the plan review and evaluation and plan elements sections.

Attachment:



UNIVERSITY of
MASSACHUSETTS
W.E.B. Du Bois Library
Amherst, MA 01003-9275

December 10, 2018

Office of Science Policy (OSP)
National Institutes of Health
6705 Rockledge Drive, Suite 750
Bethesda, MD 20892

Submitted electronically to:

<https://osp.od.nih.gov/provisions-data-managment-sharing/>

RE: Notice Number: NOT-OD-19-014 "Request for Information (RFI) on Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research"

Dear Office of Science Policy, Office of the Director, and Acting Director Bonham,

The University of Massachusetts Amherst Libraries write in response to the Request for Information (RFI) on Proposed Provisions for a Data Management and Sharing Policy for NIH Funded or Supported Research.¹

The University of Massachusetts Amherst Libraries maintain an institutional repository, "ScholarWorks", at <https://scholarworks.umass.edu/data/>. We currently host at least 68 datasets, deposited since our repository began accepting data in 2016. The Libraries' Data Working Group was established in 2010 and has regularly contributed feedback on data management plans, provided data consultations, and instructed scholars in best practices for data management. To build on this work, the Libraries hired a dedicated Data Services Librarian (Thea Atwood, one of the authors of this comment) in 2017 to improve data management capacities and competencies across campus. In that capacity, she works with research groups across campus, as well as interfacing regionally and nationally.

As described in the RFI, the NIH seeks comments on key provisions for a future policy for the management and sharing of data, to replace the 2003 NIH Data Sharing Policy.² Specifically, NIH requests public comment on I. The definition of

¹ Request for Information (RFI) on Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research, October 10, 2018, available at <https://grants.nih.gov/grants/guide/notice-files/NOT-OD-19-014.html>.

² Final NIH Statement on Sharing Research Data (February 26, 2003), available at <https://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html>.

Scientific Data, II. The requirements for Data Management and Sharing Plans, III. The optimal timing for implementation, and any other relevant topic. We write on topics I. and II. and offer some further suggestions

I. The definition of “Scientific Data”.

The proposed definition is:

Scientific Data: The recorded factual material commonly accepted in the scientific community as necessary to validate and replicate research findings including, but not limited to, data used to support scholarly publications. Scientific data do not include laboratory notebooks, preliminary analyses, completed case report forms, drafts of scientific papers, plans for future research, peer reviews, communications with colleagues, or physical objects, such as laboratory specimens. For the purposes of a possible Policy, scientific data may include certain individual level and summary or aggregate data, as well as metadata.[7] NIH expects that reasonable efforts should be made to digitize all scientific data.

[7] NIH Policy on the Dissemination of NIH-Funded Clinical Trial Information (September 16, 2016) <https://grants.nih.gov/grants/guide/notice-files/NOT-OD-16-149.html>

We appreciate the thoughtfulness of this definition, which carefully describes the material to be included in broad, functional terms, and equally carefully excludes specific types of communication in appropriately narrow and discrete terms. We have three comments regarding this proposed definition, as well as a suggestion for further study.

First, we recommend including specific reference to **negative results** as potential subjects of inclusion, possibly in the next to last sentence. This sentence could read:

For the purposes of a possible Policy, scientific data may include certain individual level and summary or aggregate data, metadata, or relevant unpublished data demonstrating negative results.

As many scientists have observed³, the lack of access to negative results and null data may result in duplicative or misdirected research, distorting assessment of scientific research and hindering the progress of research. The proposed definition properly applies to material “including, but not limited to” data supporting publications, which can be interpreted as including negative data. However, without express mention of data *not* included in publications, such as negative results / null data, researchers will not be prompted to consider preservation of this data. Indeed, due to current conventions of publishing only positive data, researchers may in many cases not feel it appropriate to include negative data without express “permission” to

³ See, e.g., Mlinaric et al, “Dealing with the positive publication bias: Why you should really publish your negative results,” *Biochemia Medica (Zagreb)* 2017 Oct 15; 27(3):030201; Weintraub, “The Importance of publishing negative results,” *Journal of Insect Science* 2016, 16(1): 109; Matosin et al, “Negativity towards negative results: A discussion of the disconnect between scientific worth and scientific culture,” *Disease Models & Mechanisms* 2014 Feb., 7(2): 171-173; and Sandercock, “Negative results: Why do they need to be published?” *International Journal of Stroke* 2012 Jan; 7(1):32-33.

do so.

Second, we recommend additional detail about the potential “**metadata**” requirements. In particular, the final policy should specify that metadata dictionaries or codebooks, if used or developed, should be made available to the public on similar terms, either published separately or as materials and methods, or included within the data repository for a given proposal.

Third, we note that there is significant ambiguity in the expectation that reasonable efforts should be made to “digitize all scientific data.” The modifier “all” in front of the phrase “scientific data”, in the context of a definition of scientific data, opens this sentence up to further interpretation: For instance, does “all” mean the kinds of scientific data that were not included in the definition? Moreover, what standards are implied by “digitiz[ing]”? Scanning text without character recognition, for instance, provides an exceedingly minimal and not very useful amount of digitization. Images may be digitized at high quality or low quality, with significant differences in usability. Digital formats may be encrypted, non-standard, or sui generis, and in other ways “digital” but not necessarily useful. We recommend clarifying this language with functional qualifications; for instance, “digitize ... in open formats that are usable by researchers” or “digitize ... in formats that are broadly available.”

Finally, we recommend further investigation of the question of preservation of **laboratory notebooks**. We agree that laboratory notebooks are properly excluded from the definition of “scientific data” in the “Data Management and Sharing Policy.” Because laboratory notebooks are idiosyncratic and may include material from many projects, as well as draft, confidential, and non-scientific content, it is inappropriate to treat laboratory notebooks as subject to NIH’s open data policies.

However, laboratory notebooks often include data, procedures, fundamental research techniques, and observations vital to reproducing research findings. The preservation of laboratory notebooks is therefore of high concern to scientists, both as individual scientists and as managers of laboratories and principal investigators of research projects.

Unfortunately, the current proliferation of electronic laboratory notebook software, as well as the use of non-dedicated software including wikis, word processing, spreadsheet, and cloud-based storage, raise concerns about long-term preservation and access to lab notebooks. The very multiplicity of options raises concerns, as the “lab notebook” environment in many laboratories is fracturing across multiple formats and styles, leaving laboratories without consistent lab notebook data. We recommend, therefore, that the NIH study and develop standards for local preservation and retention of lab notebooks. Scientists need assistance in understanding how to maintain and preserve their own lab notebooks into the future, and the progress of science will suffer without attention to this detail. This is a task that is well-suited to the broad perspective of a national-level agency.

II. The requirements for Data Management and Sharing Plans.

PLAN REVIEW AND EVALUATION

The proposal to incorporate plan review and evaluation into **Contracts**, after technical evaluation performed by NIH staff, offers the best opportunity for consistent and high-quality assessment. NIH staff would be able to develop relevant expertise in preservation and access, and could connect with peers at other agencies in developing relevant standards and guidelines for repositories, and procedures for preservation, retention, data migration, and other data management policies.

Based on our work with faculty researchers, we are not persuaded that extramural grant reviewers would be consistently well-positioned to assess the acceptability of Data Management and Sharing Plans when reviewing proposals. This is not their area of expertise, generally, and assessment of these plans may well suffer by comparison with assessment of science. We believe, therefore, that assessment of the plans is more appropriately considered part of the technical review. However, extramural reviewers have a key role to play in review and assessment of repositories.

Incorporation of review of plans into reviews by the Scientific or Clinical Director may offer an additional level of review, but should not substitute for a consistent review by staff with relevant expertise.

Finally, we note that evaluation of individual plans in context of other funding / support agreement mechanisms may be appropriate in some cases but should not substitute for routine assessment by staff with appropriate expertise. Some plans, however, may warrant additional review. For instance, proposals to establish a new repository or new method of access may reasonably benefit from *additional review* over and beyond the technical evaluation performed by NIH staff.

PLAN ELEMENTS

Plan Elements Section 2 – Related Tools, Software and/or Code.

Computational methods for generating data must be preserved, and presently, software programs and scripts are made available at any number of private repositories (such as GitHub), which can close at any time. We note that by comparison, in Plan Elements Section 3, the NIH has established a “Common Data Element Resource Portal”. We recommend development of a “Common Software Tools Resource Portal” to connect to significant or recommended software repositories. We further recommend assessment of the feasibility of developing an NIH-based software repository, or a software repository developed in conjunction with other major funding agencies, such as the NSF.

Plan Elements Section 4. Data Preservation and Access.

We recommend a working definition of “preservation” be included.

We also recommend development of standards for repositories, adoption of privately-developed standards or audits for repositories, or guidance to researchers on how to assess repositories. Researchers are currently not well-prepared to assess data security and management of repositories, including data migration, data security practices, and compliance with applicable data laws and standards. While the University of Massachusetts Amherst offers consulting services to our researchers, many institutions are not as well-positioned. NIH-developed guidance will benefit all scientists, particularly those without institutional data management consulting services.

As previously mentioned, staff doing technical reviews for data plans are well-positioned to collaborate with peers in developing or adopting standards for repositories, assessing repositories for compliance with standards, and certifying them to researchers.

Plan Elements Section 5 – Data Preservation and Access Timeline.

We strongly believe that data preservation and access timelines are an integral part of data management plans. However, researchers rarely know how to assess the length of time data should be kept or made available. Some researchers would prefer to destroy data after the project is over, for instance, while other researchers rely on IRB standards whether or not appropriate in other circumstances. Researchers also rarely receive guidance on the appropriate length of time to distribute data. NSF offers some guidance, suggesting that data should be available “at time of article publication,” but there is little guidance available on sunseting data distributions.

We therefore agree that this section needs to be included, but recommend that it incorporate reference to additional standards and support from NIH or other federal agencies.

Plan Elements Section 6 – Data Sharing Agreements, Licensing, and Intellectual Property.

We recommend, first, that NIH include recommended sample licenses, such as Creative Commons, in Section 6.3.

Second, in section 6.2, we recommend that the NIH take this opportunity to return to the issue of Material Transfer Agreements (MTAs) that affect access to key reagents

necessary to reproduction of scientific results. As an initial matter, the NIH should encourage use of the Uniform Biological Material Transfer Agreement (UBMTA).⁴ However, this policy offers an opportunity to revisit the substance of the UBMTA, and consider appropriate limitations on attempts to own scientific data resulting from use of received material. The UBMTA was published in 1995, and in the ensuing two-plus decades, open access and open data principles have become standard – as demonstrated by this RFI.

III. Additional comments

Finally, we recommend that the NIH develop a repository of successfully funded Data Management Plans. Researchers too often develop these plans “in the dark,” particularly researchers from institutions that are under-resourced or do not have access to support staff within the libraries with an expertise in data management.

CONCLUSION

We thank you for the opportunity to comment on this important matter, and hope our comments prove helpful. Please feel free to contact us about any of our comments.

Sincerely,

Marilyn Billings, MLS
Director, Scholarly Communication Department

Thea Atwood, MLS
Data Services Librarian

Erin Jerome, PHD
Open Access and Institutional Repository Librarian

Laura Quilter, JD, MLS
Copyright and Information Policy Librarian

Submitted electronically by
Laura Quilter
LQuilter@umass.edu

⁴ Uniform Biological Material Transfer Agreement (UBMTA), *Federal Register*, March 8, 1995.

Submission #174**Date:** 12/10/2018**Name:** Elaine Martin, Julie Goldman**Name of Organization:** Countway Library of Medicine, Harvard Medical School**Type of Organization:** University**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Data Management services are a key component of our library at our organization. We believe required and maintained data management and sharing plans will accelerate research in all areas and the quality of research that is produced.

I. The definition of Scientific Data

The given definition cites Uniform Administrative Requirements (<https://www.federalregister.gov/d/2013-30465/p-834>). The term 'scientific' may be seen as limiting to some researchers who may interpret their data to fall outside this area; however it does make the distinction between 'administrative' data, for example. It may be important to increase the prominence 'metadata' has in the definition, making it very clear that metadata should always be included within scientific data.

II. The requirements for Data Management and Sharing Plans

Under Plan Review and Evaluation, there should be a place to address when a project funding source conflicts with NIH. For example, which funding requirements take precedence. Under Plan 'Related Tools, Software and/or Code' should be citable, and the citation included in the Plan. This encourages others to reuse other's work, and expands the 'provide credit where credit is due' system. Under 'Data Types' the sections describing what can be shared versus what data needs to remain private, will be dictated and based on ethical and legal compliance related to the type of research. Therefore, this section may warrant its own section, instead of being included as such a small Plan element or as an 'Additional Consideration' at the end of the Plan. Under Plan 'Data Preservation and Access' there should be more explicit language about this. While the Plan may state where the data will be deposited into a central repository for access, it must also address the active management needed for the long-term preservation assurance and any costs associated with this activity

Under Plan 'Data Preservation and Access Timeline' is it possible to state a more explicit timeframes for projects to make their data available? For example, the NIH Public Access Policy

requires manuscripts that arise from NIH funds are accessible to the public on PubMed Central no later than 12 months after publication. Is there an acceptable time table for making data publicly available? Under Plan 'Oversight of Data Management' it may be important to state who on the project will become the steward of the data and the Plan should anything happen to the current PI of the project. For instance, 'if tragedy strikes, who will become the next in line?' The various components should involve both 'Active' and 'Post' Project roles. For example, there should be an active Project Manager who is tasked with the maintenance and communication of the Plan document and there should be a Contact Person for after the project has been completed for other researchers to contact should they have questions about the project, research or data. Additional Plan considerations should include how project and research plans change overtime. How has NIH thought to address researcher's needs to adapt and change their Data Management and Sharing Plans as they go through their research? The Plan is living document that should be referenced and changed as plans change. For example, what's the backup plan in case the repository dies. One final point of consideration is whether these plans should be made public with the grant submission information.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

Phased adoption could come around requiring certain parts of the Plan and requiring training in the different Plan areas. This could be an opportunity for NIH, Office of Sponsored Programs and Libraries to partner together to provide the needed resources and training needed to guide researchers through the new Plan provisions. Compliance rates with these new requirements will only happen if NIH program officers are diligent about following up with researchers about their Plans and the follow through.

Submission #175**Date:** 12/10/2018**Name:** Daniel Handwerker**Name of Organization:** NIMH (Writing in my personal capacity, not as a representative of NIMH)**Type of Organization:** Government Agency**Role:** Scientific Researcher**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

I conduct neuroscience research primarily on human volunteers using fMRI.

I. The definition of Scientific Data

In addition to the listed elements of scientific data in the Draft NIH Data Management and Sharing Policy, scientific data should include the code or algorithms used to reach published results.

II. The requirements for Data Management and Sharing Plans

I support the proposal that Data Management and Sharing Plan should be part of the funding/support application process.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

Sharing data in a useful manner is not trivial. Many scientists have not been trained to document data and processes on computers in a way that can easily transition to sharing. Adding data sharing requirements without given educational support to train scientists how to share data and fund infrastructure and staff needed to share data will limit the utility of any data that is shared. Discrete funding opportunities to fund this infrastructure (either by university or through online courses) and well as funding to increase the number of people with IT experience working on data sharing would both reduce the burden of this initiative and make sure that the data that is shared is more likely to be useful to others.

Submission #176**Date:** 12/10/2018**Name:** Jason Bret Harris**Name of Organization:** Collaborative Drug Discovery, Inc. (CDD)**Type of Organization:** Biotech/Pharmaceutical Company**Role:** Scientific Researcher**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Bioinformatics and ChemInformatics

I. The definition of Scientific Data

The below definition has been slightly modified with edits noted inside of the brackets.

Scientific Data: The recorded factual material commonly accepted in the scientific community as necessary to validate and replicate research findings including, but not limited to, data used to support scholarly publications. Scientific data do not include laboratory notebooks, preliminary analyses, completed case report forms, drafts of scientific papers, plans for future research, peer reviews, communications with colleagues, or physical objects, such as laboratory specimens. For the purposes of a possible Policy, scientific data may include [software, computer code,] and certain individual 2 level and summary or aggregate data, as well as metadata.⁷ NIH expects that reasonable efforts should be made to digitize [and make machine-actionable] all scientific data.

[Machine-actionable: For the purposes of policy, digitized data should follow standards as described in section IV to render data more easily readable by machines. This may include using certain formats and semantic standards from ontologies depending on the presence of applicable scientific terms and Data Types.]

II. The requirements for Data Management and Sharing Plans

The below statement contains general suggestions for improving this policy.

Machine-Actionable Data. Wilkinson et al. (2016) explain in their Nature article about FAIR principles that well-managed data is increasingly more 'machine-actionable'. The current policy as written is vague on how to ensure that data is made machine-actionable. It is recommended that the introduction of the policy elude to this important aspect of FAIR data. Also the

paragraph about Standards in the Plan Elements of Section IV would be more effective by stating that researchers ‘must’ follow community data standards and common data elements (CDEs) instead of merely “encouraging’ them to do so. With this in mind, the use of semantic standards from community ontologies should also be mentioned as part of the Standards section. The importance of semantics for machines is discussed by Wilkinson et al. (2016). Data annotated with a semantic standards is made machine-readable, interoperable, easier searchable, and ultimately reusable (the goal).

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

As with FAIR data, standards for managing it need to be findable too. For reference purposes, it would be helpful to compile known resources for building machine-readable data, similar to the already referenced NIH CDE Portal. The first three tools of many that come to mind are listed:

BioPortal - National Center for Biomedical Ontology - hundreds of biomedical ontologies.
<https://bioportal.bioontology.org>

CEDAR - Center for Expanded Data Annotation and Retrieval - specializes in forming templates for managing complex metadata. <https://metadatascenter.org>

BioAssay Express (BAE) - Collaborative Drug Discovery, Inc. - helps researchers quickly integrate their data with existing biomedical ontologies.

These and other resources would be best curated continually by the community and kept in a wiki-style portal endorsed or managed by NIH. Additionally, recommended repositories associated with different Data Types should be included in such a wiki.

Submission #177**Date:** 12/10/2018**Name:** Katie Steen**Name of Organization:** Association of American Universities (AAU)**Type of Organization:** University**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Given we represent research universities, all research areas are important to our members.

I. The definition of Scientific Data

We support NIH's proposed definition of scientific data and that it should not include preliminary analysis, lab notebooks, and other early outputs of the research process, and should primarily consist of "individual level and summary or aggregate data, as well as metadata."

Additionally, we request that NIH provide standards for data and metadata, to the extent possible, to ensure data is reproduceable and user-friendly. If NIH is aware of standards within disciplines that it supports, it would be helpful to share these standards with the broader research community.

II. The requirements for Data Management and Sharing Plans**Plan Review and Evaluation**

We appreciate that NIH is not proposing to factor data management plans into the overall impact score for extramural grants. Standards for data sharing vary significantly by discipline and both researchers and NIH program managers and reviewers may not have the data expertise to write or evaluate data management and sharing plans. Given the complexity and changing landscape of data sharing and that research plans sometimes change as novel discoveries are made, some plans may need to undergo considerable edits that warrant significant dialogue with the program manager and reviewers over the lifetime of the grant. It is critical that NIH creates and sustains opportunities for potential grantees and program managers to discuss the content of the plans and any concerns without it being factored into the impact score.

We are concerned that while researchers are experts in their disciplines, they may not be experts in data management and sustainability. Specifically, researchers may not have agreed upon “community data standards” in their discipline, making it difficult to propose where and for how long data should be stored. If data expertise is required on the part of the researchers, universities would have to spend additional resources on the writing of data management plans before applying for NIH funding.

We support NIH’s proposed annual reviews of data management plans and interpret this to mean reviews will not continue after the grant has closed. In addition, it is unclear in the proposed provisions, if any party (researcher or university) is responsible for compliance with the data management plan after the grant has ended. In our report recommendations for federal agencies, we suggest that federal agencies “consider the community of interest and duration of usefulness for the data in question and make retention and access requirements clear.” Clarity on this issue would be helpful to determine the kind and amount of additional resources universities will need to allocate to fulfill the agency requirements.

Plan Elements

Given the number of requirements outlined in the proposed elements of the plan, we do not believe two pages is enough for researchers to accurately and fully outline appropriate data management and sharing plans. Additionally, we think it will be very difficult to estimate the amount of scientific data that will result from NIH-funded research in advance of conducting the research. Many research projects require unexpected but necessary experiments, making it impossible to accurately estimate the amount of data resulting from the research. We also seek clarity from NIH on the standards researchers should reference in their plans and if these standards will be defined by NIH or others.

Data preservation and management standards are still evolving across the scientific research enterprise. Best practices around long-term data preservation and access are still being defined in some disciplines. If a university is required to ensure long-term access to and preservation of data, it is important that there be a mechanism to update and change how existing data is stored and plans for continued preservation. For example, if a repository shuts down, a university (assuming the onus is on the university) should be able to contact NIH and make a new plan without penalty. It would be helpful for NIH to provide guidance on what types of “newly created repositories” would be accepted. We would also suggest that NIH create its own data repository to host any NIH-funded research data. This would be especially useful for disciplines that do not already have “community repositories.” Additionally, universities need clarity on if the timeline for data preservation and access is expected to continue after the grant is closed.

Many researchers will find it difficult to anticipate the commercialization opportunities before a research project starts. If NIH requires potential grantees to describe limitations on the data use with respect to intellectual property and licensing agreements too early, researchers may

choose to claim that all research has commercialization potential, resulting in an unintended consequence of the NIH data sharing policy as currently proposed.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

Given the requirements for developing plans to manage and make data publicly accessible under NIH's proposed policy, it is critical that universities and researchers understand what additional expenses would be allowed as "reasonable costs" under the new policy. Clarity from NIH on this issue would help researchers in proposing their data management plans. We would strongly urge that "reasonable costs" be a standard part of the grant proposal submission.

Implementation of these proposed provisions and changes to data plans will require significant time and resources on behalf of the university and the researchers. To fully prepare for these changes, we request a two-year implementation period for the new policies. These two years would give universities time to facilitate additional discussions across institutions to determine meaningful and effective data access collaborations while engaging with NIH and other federal agencies. We would also strongly encourage the harmonization of data policies and the required elements of data management plans within the NIH and across federal agencies.

Attachment:



Association
of American
Universities
Inquiry · Innovation · Impact



ASSOCIATION OF
PUBLIC &
LAND-GRANT
UNIVERSITIES

COGR
Council On Governmental Relations

MEMORANDUM

TO: Office of Science Policy, National Institutes of Health

FROM: Association of American Universities
Contact: Katie Steen, katie.steen@aau.edu; (202) 789-5377

Association of Public and Land-grant Universities
Contact: Kacy Redd, kredd@aplu.org; (202) 478-6022

Council on Governmental Relations
Contact: Jackie Bendall, jbendall@cogr.edu; (202) 289-6655

DATE: December 10, 2018

Re: NOT-OD-19-014 “Request for Information (RFI) on Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research”

On behalf of the over 200 universities we represent, the Association of American Universities (AAU), the Association of Public and Land-grant Universities (APLU), and the Council on Governmental Relations (COGR) greatly appreciate the opportunity to comment on the National Institutes of Health (NIH) proposed policy provisions and proposed required data elements for data management and sharing plans.

Data access, preservation, and management are complex and emerging areas of the research pipeline. Universities agree it is beneficial to make data from federally-funded research accessible both to the public and others in the research community to accelerate scientific discovery by making data more open to scrutiny and re-analysis. We are actively working with our campuses to help them fulfill agency requirements and continue to engage with federal agencies as they develop additional data access policies.

To build on a [report](#) released by AAU and APLU in November 2017 with recommendations for universities and federal agencies, AAU and APLU hosted an NSF-funded workshop in October 2018 on [Accelerating Public Access to Research Data](#). The workshop convened federal agency representatives and 30 institutional teams comprised of senior research officers, data librarians, general counsels, information technology specialists, faculty members and other university administrators. The goals of the workshop were to better understand the challenges to data sharing and identify opportunities for collaboration and alignment, all in support of developing campus action plans to advance data sharing. Our universities found it helpful to discuss data access and management with stakeholders across the federal government and campuses. Representatives from NSF, NIH, Department of Energy, National

Institute on Standards and Technology, the Department of Defense, and OSTP also expressed that it was very valuable to hear the barriers and opportunities to data sharing from this cross-campus community.

We were only able to convene 30 institutional teams from primarily research-intensive institutions at this workshop, although 52 universities had applied to participate. Participants agreed there is need for additional workshops for agencies and institutions to discuss better and new ways to collaborate and to build upon the momentum generated by this first workshop. We recommend that NIH support and provide similar venues for universities--especially from a diverse mix of institutions that conduct federally funded research--and agencies to discuss public access to research data before releasing final provisions.

The Definition of Scientific Data

We support NIH's proposed definition of scientific data and that it should not include preliminary analysis, lab notebooks, and other early outputs of the research process, and should primarily consist of "individual level and summary or aggregate data, as well as metadata."

Additionally, we request that NIH provide standards for data and metadata, to the extent possible, to ensure data is reproduceable and user-friendly. If NIH is aware of standards within disciplines that it supports, it would be helpful to share these standards with the broader research community.

Requirements for Data Management and Sharing Plans

Plan Review and Evaluation

We appreciate that NIH is *not proposing* to factor data management plans into the overall impact score for extramural grants. Standards for data sharing vary significantly by discipline and both researchers and NIH program managers and reviewers may not have the data expertise to write or evaluate data management and sharing plans. Given the complexity and changing landscape of data sharing and that research plans sometimes change as novel discoveries are made, some plans may need to undergo considerable edits that warrant significant dialogue with the program manager and reviewers over the lifetime of the grant. It is critical that NIH creates and sustains opportunities for potential grantees and program managers to discuss the content of the plans and any concerns without it being factored into the impact score.

We are concerned that while researchers are experts in their disciplines, they may not be experts in data management and sustainability. Specifically, researchers may not have agreed upon "community data standards" in their discipline, making it difficult to propose where and for how long data should be stored. If data expertise is required on the part of the researchers, universities would have to spend additional resources on the writing of data management plans before applying for NIH funding.

We support NIH's proposed annual reviews of data management plans and interpret this to mean reviews will not continue after the grant has closed. In addition, it is unclear in the proposed provisions, if any party (researcher or university) is responsible for compliance with the data management plan after the grant has ended. In our report recommendations for federal agencies, we suggest that federal agencies "consider the community of interest and duration of usefulness for the data in question and

make retention and access requirements clear.”¹ Clarity on this issue would be helpful to determine the kind and amount of additional resources universities will need to allocate to fulfill the agency requirements.

Plan Elements

Given the number of requirements outlined in the proposed elements of the plan, we do not believe two pages is enough for researchers to accurately and fully outline appropriate data management and sharing plans. Additionally, we think it will be very difficult to estimate the amount of scientific data that will result from NIH-funded research in advance of conducting the research. Many research projects require unexpected but necessary experiments, making it impossible to accurately estimate the amount of data resulting from the research. We also seek clarity from NIH on the standards researchers should reference in their plans and if these standards will be defined by NIH or others.

Data preservation and management standards are still evolving across the scientific research enterprise. Best practices around long-term data preservation and access are still being defined in some disciplines. If a university is required to ensure long-term access to and preservation of data, it is important that there be a mechanism to update and change how existing data is stored and plans for continued preservation. For example, if a repository shuts down, a university (assuming the onus is on the university) should be able to contact NIH and make a new plan without penalty. It would be helpful for NIH to provide guidance on what types of “newly created repositories” would be accepted. We would also suggest that NIH create its own data repository to host any NIH-funded research data. This would be especially useful for disciplines that do not already have “community repositories.” Additionally, universities need clarity on if the timeline for data preservation and access is expected to continue after the grant is closed.

Many researchers will find it difficult to anticipate the commercialization opportunities before a research project starts. If NIH requires potential grantees to describe limitations on the data use with respect to intellectual property and licensing agreements too early, researchers may choose to claim that all research has commercialization potential, resulting in an unintended consequence of the NIH data sharing policy as currently proposed.

The optimal timing, including phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy.

Given the requirements for developing plans to manage and make data publicly accessible under NIH’s proposed policy, it is critical that universities and researchers understand what additional expenses would be allowed as “reasonable costs” under the new policy. Clarity from NIH on this issue would help researchers in proposing their data management plans. We would strongly urge that “reasonable costs” be a standard part of the grant proposal submission.

Implementation of these proposed provisions and changes to data plans will require significant time and resources on behalf of the university and the researchers. To fully prepare for these changes, we request a two-year implementation period for the new policies. These two years would give universities

¹ <https://www.aau.edu/sites/default/files/AAU-Files/Key-Issues/Intellectual-Property/Public-Open-Access/AAU-APLU-Public-Access-Working-Group-Report.pdf>

time to facilitate additional discussions across institutions to determine meaningful and effective data access collaborations while engaging with NIH and other federal agencies. We would also strongly encourage the harmonization of data policies and the required elements of data management plans within the NIH and across federal agencies.

We appreciate the opportunity to comment on the proposed policy provisions and data plan elements. AAU, APLU, and COGR encourage NIH to host convenings where universities and NIH staff can engage in dialogue about these issues before releasing final provisions.

Submission #178**Date:** 12/11/2018**Name:** Ricardo de Miranda Azevedo**Name of Organization:** Maastricht University**Type of Organization:** University**Role:** Scientific Researcher**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Data Science, Epidemiology, Data Management

I. The definition of Scientific Data

For developing a broad and clear set of requirements for Data Management and Sharing Plans, we hereby offer our expertise. Our interdisciplinary community has been involved with a number of initiatives related to data management. Among these initiatives, are the FORCE11 group, responsible by developing the FAIR principles for scientific data management, the GO-FAIR initiative, which aims to generate resources for the effective implementation of the FAIR principles through three pillars: GO-CHANGE (disseminate the FAIR principles across different communities), GO-TRAIN (generate training material for instructing researchers to implement the FAIR principles, and GO-BUILD (develop software for enabling the implementation of the FAIR principles. The ultimate goal of GO-FAIR is to enable the construction an internet of FAIR data. Moreover, we are also engaged on the FAIRmetrics group which aims to

Scientific data is the foundation for science. Similarly, one could say that sound is the foundation for music. Structurally defining what is scientific data; or sound; is a rather unfair task. Such terms can be used to mean several things on several way, depending on the way it gets used. Categorizing “scientific data” as in a dictionary style, will likely lead to incomplete definitions. Therefore, we propose that this definition should be based on the “lifecycle” of the data being produced in a research project. Most importantly, however, scientific data may be useful for other things than research, such as for clinical, healthcare, governmental and business purposes .

Researchers often tend to limit their idea of scientific data as being the dataset used for the analyses reported in their manuscripts. However, a substantial amount of (meta)data is also produced surrounding this “core” analysis dataset (e.g. raw data, contracts, standard operating procedures, policies). Considering that the current scientific paradigm endorses FAIR and open

data, researchers should consider all forms of supplementary (meta)data being produced during a research project.

II. The requirements for Data Management and Sharing Plans

There have been many efforts done for developing requirements for data management and sharing plans. The ideal procedure would consist of selecting elements from a basis template of a data management plan, that should be later assessed by a peer-review expert of the field.

Besides the data management plan, a FAIRness assessment would be ideal. There are currently initiatives to objectively measure compliance with the FAIR principles. Our community is involved with the FAIRmetrics group, that has developed and published a set of metrics based on each of the FAIR principles [https://\(www.fairmetrics.org\)](https://www.fairmetrics.org). We propose the inclusion of a FAIRness assessment embedded on the data sharing plan, for ensuring that the data and metadata are properly available and enhance the chances of data reuse.

Ideally, a data sharing plan must also include information clarifying all privacy-related issues of the data. Researchers collecting human data must clearly explain their will for reuse, and ensure participants that their privacy requirements will be met. In general, privacy concerns come more often from the side of researchers than participants, Prior to letting privacy concerns inhibit data sharing, we should better talk to participants and hear what they think about it. A study on clinical trials participants indicated that 93% of the sample would allow their own data to be shared with university scientists, and 82% would be willing to allow their data to be shared with scientists working in for-profit companies. Therefore, privacy should not be a reason to inhibit data sharing efforts.

Understanding the attitudes and behaviors that influence researchers to share their data is also crucial for achieving proper sharing and increases in data reuse. More than ensuring that researchers funded by the NIH will share their data, it is imperative to ensure the availability of these data (that it is ready to be reused). Another important aspect, is that the NIH must ensure that researchers will not negatively experience their additional efforts for making their data shareable. A currently study showed that perceiving the efforts of sharing as a burden are the largest cause of inhibiting data sharing. Understanding what makes researchers reuse data, should also take place in the agenda of the NIH data management and sharing policy. Ideally, focus groups should be performed with researchers so that their needs are understood for reusing data. Moreover, incentives should also be given when possible, whether by means of badging or financial.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

From an institutional point of view, the NIH should consider to involve more the researchers being funded and understand their needs and difficulties by collecting survey data, and find out what kind of infrastructure should be provided prior to implementing a new policy.

For defining the optimal timing we should distinguish between new studies that will get funded and studies already receiving funding.

For new studies, researchers should ideally plan their sharing based on data collection, based on a framework of the data life cycle. Therefore, the implementation should follow the order:

1- Inventorisation of the expected data types to be produced (Embedded on the grant proposal)

2- Data management and sharing plan could be ideally assessed in three phases:

Initial (up to 6 months after project has been approved, prior to data collection starts)

Mid-term (After data collection starts, before data collection ends; approximately 2 years for 4 year projects)

Final (Before the end of the project, but close to the end of the project, with enough information on data sharing (incl. FAIRness assessment)).

3- FAIRness assessment: performed at the third year of the project, Afterwards, a follow up evaluation should be performed, where the feedback from the first assessment should be provided.

Submission #179

Date: 12/10/2018

Name: Andrew Tein

Name of Organization: Wiley

Type of Organization: Other

Other Type of Organization: Research publishing and platform company

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

all biomedical areas; multidisciplinary

I. The definition of Scientific Data

Please see attached comments

II. The requirements for Data Management and Sharing Plans

Please see attached comments

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and

Please see attached comments

Attachment:

December 10, 2018

Carrie D. Wolinetz, Ph.D.
Associate Director for Science Policy
Office of Science Policy (OSP)
National Institutes of Health
6705 Rockledge Drive, Ste 750
Bethesda, MD 20892

Dear Associate Director Wolinetz:

I am writing on behalf of Wiley, a leading American research and education company. Wiley publishes 1,600 journals across all major disciplines and is honored to partner with over 600 non-profit societies in the United States and worldwide. Our subsidiary Atypon develops technology solutions to deliver mission-critical content to practitioners and researchers in every field, and its publishing platform hosts nearly 45% of the world's English-language scholarly journals. Wiley is actively working with its authors and society partners in enabling open science, including through our vibrant open access publishing programs and our data sharing and citation policies.

We appreciate the opportunity to provide comments in response to the National Institutes of Health (NIH) *Request for Information on Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research*. Wiley is committed to improving openness, transparency, and reproducibility of research. Fundamental to enabling reproducible research is the easy access to and ready discovery of its supporting data, made possible through a robust and universal framework that allows research data to be cited through standard reference lists. This will ensure that data is treated as a first-class research object, easily accessible as part of the scholarly literature, and that researchers are credited for their work. We look forward to collaborating with the NIH in this effort.

Definition of Scientific Data

We recognize that the definition of scientific data can vary by discipline, and that different communities may have different standards. To this end, it will be important for NIH to provide an overarching framework while also allowing some flexibility for communities that may have additional, specific standards. We appreciate that the definition in the proposed provisions, in line with the Office of Science and Technology Policy memorandum on Increasing Access to the Results of Federally Funded Scientific Research and the America COMPETES Reauthorization Act of 2010, makes a distinction between data and scholarly publications, as well as other content. This framework will help create clarity for the research community and ensure that open science policies can be appropriately tailored to address the specific characteristics of scientific data and the unique challenges and opportunities for making it more widely available.

Purpose

We commend NIH for holding the current RFI and its commitment to regularly evaluate any new data sharing policy. In light of the complexities and costs associated with data sharing, we hope NIH will continue to consult openly with the entire research ecosystem, fully evaluate the impacts of any proposed policies, and ensure that burdens on researchers and other stakeholders are minimized.

Scope and Requirements

Data sharing will require a significant, long-term investment to achieve the goals outlined in the proposed scope and requirements. We appreciate that the NIH will allow for researchers to request reasonable costs

associated with data management and sharing. This funding will be essential to ensure that researchers have the resources to fulfill any new requirements, and to enable the development of a sustainable, high-quality ecosystem of data repositories and other providers to support the preparation, archiving, preservation and curation of scientific data. As requirements and expectations can change over the course of a project, and associated data management and sharing activities may take place long after its conclusion, it will be important for NIH to ensure that sufficient resources are made available to achieve the goals of the policy.

Requirements for Data Management and Sharing Plans

As a step towards promoting data sharing, Wiley is working across its 1,600 journals to encourage adoption of one of four various levels of data sharing policies, namely 1) encouraging data sharing, 2) expecting data sharing, 3) mandating data sharing, or 4) mandating data sharing and peer reviews. While these efforts remain a work in progress, they provide an opportunity for individual journals and communities to move towards greater data sharing, tailored to their specific needs and at the appropriate pace for their authors.

We are acutely aware that no one-size-fits-all solution is appropriate or desirable in promoting open science, including data sharing. Each community, journal and individual researcher will have a unique set of requirements and circumstances that must be taken into account as part of this process to both encourage greater access while also promoting the best science. To this end, we offer several initial comments on the proposed requirements for data management and sharing plans:

- Related Tools, Software and/or Code: It is unclear why the proposed provisions suggest requiring that data plans describe whether software/computer code used to process or analyze data is free and open source, and if not, why it is needed and whether alternative free and open source software/code may be available. Researchers, just like federal agencies, businesses, and other individuals and organizations, rely on a variety of software tools which are subject to different licensing arrangements and business models. Particularly as data sets grow, researchers will likely take advantage of more sophisticated software and analytics, including based on artificial intelligence, to develop new insights which may not have been possible before. While it will be useful for researchers to indicate what software/code will be used and needed to process and analyze data and whether alternatives are available, this guidance should be applied equally without regard to underlying licensing arrangements or business models to ensure that researchers are encouraged to leverage the most appropriate software tools for a given project.
- Standards: Wiley is actively involved in contributing to the research data community as an organizational member or signatory of the following initiatives, which we encourage NIH to engage, align and partner with as it moves forward:
 - Research Data Alliance (RDA)
 - International Council for Science World Data System (ISCU-WDS)
 - ORCID
 - Initiative for Open Citations (I4OC)
 - Transparency and Openness Promotion (TOP) Guidelines
 - STM Brussels Declaration
 - FORCE 11 FAIR Data Principles
 - Joint Declaration of Data Citation Principles (JDDCP)

In August 2018, Wiley was the first of the major publishers to sign on to the ‘Enabling FAIR data commitment statement in the Earth, Space and Environmental Sciences.’ The Enabling FAIR Data initiative is an extension of COPDESS, the Coalition for Publishing Data in the Earth and Space

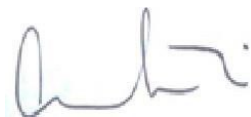
Sciences which Wiley also signed back in 2013, as a demonstration of our support of open data. This new initiative focuses on practical implementation of these principles and provides guidance and solutions for publishers and repositories to take meaningful action towards making data FAIR.

To support this and other data initiatives, Wiley will be participating in the [Scholix](#) initiative, a high-level interoperability framework for exchanging information about the links between scholarly literature and data.

- Data Preservation and Access Timeline: We encourage authors to make research data available as early as possible. However, we recognize that practice varies by field, and embargoes on data sharing are common practice in some communities. We encourage NIH to provide flexibility for each community to determine what data sharing timelines are most appropriate and acceptable for its researchers.
- Data Sharing Agreements, Licensing, and Intellectual Property: Researchers should ideally decide how their scientific data is made available. While some scientific data may not be subject to intellectual property rights, other data may be. Consistent with the 21st Century Cures Act and other federal policies, it will be important for the NIH to respect rights holders' decisions in terms of how they would like to license their intellectual property, and to ensure that the intellectual property rights associated with any scientific data used in a project that was not generated under the NIH award is not subject to the proposed data policy.
- Scientific Data Archiving: In general, we encourage authors to submit scientific data to discipline-specific, community-recognized repositories where possible, or to general-purpose repositories if no suitable community resource is available. There already exists a diverse and dynamic landscape of data repositories, serving various research communities and operating under a variety of models. We encourage the NIH to ensure its efforts are in line with and supportive of these ongoing community efforts, do not duplicate existing infrastructure, and provide flexibility for researchers to make data archiving decisions that are most appropriate for their projects.

Thank you for the opportunity to provide comments as the NIH develops a potential Data Management and Sharing Policy. Wiley looks forward to continuing to work with researchers, societies, NIH and other stakeholders to promote data sharing and enhance research communication.

Sincerely,



Andrew A. Tein
Vice President, Global Government Affairs
John Wiley & Sons (Wiley)
T: +1 201 748 7751
e-mail: antein@wiley.com

Submission #180

Date: 12/10/2018

Name: Alexander (Sasha) Wait Zaranek

Name of Organization: Curoverse Research

Type of Organization: Other

Other Type of Organization: For Profit, Open Science Business

Role: Institutional Official

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):

Genomic / Precision Medicine

Attachment:

Response to Data Management and Sharing Policy for NIH Funded or Supported Research

Our organization recommends a standardized data management and resource sharing policy to all our investigators for their research. (See **Appendix A** next page.) In developing this policy we carefully incorporated input from the Free Software Foundation, the Software Freedom Law Center, the Creative Commons Foundation, the Wikimedia Foundation (responsible for Wikipedia), the Open Source Initiative, the Software Freedom Conservancy, the Open Knowledge Foundation and others. We are sharing our own choices with the hope our policies might benefit other NIH investigators.

Our policy is a synthesis of more than twenty five years of experience with commercial Free and Open Source Software (FOSS) licensing across multiple domains. In the last decade we have extended these ideas to biomedical data and materials (DNA, cells, etc.) under an “open consent” approach pioneered at the Harvard Personal Genome Project (PGP) by Dr. Jeantine Lunshof and others.

Each new PGP organization demonstrates the replicability of the model in a new regulatory environment. Across PGPs, we have enrolled 7,000+ individuals, of which over 3,000 have publicly shared phenotype information. The PGP has sequenced genomes of hundreds of people and hundreds more are in the sequencing pipeline. To date, this is still the only publicly available resource with genomes, phenotypes and cells available for both commercial and non-commercial purposes. This has allowed both for-profit and non-profit organizations to create products from PGP data and cells.

The unique combination of public domain data, open standards and FOSS implementations could make NIH resources significantly more accessible to new audiences. At our organization high-school students, retirees and other members of the general public are able to explore non-anonymous biomedical data and do machine learning (building eye color, blood type and other classifiers) without being “qualified researchers” and with no special access permissions required. While open standards, FOSS implementations and open data are powerful individually, they are especially powerful together.

Without question not all projects will be able to implement these policies. In our experience, however, virtually every cohort includes individuals that would be willing to share their data publicly and these individuals should be allowed to do so! It seems only fair that when the public shares their data (or cells) with us that researchers share their own data, software and methods with the public. The PGP has shown we can do this responsibly and we offer our policies to other NIH investigators (and to investigators worldwide) as an example of responsible sharing.

Sincerely,



Alexander (Sasha) Wait Zaranek, PhD.
Co-Founder, Harvard Personal Genome Project (PGP)
Chief Scientist, Curoverse Research

Appendix A. Sample data management and sharing policy for our organization

Our organization is committed to sharing resources with the scientific community and the general public, so that people may collaborate in ways that have not been possible before. We think this will promote discovery and accelerate our ability to realize precision medicine. To this end, we will use the following legal tools and standards in the proposed research.

Open Consent

We build on the [open consent](#) protocol pioneered by the Harvard Personal Genome Project (PGP) and adopted worldwide by PGP affiliates and other similar "public genomics" research studies to make non-anonymous human data publicly available in an ethical manner.

- Read the Nature Reviews Genetics publication from April 2008. ([PDF](#))
- Read the original white paper on open consent from April 2007, published by George Church, Jeantine Lunshof, and Daniel Vorhaus. ([PDF](#))

Open Standards

- GA4GH - <http://genomicsandhealth.org>
- NIST GIAB - <http://jimb.stanford.edu/giab>
- CWL - <http://www.commonwl.org>

Open Data

We are creating a repository of integrated genomic, environmental, and trait datasets as well as accession numbers for accompanying cell-lines and other biological materials. These data are available in the public domain:

- CC0 - <http://creativecommons.org/publicdomain/zero/1.0/legalcode>

Open Source Software

Our software development efforts are committed to using a combination of GNU GPL and more permissive (GPL compatible) open source licenses:

- GNU AGPLv3 - <http://www.gnu.org/licenses/agpl-3.0.html>
- GNU GPLv3 - <http://www.gnu.org/licenses/gpl-3.0.html>
- Apache v2 - <http://www.apache.org/licenses/LICENSE-2.0.html>
- CC0 - <http://creativecommons.org/publicdomain/zero/1.0/legalcode>

Open Access Text and Other Media Licenses

We publish our research in open-access journals and pre-print servers as well as use other Internet based dissemination under these licenses:

- CC-BY-SA - <http://creativecommons.org/licenses/by-sa/4.0/legalcode>
- CC-BY - <http://creativecommons.org/licenses/by/4.0/legalcode>
- CC0 - <http://creativecommons.org/publicdomain/zero/1.0/legalcode>

All of the options are highly regarded legal constructs consistent with recommendations from the Open Knowledge Foundation, a reputable organization encouraging the use of standard legal constructs in open science projects. See: Molloy JC (2011) The Open Knowledge Foundation: Open Data Means Better Science. PLoS Biol 9(12): e1001195. [doi:10.1371/journal.pbio.1001195](https://doi.org/10.1371/journal.pbio.1001195)

Submission #181**Date:** 12/10/2018**Name:** Diane Lehman Wilson**Name of Organization:** University of Michigan Medical School**Type of Organization:** University**Role:** Member of the Public**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

The University of Michigan is involved in all of the areas listed and so many more. I have assisted biomedical and behavioral researchers in each of these areas and many more.

I commend and appreciate that this RFI, while divided as others have been, into specific questions, has generous data limits and upload capability so that contextual information can be shared.

I. The definition of Scientific Data

It is important for NIH to provide clear standards for data and metadata. Preliminary analyses, lab notebooks, case report forms, and other early outputs of the research process should not be included in the definition of scientific data.

II. The requirements for Data Management and Sharing Plans

With data management and sharing plans, it would be most helpful if NIH would provide templates for the sorts of plans, as well as the licensing agreements that might support such sharing. With data science changing so swiftly, NIH should allow for changes to this plan as projects progress, especially for projects that are expected to last more than a few years. The degree of detail requested in the proposed plan does not reasonably cohere with a two page requirement, unless the government itself provides the data repositories and the detailed expectations that underlie the statements that such a plan would contain.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

The timing for implementation of any new policy should be contingent on having built the infrastructure to support the data sharing it seeks, complete with templates, sample agreements etc. and full consideration of the issues of informed consent discussed below. Beyond that, the timing of data sharing expected per project should have some flexibility in response to the years involved in data collection or reasonable analysis, the sensitivity of the data, and researcher circumstances which would justify additional time before data sharing.

Informed Consent: The history of informed consent as it relates to populations without equal access to the scientific and medical enterprise is riddled with areas of governmental and corporate insouciance for the feelings, risks, and wellbeing of participants. I appreciate NIH's desire to get wide input before creating a broad data sharing policy, but to get this right this time, requires, I believe, NIH's own substantial investment in developing and testing methods of de-identification while preserving data integrity. Communities with special concerns (whether they be populations with sensitive diseases or conditions, small genetically identifiable groups, or others) should be involved in the development and vetting of such systems. Ultimately it may be determined that there are swaths of research where the balance between honoring autonomy and privacy is not really possible to tip in favor of data sharing. These issues should be considered and exposed prior to the development of a policy rather than creating an overbroad policy which ends up reinforcing or reenacting systemic inequities and injustices.

Potential participants in research should not have to feel that they are being compelled to give up their privacy. Certainly studies that began prior to whatever future date a policy may come out should not be subject to data sharing requirements that could violate earlier informed consents.

Attachment:

12/10/18

Francis S. Collins, MD, PhD
National Institutes of Health
Bethesda, MD

Submitted electronically: <https://osp.od.nih.gov/provisions-data-managment-sharing/>

Re: RFI on Proposed Provisions for Draft Data Management and Sharing Policy

Dear Dr. Collins:

I am writing as a private citizen with more than seven years' experience as a staff member in the University of Michigan Medical School Office of Regulatory Affairs. I appreciate the webinar that was offered in association with this Request for Information and the opportunity to comment on potential future data sharing policies. I know that you seek comments from researchers themselves as well as institutions, but as the researchers with whom I work are extremely busy with their actual scientific work, I hope that submitting these comments will give voice to ideas faculty and other staff have shared with me.

While I myself have never received research funding from the federal government, I have helped hundreds of professors with their ClinicalTrials.gov submissions, trained similarly large numbers of staff about International Committee of Medical Journal Editors expectations, have participated for years on the Clinical Trials Registration and Results reporting task force (formerly part of the CTSA regulatory knowledge subgroup) and am involved in ongoing efforts with other parts of campus to facilitate data sharing plans for clinical trials.

As a private citizen with training in law and public policy, I commend the NIH's efforts to seek to maximize the utility of the science that it funds by encouraging data sharing. That said, as a staff member, I have seen up close the frustration, confusion, and sometimes massive investments that my own and peer universities expend trying to keep up with frequently changing standards, especially those which require huge investments of infrastructure as this one will.

I have also witnessed for-profit companies try to enter data management fields, claiming they will save institutions time and resources, insufficiently living up to their promises, and having tremendous difficulties adapting their products to keep up with changing regulatory requirements. I am very concerned that this proposed policy may create another such market, and even if it does not, that this may prove to be one more area where increasing infrastructure requirements leave smaller institutions even farther behind large ones, or that multiple institutions may compete to create acceptable repositories, dividing the information data banks and therefore causing still more inefficiencies for those seeking access to the information. Similarly, I have heard many researchers say that pilot studies should not be subjected to the same expectations as larger scaled studies, and that doing so imposes a surtax that sometimes outweighs the very benefit of accepting a small grant and doing the work involved. A well-designed policy may avoid this concern, but a poorly designed policy can make this concern even greater.

Thus the request for information in this context is especially welcome and appreciated. In this area, after extensive work with different clusters of NIH-funded investigators, the government should develop templates, tools and topic-specific repositories (or a larger single repository with different components tailored to different data types) before mandating broad sharing. If the tools for compliance are readily available to researchers to understand what is expected before beginning their proposals and whereby the costs involved are also identifiable in advance, both public and scientific respect for the data sharing enterprise will be stronger. Simply put, if you build the cart, put the horse in front of it, and feed the horse, researchers will be happy to load the cart up with the data and discoveries you seek. If those elements are lacking, the cart will not be loaded and able to move.

I also wish to commend and specifically thank NIH that this RFI, while divided as others have been, into specific questions, has generous data limits and upload capability so that contextual information can be shared.

Regarding specific areas of information request.

The Definition of Scientific Data and Timing of Data Sharing Requirements

It is important for NIH to provide standards for data and metadata. Preliminary analyses, lab notebooks, case report forms, and other early outputs of the research process should not be included in the definition of scientific data. The timing for when data sharing is expected should have flexibility, in appreciation of different amounts of time required for different types and scales of analyses, differing publishing timelines and the realities of faculty fulfilling many roles contemporaneously. Therefore the timing that might feel appropriate in a corporate world may not be realistic in an academic environment where a) grant and educational cycles may overlap in complex ways and b) individual researchers may not be as fungible as corporate personnel in large institutions and therefore requirements should not be so rigid that faculty cannot live up to their responsibilities as NIH funded researchers and still face the human factors of child-bearing, elderly parent care, and their own sickness. I have interviewed former faculty with good research potential or history leave academia altogether (with early retirement, or for other types of positions altogether) in part specifically because the requirements from NIH have become increasingly top heavy.

In sum, if the government wants a particular outcome (perpetually or long-term data preserved indefinitely for future mining and use), it should build, or at least develop and fund the system needed (if, for example, it were done through a multi-institutional consortium) with enough care and flexibility that it does not impinge the research projects being funded. Such government development would have to contain the long term funding to maintain and perpetuate the data maintenance as long as the government determines to be desirable.

The Requirements for Data Management and Sharing Plans

With data management and sharing plans, it would be most helpful if NIH would provide templates for the sorts of plans, as well as the licensing agreements that might support such sharing. With data science changing so swiftly, NIH should allow for changes to individual plans as projects progress, especially for projects that are expected to last more than a very few years. The degree of detail requested in the proposed plan does not reasonably cohere with a two

page requirement, unless the government itself provides the data repositories and the detailed expectations that underlie the statements that such a plan would contain.

Timing:

The timing for implementation of any new policy should be contingent on having built the infrastructure to support the data sharing it seeks, complete with templates, sample agreements etc. and full consideration of the issues of informed consent discussed below. Beyond that, the timing of data sharing expected per project should have some flexibility in response to the years involved in data collection or reasonable analysis, the sensitivity of the data, and researcher circumstances which would justify additional time before data sharing.

Informed Consent: The history of informed consent as it relates to populations without equal access to the scientific and medical enterprise is riddled with areas of governmental and corporate insouciance for the feelings, risks, and wellbeing of participants. I appreciate NIH's desire to get wide input before creating a broad data sharing policy, but to get this right this time, requires, I believe, NIH's own substantial investment in developing and testing methods of de-identification while preserving data integrity. Communities with special concerns (whether they be populations with sensitive diseases or conditions, small genetically identifiable groups, or others) should be involved in the development and vetting of such systems. Ultimately it may be determined that there are swaths of research where the balance between honoring autonomy and privacy is not really possible to tip in favor of data sharing. These issues should be considered and exposed prior to the development of a policy rather than creating an overbroad policy which ends up reinforcing or reenacting systemic inequities and injustices.

Potential participants in research should not have to feel that they are being compelled to give up their privacy. Certainly studies that began prior to whatever future date a policy may come out should not be subject to data sharing requirements that could violate earlier informed consents.

Thank you again for this opportunity to share concerns and thoughts as you continue forward in the laudable quest to make the greatest and best use of public funds to support meaningful science in contributing to human health.

Sincerely,

Diane Lehman Wilson

Regulatory Manager

University of Michigan Medical School Office of Regulatory Affairs

Submission #182**Date:** 12/10/2018**Name:** Sarah Greene**Name of Organization:** Health Care Systems Research Network**Type of Organization:** Healthcare Delivery Organization**Role:** Scientific Researcher**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Health Services Research, Epidemiology, Behavioral Science, Data Science, Implementation Science

I. The definition of Scientific Data

We appreciate the opportunity to comment. As a network of research centers embedded in 18 health care delivery systems, the data we use most often for research is derived from electronic health records, health care claims, other administrative data generated through health care transactions, and patient-generated data. We also collect and utilize biospecimen data, which may be generated during health care encounters, or in the context of specific studies. As such, we believe the currently proposed definition could be broadened and made more specific to include the types of data that are generated during the course of patient care, and are then transformed for the purposes of studying specific research questions. Moreover, the eventual policy should be more explicit about whether "scientific data" refers only to raw data, or also to highly circumscribed analytic files that are created for a specific study. These latter data sets require deep familiarity with data quality, provenance and context in order to be interpreted accurately.

II. The requirements for Data Management and Sharing Plans

Given the significant investments by NIH to fund thousands of important research studies, the attention to sharing is welcome and essential. Nevertheless, the data management and sharing plans as conceived in the proposed policy provisions do not fully take into account the range and variety of data that could be managed and shared. We urge NIH to further delineate the differences between types of data/data sets, and provide distinct guidance for categories of data. The processes and procedures for sharing biological data should be considered and constructed separately from the processes and procedures for sharing health system data, for example. Considerations such as data use agreements, business associate agreements, and

variability in IRB requirements related to data storage, archival, and secondary use, are enduring aspects of working with health system-derived data, and thus warrant different requirements.

Also, given the breadth of possible elements of a data management and sharing plan, as listed on pages 4-6 of the proposed provisions, a 2-page limit for a data sharing plan is not sufficient, particularly for complex multi-site interventional trials (whether pragmatic or exploratory).

Managing permissions related to sharing is a tremendous and arduous undertaking when one considered the variety and volume of data sets that could eventually be made available, and the variable permissions rendered by the Institutional Review Board, other compliance body, or health system decision-makers/stakeholders. Additionally, it is not clear what would be considered compliant with the policy. A fully deidentified data set with minimum necessary data might not enable others to reproduce analyses; whereas a limited data set might have more data elements but stricter permissions set by the originating institutions.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards

In our experience, over 25 years of conducting collaborative epidemiological and health services research, a critical aspect of the infrastructure is the lead data analyst. That individual has an unmatched depth of familiarity with the project, however, may only be funded to support the analyses on the original grant. The notion of analysts being "on call" for an unspecified duration, in order to address ad hoc questions about a given data set, is difficult to fathom and even more difficult to resource adequately. Collecting metadata can mitigate this to some degree,

Also, some data, if shared could reveal potentially concerning information about a health system, especially if unblinded. An example for consideration is rates of adverse drug events by site, or simply rates of readmission after hospitalization. Other types of data (e.g., formulary composition) may be used in a research context with permission from the health system, but these are proprietary business data that could affect a health system's market competitiveness.

We encourage NIH to consider the notion of the relative freshness or staleness of a given data set. Secondary use is vital, but if a data set is static and made available for sharing, and newer data are available that are more current or an improvement on the data, is it still worth sharing?

An important consideration that may lend itself to phased adoption relates to helping secondary users find the right data set to fit their needs. We wonder how NIH plans to support visibility and discoverability? In our experience, if significant work is put into a data set for sharing and it's never re-used, it is a stranded asset, and resources used to make it shareable

could have been put to better use elsewhere. The valuation of the myriad data sets that are generated in research is a pivotal question to address, and phasing / piloting tactics to help users find the data could be beneficial

Submission #183**Date:** 10/12/2018**Name:** William Hersh**Name of Organization:** Oregon Health & Science University**Type of Organization:** University**Role:** Scientific Researcher**Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology):**

Biomedical Informatics and Data Science

I. The definition of Scientific Data

My main concern is what appears to be a relatively narrow scope of what constitutes scientific data. The definition above implies that such data is only that which is collected in active experimentation or observation. This ignores the increasing amount of scientific research that does not come from experiments, but rather is derived from real-world measurements of health and disease. This includes everything from data routinely collected by mobile or wearable devices to social media to the electronic health record (EHR). A growing amount of research analyzes and makes inferences using such data.

It could be argued that this sort of data derived “from the wild” should adhere to the provisions above. However, this data is also highly personal and usually highly private. Would you or I want our raw EHR in a data repository? Perhaps connected to our genome data? But if such data are not accessible at all, then the chances for reproducibility are slim.

There is also another twist on this, which concerns data used for informatics research. In a good deal of informatics research, such as the patient cohort retrieval work I do in my own research [1], we use raw, identifiable EHR data. We then proceed to evaluate the performance of our systems and algorithms with this data. Obviously we want this research to be reproducible as well.

There are solutions to these problems, such as Evaluation as a Service [2] approaches that protect such data and allow researchers to send their systems to the data in walled-off containers and receive aggregate results. Maybe the approach in this instance would be to maintain encrypted snapshots of the data that could be unencrypted in highly controlled circumstances.

In any case, the NIH Data Management and Sharing Policy for NIH Funded or Supported Research is a great starting point but should take a broader view of scientific data and develop policies to insure research is reproducible. Research done with data that does not originate as scientific data should be accounted for, including when that data is used for informatics research.

References

1. Wu, S, Liu, S, et al. (2017). Intra-institutional EHR collections for patient-level information retrieval. *Journal of the American Society for Information Science & Technology*. 68: 2636-2648.
2. Hanbury, A, Müller, H, et al. (2015). Evaluation-as-a-service: overview and outlook. arXiv.org: arXiv:1512.07454. <https://arxiv.org/abs/1512.07454>.