Compiled Public Comments on NIH Request for Information: Strategies for NIH Data Management, Sharing, and Citation

November 14, 2016 – January 19, 2017

Public Comments

1. Matt Weatherford, University of Washington
2. Metacogna, Countervalliance, LLC
3. Andrew McCulloch, University of California, San Diego
4. Charles Horn, University of Pittsburgh
5. Mary Carskadon, Brown University
6. Chris Myers, University of Utah
7. Concerned Citizen
8. Kaare Mikkelsen, Aarhus University
9. Brett Duce, Princess Alexandra Hospital
10. Alexander Tsai, Massachusetts General Hospital
11. Leila Tarokh, University Hospital of Child and Adolescent Psychiatry and Psychotherapy
12. Jeffrey Petruska, University of Louisville
13. Logan Schneider, Stanford University
14. Anna May, University Hospital of Child and Adolescent Psychiatry and Psychotherapy
15. Ellen Wijsman, University of Washington
16. Gianluigi DeLucca, Medicoimpianti
17. Eyleen O' Rourke, University of Virginia
18. FASEB
19. David H. Barker, Brown University
20. Ramanathan Natesh, Indian Institute of Science Education and Research
21. MacKenzie Smith, University of California Davis Library
22. Mark Burkard, University of Wisconsin – Madison
23. Daniel S. Katz, University of Illinois Urbana – Champaign
24. Anton Popov, National Technical University of Ukraine
25. Jonathan Petters, Virginia Polytechnic Institute and State University
26. David Hansen, Duke University Libraries
27. Robert Thomas, Beth Israel Deaconness Medical Center
28. Kevin Read, NYU Health Science Library
29. John Conway, R&D Strategy and Solutions
30. Michael Rueschman, Brigham and Women's Hospital
31. Rich Platt & Adrian Hernandez, NIH Health Care Systems Research Collaboratory
32. Sara Mariani, Brigham and Women's Hospital
33. Chris Bourg, MIT Libraries
34. Jose Luis Carrilo Alduenda, Academia Mexicana de Medicina del Dormir
35. Mara Mather, University of Southern California
36. Shawn Murphy, Partners Healthcare
37. Daureen Nesdill, University of Utah Libraries
38. Madhvi Upender, Awarables
39. Livia Dinu, Engineering Custom Solutions, Inc.
40. Steven Ruggles, University of Minnesota
41. James Poterba/Jonathan Skinner, National Bureau of Economic Research
42. Matthew Dougherty, Baylor College of Medicine
43. Holly Falk-Krzesinski, Elsevier
44. Vincent Mor, Brown University
45. David Carr, Wellcome Trust
46. Stephanie Marvin, Brigham and Women's Hospital
47. Jeffrey R. Smith, American Medical Informatics Association
48. Rebecca Boyles, RTI International
49. Shirley Y. Hill, University of Pittsburgh School of Medicine

50. Iain Hrynaszkiewicz, Springer Nature
51. Ara Tahmassian, Harvard University
52. Carmen Nitsche, Pistoia Alliance
53. Mary Ellen Davis, Association of College and Research Libraries
54. Wendy Pradt Lougee, University of Minnestoa
55. Chuck Cook, EMBL-European Bioinformatics Institute
56. Valerie Jackson, Radiological Society of North America
57. Sarah Wright, Cornell University Research Data Management Service Group
58. Matthew Spitzer, Center for Open Science
59. Mary M. Langman, Medical Library Association and Association of Academic Health Sciences Libraries
60. Neil Chue Hong, Software Sustainability Institute
61. Ary L. Goldberger, Beth Israel Deaconess Medical Center/Harvard Medical School
62. Andrew Smith, Elixir
63. Di Cross and Nigel Robinson, Clarivate Analytics
64. Mary Jo Hoeksema, Population Association of America
65. David Lam, University of Michigan
66. Steven M. Girvin, Yale University
67. Susan Redline, Brigham and Women's Hospital
68. Ashok Krishnamurthy, University of North Carolina at Chapel Hill
69. Virginia Steel, UCLA Library
70. Kacy Redd, AAU, APLU, and COGR
71. Shiqiang Tao, University of Kentucky
72. Tim Clark, FORCE11
73. Daniel Valen, Figshare
74. The YODA Project, Yale University
75. Dan Mobley, Brigham and Women's Hospital
76. Sayeed Choudhury, Johns Hopkins University Sheridan Libraries
77. Sarah Brookhart, Association for Psychological Science
78. Ross McKinney, Association of American Medical Colleges
79. James Bryant, Ishpi Information Technologies, Inc.
80. James D. Luther, Duke University
81. John Michael DeCarlo, IBM
82. Letisha R. Wyatt, Oregon Health & Science University
83. Shaun Purcell, Brigham & Women's Hospital
84. John Tagler &Michael Mabe, AAP Professional and Scholarly Publishing Division & International Association of STM Publishers
85. Alexander Sherman, Massachusetts General Hospital
86. Abigail Goben, University of Illinois at Chicago
87. Neils Volkmann, Sanford Burnham Prebys
88. Rebecca Reznik-Zellen, University of Massachusetts Medical School
89. Melissa Haendel, The Monarch Initiative
90. Mark Gerstein, Yale University
91. ASCO, American Society of Clinical Oncology
92. Veronique Kiermer, PLOS
93. Margaret Levenstein, Inter-university Consortium for Political and Social Research
94. GQ Zhang, University of Kentucky
95. Heidi Imker, University of Illinois at Urbana-Champaign

Submission Date

11/15/2016

Submitter Name

Matt Weatherford

Name of Organization

University of Washington

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

Social Science, Computer Science

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

Data needs to be kept around at least 25 years and should be retained for a longer period depending on citations of that data. Both the Data and its codebooks should be maintained in a DDI-approved format and uploaded to a repository Storage is just a matter of disk space. Mulitple copies should be stored in various places around the internet. If things are to be preserved, there needs to be more than one copy floating around

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

All grants should include a long-term data plan. All grants should include funds specifically for cleaning the data and making good code books. We need good, easily-followed standard for how data will live for 25 years at rest and so that future researchers can understand how to make sense of the data. Data is garbage without the appropriate context and codebooks. There needs to be a penalty for grants awarded that do not produce a usable data product in the end. How about a "Score" for the researcher or organization that is done a year after the grant is wrapped up?

4. Any other relevant issues respondents recognize as important for NIH to consider

There are important new developments in open source computing platforms including the concept of a "Container" that could contain all the data and tools needed to re-run the research. These should also be considered in the long term as a possible archival method. But note that there is no substitute for great codebooks and clean data produced in a standards compliant format such as DDI or what IPCSR recommend.

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

Having a Persistent Unique Identifier that resolves to a software package must take version #'s in to account - i.e. a new version of the software or data should be possilbe to produce with a new DOI

b. Inclusion of a link to the data/software resource with the citation in the report

yes

c. Identification of the authors of the Data/Software products

yes

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

The emphasis here should be on what is needed to reproduce the results of the study

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

yes, and remember the "version #" of the dataset of software also - if the digital repository updates the data or software, they must keep the old version around and get a new DOI for the new version

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

NIH should regularly (as a % of funding?) fund efforts to reproduce important research - never at the same facility or university, however! ;) This could be a great way to fund early career scientists and also to make sure that the quality of the research out there is good.

4. Any other relevant issues respondents recognize as important for NIH to consider

There are important new developments in open source computing platforms including the concept of a "Container" that could contain all the data and tools needed to re-run the research. These should also be considered in the long term as a possible archival method. But note that there is no substitute for great codebooks and clean data produced in a standards compliant format such as DDI or what IPCSR recommend.

Additional Comments

Submission Date

11/15/2016

Submitter Name

METACOGNA

Name of Organization

COUNTERVAILLANCE LLC

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

Umbrella/All

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

IGH Reductions / BPL

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

Forever

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

We need to close our FAR 15.6

4. Any other relevant issues respondents recognize as important for NIH to consider

IGH Reduction

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

Reduction of the IGH curve.

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

Included

b. Inclusion of a link to the data/software resource with the citation in the report

http://metacogna.com http://intraneura.com http://countervaillance.com

c. Identification of the authors of the Data/Software products

Kyle Lussier

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

We use new models.

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is

stored and can be found and accessed

Links

Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

Support our platform over next 30 years through the phases.

3. Any other relevant issues respondents recognize as important for NIH to consider

IGH Reduction

Additional Comments

Submission Date

11/16/2016

Submitter Name

Andrew D. McCulloch

Name of

Organization UC San

Diego

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

Cardiovascular Science

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

The highest priority data to be shared are those that are already being presented in publications but not archived. The best way to achieve this would be for the NLM to develop an online manuscript preparation portal and require researchers to use it for all their manuscripts. This would not be costly and could cover the majority of data types quite easily. Instead of turning our raw data into tables and graphs or illustrations, we would upload raw images and tables to an NLM server, annotate the data with existing ontologies and use online tools to generate tables, figures and statistical analyses. There are also standards for describing mathematical models (e.g. CellML and SBML) and experimental protocols that could be used to help structure Methods sections. Rather than generate a plot offline, create the commands to generate the plot from raw data uploaded. The same for statistical analysis: choose from a menu of scripts or create your own R script that performs the statistical analysis. The result would be a conventional looking manuscript but with all the the raw data and analysis accessible beneath the formatted manuscript view. Journal reviewers and readers could then examine raw blots, micrographs or data (not just representative choices), see how results look if plotted a different way, re-run statistical analyses with different methods, search for other papers using similar protocols. This will also allow like types of data to be found and federated inot new databases.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

The cost of data storage will continue to decrease to the point where the cost of lost data will be greater than the cost of keeping it indefinitely.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

I think the assumption that data resources need to be large and centralized and therefore highly standardized and curated has raised the cost and resource requirements for new initiatives. If we captured more data at the source (i.e. the time of manuscript preparation) and used even available ontologies and standards to annotate them, we would build the foundations of specialized curated databases while streamlining science and promoting rigor, transparency and reproducibility

4. Any other relevant issues respondents recognize as important for NIH to consider

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

The current RPPR provides ample opportunity and encouragement to report data sharing. I think a much larger problem than the often suggested desire of scientists not to share data (I do not see this nearly as often as other suggest it happens) is the lack of an easy way to do it that falls within the normal flow of scientific data reporting. Much valuable data gets lost or condensed between the raw source and the final manuscript, yet enormous effort is devoted to generating manuscripts which conform to unnecessarily abridged and outdated conventions. The NLM has an opportunity develop a portal that would make the process of generating a manuscript simpler and avoid the need for authors to condense and selectively filter the data they submit while still resulting in a manuscript conforms to the expectations of reviewers, publishers and readers.

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

This is simple and worthwhile. RRIDs are easy to use and will only improve when more widely adopted.

b. Inclusion of a link to the data/software resource with the citation in the report

These should already be standard

c. Identification of the authors of the Data/Software products

The link can provide that information. If authors feel they need their name cited they can always write a paper and ask users of the resource to cite the paper.

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

4. Any other relevant issues respondents recognize as important for NIH to consider

Additional Comments

Submission Date

11/18/2016

Submitter Name

Charles Horn

Name of Organization

University of Pittsburgh

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

Neuroscience

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

This answer is largely dependent on research domain. Often data sharing is defined very narrowly in terms of genomics or clinical trial data; however, other types of data are of equal importance, including physiology and non-human animal.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

This is a very difficult question to answer. Twenty years might be appropriate.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

I do not see large barriers to providing this service. You should consider https://zenodo.org/ as a model.

4. Any other relevant issues respondents recognize as important for NIH to consider

None

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

There is a wide variability on the details provided by researchers on sharing of data and software in RPPRs, with little enforcement of standards. In many cases it is not clear that sharing can take place. And, if grantees choose not to share at some later time, is there anyone to check this activity?

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

This is essential. https://zenodo.org/ provides this service.

b. Inclusion of a link to the data/software resource with the citation in the report

This is also essential but it need to be link that is stable and well maintained.

c. Identification of the authors of the Data/Software products

Essential

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

No comment

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

No comment

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

You could create a section in the NIH Bio sketch that highlights this activity.

4. Any other relevant issues respondents recognize as important for NIH to consider

None

Additional Comments

Submission Date

11/20/2016

Submitter Name

Mary Carskadon

Name of Organization

EP Bradley Hospital Sleep Lab, Alpert Medical School of Brown University

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

Sleep and circadian rhythms

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

Sleep recording files that include raw physiological data, as well as processed (e.g., edf) files. The data must be kept with clear descriptions of recording parameters, participant/patient and other relevant information as to demographics (age, sex, sleep-schedule, prior sleep-wake, diagnosis, and any relevant information about the person (test scores, sleep-wake diaries, actigraph recordings).

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

This type of data is 'ever green' and should be kept indefinitely. The key is to have the data base well curated and wherever possible to have tools for interrogating the files and preprocessing, as well as performing exploratory analyses.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

For a resource such as the NSRR, adequate funding and committed personnel are critical. I am uncertain of the appropriate mechanism, i.e., whether NIH should support such a resource in the research community or house it intramurally. Ready and easy access is importing.

4. Any other relevant issues respondents recognize as important for NIH to consider

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

b. Inclusion of a link to the data/software resource with the citation in the report

c. Identification of the authors of the Data/Software products

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

4. Any other relevant issues respondents recognize as important for NIH to consider

Additional Comments

Submission Date

11/21/2016

Submitter Name

Chris Myers

Name of Organization

University of Utah

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

Synthetic Biology

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

Data stored in standard formats such as SBML, SBOL, BioPax, SED-ML, NeuroML, CellML, etc. that can be more easily utilized to reproduce scientific results in systems and synthetic biology.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

Perpetually. It should be stored in public repositories that have guaranteed funding to keep them up and running.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

The biggest barriers are funding support, data curation, and incentives to the data creators to store their data.

4. Any other relevant issues respondents recognize as important for NIH to consider

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

The impact would be higher quality research that is more easily reproduced. Any money spent to produce work that cannot be reproduced is wasted resources. Therefore, this should be a very high impact endeavor.

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

This would be very useful.

b. Inclusion of a link to the data/software resource with the citation in the report

Again, very useful.

c. Identification of the authors of the Data/Software products

Again, very useful.

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

Not entirely sure about this one.

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

Yes, this is useful too.

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

As mentioned above, making this a prime requirement for PIs and providing resources to validate the quality and reproducibility of the data is key.

4. Any other relevant issues respondents recognize as important for NIH to consider

Additional Comments

Submission Date

11/30/2016

Submitter Name

Concerned Citizen

Name of Organization

N/A

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

Nothing specific

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

Genomic data.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

Indefinitely.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

A culture of sharing needs to be developed, incentivized, and enforced.

4. Any other relevant issues respondents recognize as important for NIH to consider

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

b. Inclusion of a link to the data/software resource with the citation in the report

c. Identification of the authors of the Data/Software products

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

4. Any other relevant issues respondents recognize as important for NIH to consider

Additional Comments

Submission Date

12/01/2016

Submitter Name

kaare mikkelsen

Name of Organization

aarhus university

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

Biotechnology

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

generalized data with neuroscience/clinical relevance. I.e. data from broadly used paradigms, which may be combined with other sources.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

until the value of the data, relative to other comparable data sources, has diminished significantly. this could happen because better equipment becomes available, or the type of data "goes out of fashion".

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

narrow definitions of "anonymization".

4. Any other relevant issues respondents recognize as important for NIH to consider

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

b. Inclusion of a link to the data/software resource with the citation in the report

this is vital

c. Identification of the authors of the Data/Software products

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

this should be relative to the volume of the work in which the citation takes places, and how many works would have to be cited. no point in citing 100 sources by name in a small report, but for a large textbook, I see no problem in citing 20.

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

4. Any other relevant issues respondents recognize as important for NIH to consider

Submission Date

12/01/2016

Submitter Name

Brett Duce

Name of Organization

Sleep Disorders Centre, Princess Alexandra Hospital

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

Sleep and Sleep Disorders

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

Clinical data from large prospective trials. These type of trials, especially in sleep, are very few and extremely valuable. They are hard to design and develop and always are useful to other researchers.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

The data should remain available for at least 15-20 years after the trial has ceased. It allows researchers to test new perspectives or apply emerging paradigms to this data. It will also save money because as new insights into pathophysiological processes emerge, they can be tested against a "control" dataset.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

Cost is always going to be a factor. Imposing a cost on any researcher to access data is usually not workable - gets very difficult for cost centres in various institutions and countries to be able to pay in a practical sense. I would suggest a crowd funding arrangement. Make it voluntary like wikipedia and other such groups. I think they tend to get a lot more revenue in that form. The people that are using it can then make nominal donations to keep it going. You owuld surprised how many people would donate even if they don't really use the data just because they view this as valuable to the research area.

4. Any other relevant issues respondents recognize as important for NIH to consider

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

An extended guideline should developed to improve the way in which all grantees report on the development of such data, databases and software. Perhaps with respect to databases they must also try to provide an overview of the table structure and table relationships.

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

Essential

b. Inclusion of a link to the data/software resource with the citation in the report

Essential

c. Identification of the authors of the Data/Software products

Audit trail must be open to all - unless some aspects are particularly sensitive.

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

Thsi is useful as sometimes a number of trials are added together to capture a number of demographics. However it is important to have that granularity to allow researchers to specific examine a particular demographic or locality.

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

Essential

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

Collaboration with other large granting bodies around the world eg NHMRC in Australia

4. Any other relevant issues respondents recognize as important for NIH to consider

Additional Comments

Submission Date

12/04/2016

Submitter Name

Alexander Tsai

Name of Organization

Massachusetts General Hospital and Harvard Medical School

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

Psychiatry

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

4. Any other relevant issues respondents recognize as important for NIH to consider

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

b. Inclusion of a link to the data/software resource with the citation in the report

c. Identification of the authors of the Data/Software products

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

4. Any other relevant issues respondents recognize as important for NIH to consider

Additional Comments

Submission Date

12/05/2016

Submitter Name

Leila Tarokh

Name of Organization

University Hospital of Child and Adolescent Psychiatry and Psychotherapy

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

Sleep

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

Sleep and neuroimaging data combining psychiatric diagnoses

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

New analysis methods are continually being developed, with emerging research questions. Therefore, a span of at least 10 years is required.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

Typically, the PI needs to publish extensively from the data set, so that can often be a barrier to sharing the data

4. Any other relevant issues respondents recognize as important for NIH to consider

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

b. Inclusion of a link to the data/software resource with the citation in the report

c. Identification of the authors of the Data/Software products

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

4. Any other relevant issues respondents recognize as important for NIH to consider

Additional Comments

Submission Date

12/05/2016

Submitter Name

Jeffrey Petruska

Name of

Organization

University of

Louisville

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

Neurobiology

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

Large datasets of any sort (imaging, -omics, electrophysiology). SAVES MONEY and TIME!

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

Fully available at least 5 years after embargo is over, for 10 additional years archived and accessible. Maintenance should be by centralized infrastructure (NOT individual labs!), which can be either Federal or Private-Public partnership. Long-term implications include better use of expensive data.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

Dedicated staff and infrastructure. This is not commonly-available for individual labs, or even Universities. Support can come from fee tagged onto qualifying grants. If a qualifying dataset is generated, it must be deposited AND the funding agency alerted so that they can dedicate some set amount to the sustaining agency.

4. Any other relevant issues respondents recognize as important for NIH to consider

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

The impact is likely to be minimal without a centralized structure in which items are deposited, or a description and link is deposited. Offer a mechanism and guidance, but do not mandate some enhanced level of reporting. If people want to gain the advantage, they will participate.

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

These are a must. ALSO, the depositor MUST provide accurate and useful descriptions of the deposited materials. This MUST have a curated commentary sections as well, to allow for user feedback. There should also be back-end repercussions if there is enough believable user feedback suggesting there is a problem with the data descriptions and organization. This could simply be the possibility of opening an investigation into data integrity.

b. Inclusion of a link to the data/software resource with the citation in the report

MUST

c. Identification of the authors of the Data/Software products

MUST

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

4. Any other relevant issues respondents recognize as important for NIH to consider

Additional Comments

Submission Date

12/05/2016

Submitter Name

Logan Schneider

Name of Organization

Stanford University

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

Sleep

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

Polysomnograms (PSGs). Polysomnograms are very expensive and labor-intensive studies resulting in a limited number of studies being able to accumulate large data sets. Furthermore, the relative dominance of obstructive sleep apnea as a sleep disorder results in a lack of robust data sets representative of the spectrum of sleep disorders or as a repository of convenience controls.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

Because the raw data are timeless, and it's truly the interpretation that changes, indefinite warehousing of the raw data (usually in EDF format) would be ideal, whereas scoring files could be made available upon request, once the scoring definitions/criteria change (approximately every 5-10 years).

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

Barriers include the IT infrastructure and staff needed to curate, store, and protect the data. Centralizing this process, will allow for not only a finite staff, but can also ensure standardization of data quality and sharing procedures.

4. Any other relevant issues respondents recognize as important for NIH to consider

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

b. Inclusion of a link to the data/software resource with the citation in the report

c. Identification of the authors of the Data/Software products

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

4. Any other relevant issues respondents recognize as important for NIH to consider

Additional Comments

Submission Date

12/06/2016

Submitter Name

Anna May

Name of Organization

University Hospitals Cleveland Medical Center

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

sleep medicine

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

sleep study and actigrpahy data are of high priority additional clinical data (CPAP/APAP/BPAP adherence) would also be very valuable

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

at least 10-20 years

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

server space is costly, obtaining IRB approval for fully de-identified datasets is...wasteful of everyone's time and effort since these are exempt datasets

4. Any other relevant issues respondents recognize as important for NIH to consider

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

b. Inclusion of a link to the data/software resource with the citation in the report

c. Identification of the authors of the Data/Software products

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

4. Any other relevant issues respondents recognize as important for NIH to consider

Additional Comments

Submission Date

12/07/2016

Submitter Name

Ellen Wijsman

Name of Organization

University of Washington

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

Genetics/genomics of neurosciences-related disorders

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

Genomic data (from the whole genome, only), including STR data, SNP data, DNA sequence data (WGS, WES). Data on unrelated subjects and pedigree data stored in a standard format that expands to large pedigrees (family, id, father_id, mother_id, sex), necessary demographic information, phenotype data.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

15-20 years, electronically, should be stored at dbGaP. There should be one, and ONLY one, unit at NIH that is responsible for databasing genomic data. It is inefficient to have multiple sites doing this because of the steep learning curve and need for a fairly large staff. Also, my experience with the multiple units that currently do this suggests that the only unit that is doing the work in a fashion that provides reliable data is dbGaP. Regarding data use, it seems to me that investigators should not have to resubmit a renewal for data use for 3 years. A 1-year cycle adds huge burden on the local IRBs and investigators, with no obvious gain in terms of scientific outcome.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

There are problems at NIH with communication between units that report to different institute directors. As a result, parallel (but operationally very different) data collection systems get put into place by different institutes. This adds burden and cost on investigators who need to prepare for different institution-specific submission procedures. It seems difficult for NIH to come to a solution regarding how to share (coded) IDs between dbGaP and the institutes that fund the phenotypic data collection, which may be a longitudinal real-time data set that does not fit neatly into the setting used by dbGaP. A solution would be to figure this out so that the same ID can be used by a phenotype-specific data base at the funding institute, and a single data request could be made to allow access to both phenotype data through an institute-specific database, and genomic data through dbGaP. Right now this does not seem to be possible, but as someone who does a lot of database and other computer work, I cannot see why it should not be possible.

4. Any other relevant issues respondents recognize as important for NIH to consider

One problem I see regarding data re-use is that NIH seems to be willing to fund the cost of data collection, but not the cost of data analysis or development of analysis tools. The assumption seems to be that data submitted for sharing can be used quickly and easily right out of the box. This is often the vision of clinicians and other people who are not skilled data analysts and/or programmers. However, in my experience (as someone who has been responsible for data analysis and databasing for >30 years), it takes awhile to get to know a data set, and no competent statistician will analyze a data set without some at least minimal QC analyses, because so often there are problems with at least some of the data. To get information out of re-use, or to encourage making software available, means recognizing the real costs involved in these activities. The statisticians, bioinformaticists, and computer programmers should not have to work for free when the laboratory and clinical scientists get funded to collect the data

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

Make no changes to current reporting procedures. This all sounds like added paperwork, with no obvious useful outcome to the scientific world. Quite a lot of software is announced initially as part of a publication - if it is useful, it will get cited. Word of mouth also spreads software availability. What NIH *could* do if they want useful software, is to provide smallish grants for software maintenance. Initial releases of software are rarely already fully developed. Someone who writes a computer program is more likely to put out additional versions (that are increasingly useful to the community) if (1) initial releases are easy, (2) they can evaluate interest through user feedback, and (3) they have a source of funds to maintain and improve the programs that look like they are filling a useful niche.

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

I do not see how this would be useful. There are a number of online open software repositories that have their own ways to do this. Don't add another.

b. Inclusion of a link to the data/software resource with the citation in the report

Sure, this would be trivial, and not much of a burden.

c. Identification of the authors of the Data/Software products

It wouldn't be hard to add this to the reporting, but I don't see how it would make any difference to the scientific community.

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

This is too study-specific to comment on in detail. Given the time and effort it takes to submit (or retrieve) data sets, I don't think most very small data sets are worth trying to collect for public sharing. But some studies have a data collection phase followed by an analysis phase, while others collect data continuously over time, with periodic freezes of the data for analysis purposes.

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

How and where software is stored is a moving target. Asking for information about how to obtain the software is fine, but let the authors decide the platform and details.

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

See comments above: provide longer-term support for development and maintenance of software, if you want to see it used. Documenting and maintaining software is a really big job. Lots of people will contact the author for support (many of whom have absolutely no computer skills), and responding takes a lot of time or at least staff support to help. Not every piece of software should be shared! If the software hasn't been thoroughly tested and documented, it can be dangerous to provide to naive users.

4. Any other relevant issues respondents recognize as important for NIH to consider

One problem I see regarding data re-use is that NIH seems to be willing to fund the cost of data collection, but not the cost of data analysis or development of analysis tools. The assumption seems to be that data submitted for sharing can be used quickly and easily right out of the box. This is often the vision of clinicians and other people who are not skilled data analysts and/or programmers. However, in my experience (as someone who has been responsible for data analysis and databasing for >30 years), it takes awhile to get to know a data set, and no competent statistician will analyze a data

set without some at least minimal QC analyses, because so often there are problems with at least some of the data. To get information out of re-use, or to encourage making software available, means recognizing the real costs involved in these activities. The statisticians, bioinformaticists, and computer programmers should not have to work for free when the laboratory and clinical scientists get funded to collect the data.

Additional Comments

Submission Date

12/07/2016

Submitter Name

gianluigi delucca

Name of Organization

medicoimpianti

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

Sleep

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

Polysomnography. It is the only type of analysis that allow multiple systems/components of real time physiology available in and out controlled environnements both as long term (several years) and short term (seconds) time base.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

Forever. One night without video would be half a Gigabyte.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

The format of the data is often manufacturer related. Controlled transfer to EDF is often an issue.

4. Any other relevant issues respondents recognize as important for NIH to consider

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

Raw data sharing is the main point. First because it is the only way to replicate the analysis, second because other ideas and/or methods may be applied anytime in the future.

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

b. Inclusion of a link to the data/software resource with the citation in the report

c. Identification of the authors of the Data/Software products

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

4. Any other relevant issues respondents recognize as important for NIH to consider

Submission Date

12/07/2016

Submitter Name

Eyleen O'Rourke

Name of Organization

Univ of Virginia - Biology

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

Genetics of obesity and aging

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

Sequencing data (genomics,transcriptomics, ChIPseq, etc) in addition to image-based screening data should be prioritized. There is lots of redundancy in sequencing projects, thus: 1) sharing would enable cross-validation which would increase accuracy and strength of conclusions, 2) researchers can mine data sets that no single research lab could afford to generate. High-content screening (HCS) data sharing would enable re-screening for alternative phenotypes without redoing costly primary screens.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

NIH should create a searchable data base to enable searches for single genes and single compounds, or sets of genes with common features (every gene responding to ER insult or all compounds affecting ER function). The output would for example be: 1) list of sequencing or HCS projects in which the query genes were significantly associated to disease, or differentially expressed after drug or other treatments. In addition to single gene or single drug queries, users should also be able to retrieve all gene inactivations or compounds affecting a process (i.e. all compounds compromising DNA stability, mitochondrial function, etc.) This user friendly searchable website should enable uploading and downloading of large data sets.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

Uploading terabytes of sequencing or HCS data would be time consuming and costly with current network technologies. Approaches similar to Amazon snowball would be an alternative to standard transfer over the internet.

4. Any other relevant issues respondents recognize as important for NIH to consider

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

b. Inclusion of a link to the data/software resource with the citation in the report

c. Identification of the authors of the Data/Software products

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and

when might each distinct data set underlying a study be cited and reported separately

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

NIH should provide the actual language, general format, and web framework. An NIH repository with searchable tools were researchers are mandated to upload the data seems a long-term solution since it does not depend on single labs or institutions willingness or capacity to make raw data publicly available.. However, this site has to provide user-friendly easy to search, upload, and download data tools. Otherwise it would not work/be used. GEO is an example of a system that is not working well for the research community. Only the top 250 genes can be retrieved and the process is cumbersome and slow. Other researchers may need the full dataset to rerank the data for different research questions.

4. Any other relevant issues respondents recognize as important for NIH to consider

Additional Comments

Submission Date

12/07/2016

Submitter Name

FASEB

Name of Organization

Federation of American Societies for Experimental Biology

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

Biological and biomedical sciences

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

Improving data management and increasing data access can create new scientific opportunities. The Federation of American Societies for Experimental Biology (FASEB) agrees with NIH that many factors must be considered when determining what data should be shared and under what conditions. We also applaud NIH for seeking to identify the types of data that are of the highest priority to share. For this purpose, FASEB recommends the potential utility of a dataset to be used as a major criterion. NIH's approach to data sharing should recognize that access to some datasets may not be worth the cost of sharing and long-term preservation. (Please see our attached Statement for additional information.)

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

The diversity of data types, research areas, and resources available make it challenging to identify data accessibility strategies that are practical and relevant for all fields of research supported by NIH. Therefore, the Federation of American Societies for Experimental Biology (FASEB) encourages NIH to employ flexible approaches that allow investigators and NIH to establish reasonable expectations for a particular research project. Data management plans (DMPs) are an important tool for clarifying these expectations while still providing flexibility at the research project level. (Please see our attached Statement for additional information.)

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

Making datasets accessible – including the skilled human labor necessary to prepare and maintain data, technological infrastructure, and continued development of databases and effective search platforms – is costly. Expansion of requirements for data sharing will require commensurate financial and staff support from research sponsors. The Federation of American Societies for Experimental Biology (FASEB) affirms that NIH should, at a minimum, provide sufficient support to fully comply with all applicable data management and access requirements as part of a project's funding. Additional resource needs and barriers exist; to identify common challenges, NIH can examine data management plans (DMPs) in aggregate and continue seeking feedback from the scientific community. (Please see our attached Statement for additional information.)

4. Any other relevant issues respondents recognize as important for NIH to consider

Integrated community-based solutions are needed to enhance data management and access. The Federation of American Societies for Experimental Biology (FASEB) encourages NIH to work with the research community and other stakeholders to develop and refine its approaches to data sharing. Research sponsors, investigators, institutions, and scientific journal can all contribute to and benefit from advancing data management and access. Coordination with stakeholders should also extend to other biological research sponsors. Many challenges and needs are cross-cutting. NIH can harmonize policies for data management plans (DMPs) with other federal agencies as well as develop unified portal

systems for data discovery. NIH and other sponsors should also conduct continuous assessment of data sharing policies and requirements to ensure that they do not delay the adoption of improved practices or new technologies. (Please see our attached Statement for additional information.)

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

Data and software citation in RPPRs and competing grant applications can serve as a limited source of professional recognition for investigators that share these research products. (Please see our attached Statement for additional information.)

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

b. Inclusion of a link to the data/software resource with the citation in the report

c. Identification of the authors of the Data/Software products

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

To incentivize data sharing, the Federation of American Societies for Experimental Biology (FASEB) recommends that NIH take actions that maximize the value of shared datasets and assure professional recognition for making them accessible. Investigators have little incentive to share data if they do not think other researchers will be able to find, use, or cite it.
FASEB recognizes NIH's efforts in these areas and encourages NIH to build upon this important work. NIH's support for the development and provision of resources and services that enable data sharing should be sustained. This includes providing funding for databases and the development of training modules as well as NIH staff assistance to identify available resources and repositories. Continued effort is also needed to promote productive data sharing, from building better systems for data discovery to facilitating the development of community-based data standards. As noted in this request for information, data citation can encourage sharing by ensuring that credit is attributed to the investigators responsible for generating the data. There are further strategies that NIH can employ to advance citation practices. For example, all NIH databases and data catalogs could provide standardized, exportable citation information for datasets. These export tools make it easier for investigators to accurately cite data and provide a visible reminder to do so. (Please see our attached Statement for additional information.)

4. Any other relevant issues respondents recognize as important for NIH to consider

Integrated community-based solutions are needed to enhance data management and access. The Federation of American Societies for Experimental Biology (FASEB) encourages NIH to work with the research community and other stakeholders to develop and refine its approaches to data sharing. Research sponsors, investigators, institutions, and scientific journal can all contribute to and benefit from advancing data management and access. Coordination with stakeholders should also extend to other biological

research sponsors. Many challenges and needs are cross-cutting. NIH can harmonize policies for data management plans (DMPs) with other federal agencies as well as develop unified portal systems for data discovery. NIH and other sponsors should also conduct continuous assessment of data sharing policies and requirements to ensure that they do not delay the adoption of improved practices or new technologies. (Please see our attached Statement for additional information.)

Additional Comments

FASEB Statement on Data Management and Access.pdf (434 KB)

FASEB
Federation of American Societies for Experimental Biology

Representing Over 125,000 Researchers

301.634.7000
www.faseb.org

9650 Rockville Pike
Bethesda, MD 20814

March 1, 2016

FASEB Statement on Data Management and Access

The Federation of American Societies for Experimental Biology (FASEB) affirms the importance of data management and access to scientific progress. Good data practices are necessary to achieve the maximal benefit of research for all stakeholders. Technological advances are expanding the ways investigators collect, utilize, and share data, leading to new knowledge and discovery. The diversity of data types, research areas, and resources available make it challenging to identify data management and accessibility strategies that are practical and relevant for all life science fields. Therefore, FASEB advocates for flexible and customizable approaches that allow investigators and research sponsors to establish reasonable expectations for a particular research project. Moving forward, the scientific enterprise will need to develop integrated community-based solutions. The following principles and recommendations are meant to help guide stakeholder efforts to advance data management and access in the biological and medical sciences.

Guiding Principles

- Improving data management and increasing data access can create new scientific opportunities
- Efforts to increase data access should consider the infrastructure required for data management, standardization, discovery, access, citation, reuse, sustainability, and long-term preservation – all of which are necessary for productive data sharing
- Expansion of requirements for data management and access will require commensurate financial and staff support from research sponsors
- Flexibility and adaptability are essential for any data management or access policies; the varied and rapidly evolving data landscape necessitates customized strategies
- Efficient and enabling data access requires a balanced approach that prioritizes datasets of high potential utility; policies should recognize that access to some datasets may not be worth the cost of sharing and long-term preservation
- Regulatory and administrative burden should be minimized so that any requirements ultimately promote science rather than hinder research

I. Data Management Plans

Data management plans (DMPs) are an important tool for promoting quality data management and appropriate data access. They can serve as a helpful planning exercise at the beginning of a project and focus attention on data sharing goals. Submission of a DMP can clarify expectations between investigators and their research sponsor. Flexibility and adaptability can be achieved by having individual investigators develop a DMP specific to their research area, data types used, and resources available.

Research sponsors may also enlist DMPs for secondary uses of benefit to the research community, such as identifying common resource needs and other barriers.

1. DMP Requirements: To attain the benefits of DMPs without creating unnecessary burden, DMPs should be short summary documents that address the most essential aspects of data management and access. In most cases, one to two pages should be sufficient, although additional information could be requested *just-in-time* for select circumstances. FASEB recommends the following DMP content requirements:

a. Description of the data and metadata to be collected
b. Overview of data management practices
c. Summary of any data sharing restrictions (confidentiality, intellectual property, etc.)
d. For **shared data**, information about when it will be made available, where it will be stored, how it will be maintained, and how others will be able to find, access, and reuse it
e. For **data that will not be shared**, justification for not making it accessible (which may include considerations of feasibility, data utility, etc. as well as sharing restrictions)

2. DMP Compliance Reporting: It will take time for funding sponsors and research communities to establish what constitutes reasonable practices and expectations for the many different areas of research and types of data. Therefore, FASEB recommends that research sponsors delay any DMP enforcement actions for the first five years. During this time, sponsors should:

a. Identify and address common barriers and emerging problems
b. Establish a process for modifying or updating DMPs
c. Ensure there is sufficient flexibility and adaptability built into all requirements
d. Standardize policies and reporting requirements with other sponsors (particularly among federal funding agencies)
e. Outline the roles and responsibilities of all parties for data management after the grant ends

Once sponsors establish harmonized and well-vetted DMP policies, continuous assessment will be necessary to ensure that the policies do not delay the adoption of improved practices or new technologies.

II. Roles of Stakeholders in Improving Data Management and Access

FASEB recognizes that improving data management and access is an important and evolving challenge for the research community. As science and technology advances, so must data practices. Research sponsors, investigators, institutions, and scientific journals can contribute to and benefit from advancing data management and access strategies.

A. Research Sponsors

1. Incentives: Sponsors should encourage investigators to improve management of and access to their research data. This should include actions that maximize the value of shared data and assure professional recognition for making datasets accessible. As a first step, research sponsors should provide investigators with resources and services, including:

a. Creation of and continued support for additional databases and repositories
b. Help identifying any relevant databases, repositories, and other resources
c. Development of training modules on data management and sharing
d. Assistance in procuring a unique digital object identifier (DOI) for shared datasets

In the long-term, research sponsors should also address the resources and measures needed to promote productive data sharing:

a. Development of a single unified portal system to discover datasets and a unified system for metadata submission to data catalogues
b. Advancement of data citation practices, including the provision of standardized, exportable citation information for datasets included in the sponsor's data catalogues or databases
c. Facilitation of the development of community-based data standards by convening stakeholders, as needed. Scientific societies can assist by identifying experts, providing thoughtful feedback, and disseminating proposed standards
d. Provision of long-term data storage options when no relevant database exists (such as sponsor-based databases or "dark" storage to serve as a back-up)

2. Policies and Rules: Research sponsors also can use policies and rules to ensure that data produced through supported research are appropriately managed and made accessible. FASEB affirms that a sponsor's expectations should be commensurate with the resources available to the investigator and the sponsor's own support for such resources. At a minimum, sponsors should provide sufficient support to fully comply with all applicable data management and access requirements *as part of a project's funding*.

To effect positive change, research sponsors must also carefully balance the costs and benefits of data access when developing and amending policies. Making datasets accessible – including the skilled human labor necessary to prepare and maintain data and metadata, technological infrastructure, and continued development of effective search platforms – is costly. Some datasets have little value for reuse or a short "shelf-life"; requirements to share and preserve such data could create inefficiencies in research funding and resource distribution. FASEB recommends that sponsors ensure their data access policies prioritize data with the highest potential for reuse.

B. Investigators

3. Data Management: Quality data management is an essential component of productive data sharing. Poor practices can render a potentially valuable dataset useless. At a minimum, FASEB recommends that investigators ensure the following data management practices are established and maintained within their own laboratory:

a. Regular back-up of *digital* data onto a well-maintained server, cloud, or separate machine (ideally an automated process utilizing offsite backup storage)
b. Standardized meta-data collection and documentation for *common* data types used or produced within the laboratory
c. Sufficient documentation to facilitate dataset retrieval several years after collection
d. When possible, use of unique identifiers in metadata fields (e.g., an ORCID iD for individuals)
e. Prompt training of all research team members on the laboratory's data management practices

4. Data Sharing: If there are no restrictions or other considerations that would preclude sharing, investigators should submit *key* data from their research to a relevant database or repository. If no publicly-accessible topical or data type-based repository exists, investigators should establish plans for making the data available. This might include sharing upon request, publishing supporting data in the supplemental materials of an article, or depositing data into a non-specific database.

C. Scientific Journals

5. Role in Compliance: The point of publication occurs too late in the research process to effectively address many issues related to good data practices. Furthermore, most journals do not have the capacity to confirm author compliance with any applicable DMPs and policies. Therefore, FASEB strongly recommends that the federal government and other research sponsors avoid requiring journals to assure compliance with DMPs, confirm data are and remain accessible, or provide database services. Such activities would be more effectively and efficiently managed by the sponsor and grantees.

6. Professional Norms: FASEB affirms the role of scientific journals in promoting good practices and encourages them to: (1) request that authors include the DOIs for and/or web addresses of datasets in their original manuscript; and (2) uphold that, in the absence of any restrictions on sharing or similar concerns, investigators are responsible for making the underlying data available upon request.

D. Research Institutions

7. Resources: Institutions should provide the technological infrastructure necessary for quality data management and compliance with DMPs. At a minimum, investigators must be able to attain the professional data norms within their field of research.

8. Professional Culture: Institutions should also foster an atmosphere where of quality data management and appropriate data sharing are standard practice. To establish and maintain such an environment, FASEB recommends that institutions ensure the following:

a. Appropriate data training is available for all individuals conducting research
b. Institutional resources for data management can be easily identified and utilized
c. Investigators are encouraged to collaborate on improving data practices at the institution and within their discipline.

Submission Date

12/07/2016

Submitter Name

David H. Barker

Name of Organization

Rhode Island Hospital / Alpert Medical School, Brown University

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

Pediatric Psychology

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

Individual data from clinical trials to help identify heterogeneity in treatment effects.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

Consider a criterion other than time. Perhaps the maturity of the research questions that the data can address. For example, areas with sparse information may need to have the data retained for more time than research areas that have an abundance of data.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

My experience is that most clinical researchers do not allocate enough time or resources on a grant to ensure that the data are prepared for general use. It would help to have standards for the data and meta-data, a common resource to store the data perhaps by grant number, and explicit budget lines for data management.

4. Any other relevant issues respondents recognize as important for NIH to consider

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

Collecting, cleaning, analyzing, and preparing data for general consumption can be a long process. Clear benchmarks would help track progress in the endeavor. Clear standards for what product is expected at the end of the process would also help. Finally, having a location to either upload the data or instructions on how to access the data (beyond contacting the PI) would be helpful.

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

A DOI would be helpful, but data and software development is often iterative. Clear standards around what is required before a DOI is assigned would be helpful.

b. Inclusion of a link to the data/software resource with the citation in the report

If this is required, it will be important to have a secured location for investigators to upload their data or require institutions to provide the service for investigators. It will be a burden to have investigators managing the security, reliability and access to the data.

c. Identification of the authors of the Data/Software products

Great idea

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

The number of unique data citations could be part of the grant proposal--investigator teams are likely to be the best at identifying meaningful distinctions within their field of research.

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

There need to be common repositories that investigators can use to house the data. The challenges of maintaining sharing capacity when there are changes in location, funding, and institutional policies are daunting.

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

Ultimately, the largest incentive will be recognition for developing usable data that is used and cited by others.

4. Any other relevant issues respondents recognize as important for NIH to consider

Additional Comments

Submission Date

12/09/2016

Submitter Name

Dr. Ramanathan Natesh

Name of Organization

Indian Institute of Science Education and Research Thiruvananthapuram

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

Structural Molecular Biology, Infectious Disease, Cancer, CVD.

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

Government funded research results to be given to academic users. Value is societal obligation for peoples' tax money funded research.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

Patent need's may be considered for limited no. of years. Example 2 years to maximum 5 years. But for academic use there should be no restrictions of new findings.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

Government must fund institutions like NIH, EBI and it is very important for scientific community and humanity that benefits from basic research. Should not expect short term benefits. Should think of ultimate goal and health benefits to humanity in long term.

4. Any other relevant issues respondents recognize as important for NIH to consider

All data relevant to academic research funded by tax payers money has to be in public domain. However the researcher may be given some time say 2 to 5 years in order to fully utilise the potential and also raise future funding for the lab.

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

Absolutely important to report/cite the usage of public databases and software and also share the results obtained by an academic research funded by public money. Some liberty of 6 months to 1 year may be given to researcher themselves, as to when to report from the time of novel results obtained.

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

Due to increacing degeneracy in many different softwares/databases, just like ORCID if there are DOI's that will enhance the productivity and identification of the exact data/software.

b. Inclusion of a link to the data/software resource with the citation in the report

Absolutely. It will be good to included such link to resources

c. Identification of the authors of the Data/Software products

Yes. But this will be available in the link for data/software.

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

Granularity is needed only for detailed researcher, eg. Scientific Grant Officer or database of funders. But for funding application reviewers this may not be that important. Individual database/software sites can maintain high resolution information similarly funding agencies can also maintain Granularity or high resolution in terms of data citations. Some sort of metrics about the applicant can be developed by database researchers or Scientific Grant officers in order to be maintained at the funding agencies site and can be provided to Grant application Reviewers.

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

YES Very important and top priority must be given to this.

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

Not sure, how to do this, but it is an important topic to be considered. Porbably incentivize data after the term of project would be a better option.

4. Any other relevant issues respondents recognize as important for NIH to consider

All data relevant to academic research funded by tax payers money has to be in public domain. However the researcher may be given some time say 2 to 5 years in order to fully utilise the potential and also raise future funding for the lab.

Additional Comments

Submission Date

12/12/2016

Submitter Name

MacKenzie Smith

Name of Organization

UC Davis Library

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

Data Management

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

This and the following responses are submitted on behalf of the UC Davis Library. Data sharing serves two main goals: enabling its reuse and enhancing reproducibility. While, data sharing as a means to facilitate reproducibility has great promise for enhancing the transparency, quality, and reliability of scientific information (e.g., by encouraging due diligence on the part of the researchers), opening the door to reuse also serves to lay the groundwork for new, and possibly unanticipated, research and discoveries. Accordingly, data with the most promise for reuse should be prioritized even as we work to encourage widespread sharing for reproducibility purposes across other kinds and classes of data. In our view, the two categories of highest potential for reuse are Big Data (>100 GB) and longitudinal studies. Because of their scope, both categories are likely to be sources for analysis beyond the originally planned project. Big Data are likely to contain multiple unexplored dependencies. Longitudinal studies by their nature will contain data relationships that may be understood only after additional time and as a product of meta-analysis. In some cases, longitudinal studies are conducted around a specific historical event, thus impossible to collect again in the future. In addition, collecting such data is costly both in resources and time required to complete the dataset, making its subsequent reuse particularly efficient.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

Factors limiting the long-term sharing of data are cost, metadata sufficiency, and the ability to discover and access the data file in the long-term. Metadata, access and preservation are addressed in section I.4. Because the costs of storing and making available Big Data are the most likely to prejudice its long-term availability, particular attention must be paid to the benefits of its continued availability. Therefore, we recommend that the NIH develops standards for data decommissioning. For example, Big Data should remain available until a dataset that covers the same parameters and variables but of higher quality is obtained and be favored for analysis. While each dataset has historical value, the decision to maintain and sustain the dataset for beyond a reasonable middle-term timeframe should be left to those bearing the costs of its continued availability. As storage cost decreases in the future, it may become a non-issue, and decommissioning need may cease. Longitudinal data, meanwhile, are often more reasonable in size (between 1 and 100 GB) and are not likely to pose a financial challenge. To the contrary, those datasets are often very expensive to reproduce, offsetting their more modest costs. Such datasets should be available indefinitely. The biggest burden on sharing small data (<1 GB) from individual experiments is the time cost of generating sufficient metadata for independent long-term reuse. The storage costs are negligible.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

The main obstacles to data sharing are the costs (in both time and treasure) associated with curation, discovery, and storage of datasets. These should be addressed and accounted for up front, with the costs of preservation and curation reflected in the budget accompanying the proposal. Other recommended best practices follow below: Subject-level

repositories should be chosen whenever possible due to well-established curation practices. If not possible, general purpose data repositories that have been vetted by the NIH or meet certain independent standards, such as Data Seal of Approval, should be used. Since storage costs can be significant with large datasets, we recommend proposals to be accompanied by data management plans that highlight the size of the total data collection and the biggest expected size of a single file. For files of size greater than 2 GB or collections greater than 20 GB, we recommend requiring that researchers discuss their datasets with their repository of choice to confirm the deposit can be accommodated and to determine the projected costs of its preservation over time. For longitudinal studies, it is important to carefully reflect on the variables that will be studied. We recommend those variables to be listed in a data management plan prior to the start of the project. Simple spreadsheets are not effective tools for managing these kind of data due to the number of entries over time and the lack of version control or logging. For such studies, we recommend databases be used, and associated costs be accounted for.

4. Any other relevant issues respondents recognize as important for NIH to consider

In order for data sharing to realize its potential as an enabler of both reuse and reproducibility, it will be important to remove outstanding concerns and inconsistencies regarding rights and licensing. Presently, much shared data are subject to ad hoc licensing arrangements that encumber subsequent reuse in an unpredictable and often unclear manner. To resolve these uncertainties and facilitate the goals of data sharing, we fully endorse the findings of the RDA-CODATA "Legal Interoperability Of Research Data" principles and guidelines (available at https://www.rd-alliance.org/rda-codata-legal-interoperability-research-data-principles-and-implementation-guidelines-now) that call for the application of public domain waivers to shared data, and particularly the application of the Creative Commons Zero waiver (CC0). Finally, data sharing can, in some cases, have substantial privacy and patient privacy complications. Mechanisms to ensure that data sharing plans consider and address any such concerns in advance of data being made publicly available will be essential to the long-term success of any such program.

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

Inclusion of software and data products in RPPR is an excellent first step. To achieve the desired impact NIH needs to be transparent in how it values datasets and software. Will a citation of a dataset of software be given the same weight as a publication or a fraction of thereof? This is an opportunity for NIH to underscore the importance of non-traditional research products that can be reused and therefore enable new research beyond the completed project. One of the challenges to recognizing software and data as valuable research output is the cultural norm to only value publications and refer to the use of data and software through the initial publication that described them. In order for this norm to change, software and data need to receive recognition divorced from traditional publications. If non-peer-reviewed journal article products of research are to be integrated in the RPPR, then researchers will likely benefit from an automated process for feeding in citations, especially if documentation of compliance is desired. For instance, the NIH public access policy has an automated documentation process for publications in the RPPR via MyBibliography. We recommend an integration of MyBibliography with ORCID or DataCite to automate citing datasets and software.

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

We strongly recommend the use of permanent identifiers, e.g. DOI, in data/software citations. Data and software need to be recognized apart from the publication in which they are first described. Only then we can trace and place value in those products and determine how they compare to traditional publications. In addition, the registration of DOIs with DataCite enables data discoverability--one of the greatest challenge for modern research. We would like to draw the attention of NIH to the following citation recommendations: Dataset citation recommendations https://www.force11.org/group/joint-declaration-data-citation-principles-final https://www.datacite.org/cite-your-data.html The Research Data Alliance standards for dynamic dataset citation: https://www.rd-alliance.org/group/data-citation-wg/outcomes/data-citation-recommendation.html Force11 software citation recommendation:

https://www.force11.org/software-citation-principles

b. Inclusion of a link to the data/software resource with the citation in the report

A DOI by definition will be associated with a persistent URL. Therefore, if DOI is used as per our above recommendation, the inclusion of a link will be redundant.

c. Identification of the authors of the Data/Software products

Authors of Data/Software products should be unambiguously identified, similarly to publications. To this end, we highly recommend requiring the use of unique identifiers for all authors, preferably an ORCID id. We recommend that human contributors to the work be recognized as authors. Where a non-author party, such as an employer or institution, claims rights over a data or software product, we recommend providing an additional identification crediting those rights. In any case, information regarding institutional affiliation is useful for host institutions, helping compliance monitoring efforts, awareness of research outputs, and archival efforts; including such information should be encouraged in all cases.

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

The principles we would hope to see in data citation practice are: 1. Cite to all third-party data sets, subject to the below provisos. 2. Avoid citation stacking. The most successful and productive uses of open Big Data might involve integrations of already integrated data sets, creating the possibility of hundreds, if not thousands, of contributors. As a general rule (subject to norms-based exceptions), citations should be made only to the data sets as used by the researcher. That is, if a researcher makes use of an aggregation of data sets prepared by third party, the citation should be only to the aggregation rather than to the authors of its constituent parts. 3. Allow norms to govern and let them evolve. Citation practice in other research outputs has historically been norms-based, evolving to meet researcher and community needs. Presently, the practicalities of data and software citation are in flux, and practice should be allowed to shift in response to changing tools, methods, and norms. The NIH is well positioned to help provide baseline guidance at this crucial stage for data and software as research outputs, but should take care to provide the needed flexibility in the details of this practice.

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

Our recommendations above reflects our view that DOIs are the best available means to ensure clear and accurate citation practice, and to identify the whereabouts of a given resource, and should be used wherever available. Naming and identifying the repository separately from the DOI might also be of some value in increasing awareness of and information about available repositories, and certainly would be an appropriate supplement to a citation. Moreover, where DOIs or other persistent unique identifiers are unavailable or unavailing, citations should provide a URL and other information, including identifying the repository holding the best and most available copy.

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

NIH has the power to emphasize the importance of data and software sharing by granting them higher impact in deciding to award a grant. It can also set an example for using those products when deciding on the tenure and promotion of its intramural scientists. The example set by NIH is likely to be followed by many departments that receive their funding predominantly from NIH. The impact of publications is measured through citations; the impact of datasets and software should be measured by reuse.

4. Any other relevant issues respondents recognize as important for NIH to consider

In order for data sharing to realize its potential as an enabler of both reuse and reproducibility, it will be important to remove outstanding concerns and inconsistencies regarding rights and licensing. Presently, much shared data are subject to ad hoc licensing arrangements that encumber subsequent reuse in an unpredictable and often unclear

manner. To resolve these uncertainties and facilitate the goals of data sharing, we fully endorse the findings of the RDA-CODATA "Legal Interoperability Of Research Data" principles and guidelines (available at https://www.rd-alliance.org/rda-codata-legal-interoperability-research-data-principles-and-implementation-guidelines-now) that call for the application of public domain waivers to shared data, and particularly the application of the Creative Commons Zero waiver (CC0). Finally, data sharing can, in some cases, have substantial privacy and patient privacy complications. Mechanisms to ensure that data sharing plans consider and address any such concerns in advance of data being made publicly available will be essential to the long-term success of any such program.

Additional Comments

Submission Date

12/13/2016

Submitter Name

Mark Burkard

Name of Organization

University of Wisconsin--Madison

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

Oncology

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

Clinical data. In the era of precision medicine, it is critical that the NIH seek to make de-identified detailed clinical information available and link this to genomics (disease course, comorbid conditions, drug therapies, etc.). The value of sharing this data is provide a large real-world data repository for precision medicine.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

Indefinite. As with RCSB Protein Databank or genomics portal, the data should be continuously available unless deemed obsolete at some later date in time. In the era of expanding storage capacity, short-term resources are more of a barrier than long-term resources.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

Three barriers that can be overcome: (1) IRB/ethics-- standardized written informed consent for data deposit for all prospective and registry studies. (2) Compliance-- require data deposit as a prerequisite for publication. (3) Standardized clinical data format--NIH should adopt or help organize an established format for clinical data.

4. Any other relevant issues respondents recognize as important for NIH to consider

In my opinion the considerations of DOI, RPPR, etc. is less important than the issue of standardizing data, providing a repository and data portal, and requiring data to be deposited at time of publication.

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

The impact of RPPR reporting is expected to be less effective than requiring reporting/depositing data for publication. Also the advantage of depositing in standardized database is data is more accessible than if uploaded in proprietary format as supplemental material.

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

An ID or link to data resource is important. A DOI is no more useful than a central repository linked to the publication.

b. Inclusion of a link to the data/software resource with the citation in the report

Yes the report must provide identifier and/or link as a prerequisite for publication.

c. Identification of the authors of the Data/Software products

Yes, ideally linked to publication as well. If NIH provides funds for software, source code must be freely available.

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

Ini long-standing tradition and using the analogy to RCSB, the publication should be cited, and the data should be accessible from this citation through a common portal in a standardized format. Thus the study alone should be cited, and is sufficient. The methods should detail how the data was accessed (i.e. the identifier and link to data portal).

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

Every resource should have a unique ID linked to publication, similar to RCSB. The publication should provide the IDs of the resources accessible in the portal.

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

Require them for PMCID deposit for studies that meet specific criteria (in case editors do not require it).

4. Any other relevant issues respondents recognize as important for NIH to consider

In my opinion the considerations of DOI, RPPR, etc. is less important than the issue of standardizing data, providing a repository and data portal, and requiring data to be deposited at time of publication.

Additional Comments

Submission Date

12/17/2016

Submitter Name

Daniel S. Katz

Name of Organization

University of Illinois Urbana-Champaign

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

All

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

We recommend that NIH should recognize the critical importance of research software, both as it supports other research products (e.g., manuscripts, data) and as a primary research product, by endorsing and implementing the Software Citation Principles [1]. [1] Smith AM, Katz DS, Niemeyer KE, FORCE11 Software Citation Working Group. (2016) Software citation principles. PeerJ Computer Science 2:e86 DOI: 10.7717/peerj-cs.86

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

We recommend that NIH adopt (current) best practises for handling research software: Use a community platform for hosting/versioning software Archive said software in, e.g., Zenodo, figshare Use metadata that are compatible with and translateable to CodeMeta (https://codemeta.github.io/) metadata files

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

We recommend that NIH consider allocating funds for long-term and sustained effort on these topics. The Software Citation Principles [1] makes specific recommendations which places requirements on sharing and stewardship. Quoting directly from [1]: 4. Persistence: Unique identifiers and metadata describing the software and its disposition should persist—even beyond the lifespan of the software they describe. 5. Accessibility: Software citations should facilitate access to the software itself and to its associated metadata, documentation, data, and other materials necessary for both humans and machines to make informed use of the referenced software. [1] Smith AM, Katz DS, Niemeyer KE, FORCE11 Software Citation Working Group. (2016) Software citation principles. PeerJ Computer Science 2:e86 DOI: 10.7717/peerj-cs.86

4. Any other relevant issues respondents recognize as important for NIH to consider

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

b. Inclusion of a link to the data/software resource with the citation in the report

c. Identification of the authors of the Data/Software products

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and

when might each distinct data set underlying a study be cited and reported separately

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

Data and Software Management Plans must be public and should be machine-readable. see: https://danielskatzblog.wordpress.com/2016/04/13/data-and-software-management-plans-must-be-public-and-should-be-machine-readable/

4. Any other relevant issues respondents recognize as important for NIH to consider

Additional Comments

peerj-cs-86.pdf (2090 KB)

Software citation principles

Arfon M. Smith[1,*], Daniel S. Katz[2,*], Kyle E. Niemeyer[3,*] and FORCE11 Software Citation Working Group

[1] GitHub, Inc., San Francisco, California, United States
[2] National Center for Supercomputing Applications & Electrical and Computer Engineering Department & School of Information Sciences, University of Illinois at Urbana-Champaign, Urbana, Illinois, United States
[3] School of Mechanical, Industrial, and Manufacturing Engineering, Oregon State University, Corvallis, Oregon, United States
* These authors contributed equally to this work.

ABSTRACT

Software is a critical part of modern research and yet there is little support across the scholarly ecosystem for its acknowledgement and citation. Inspired by the activities of the FORCE11 working group focused on data citation, this document summarizes the recommendations of the FORCE11 Software Citation Working Group and its activities between June 2015 and April 2016. Based on a review of existing community practices, the goal of the working group was to produce a consolidated set of citation principles that may encourage broad adoption of a consistent policy for software citation across disciplines and venues. Our work is presented here as a set of software citation principles, a discussion of the motivations for developing the principles, reviews of existing community practice, and a discussion of the requirements these principles would place upon different stakeholders. Working examples and possible technical solutions for how these principles can be implemented will be discussed in a separate paper.

Subjects Digital Libraries, Software Engineering
Keywords Software citation, Software credit, Attribution

SOFTWARE CITATION PRINCIPLES

The main contribution of this document are the software citation principles, written fairly concisely in this section and discussed further later in the document (see Discussion). In addition, we also motivate the creation of these principles (see Motivation), describe the process by which they were created (see Process of Creating Principles), summarize use cases related to software citation (see Use Cases), and review related work (see Related Work). We also lay out the work needed to lead to these software citation principles being applied (see Future Work).

1. Importance: Software should be considered a legitimate and citable product of research. Software citations should be accorded the same importance in the scholarly record as citations of other research products, such as publications and data; they should be included in the metadata of the citing work, for example in the reference list of a journal article, and should not be omitted or separated. Software should be cited on the same basis as any other research product such as a paper or a book, that is, authors should cite the appropriate set of software products just as they cite the appropriate set of papers.

Submitted 24 June 2016
Accepted 23 August 2016
Published 19 September 2016

Corresponding author
Daniel S. Katz, d.katz@ieee.org

Academic editor
Silvio Peroni
DOI 10.7717/peerj-cs.86

Copyright
2016 Smith et al.

Distributed under
Creative Commons CC-BY 4.0

OPEN ACCESS

2. Credit and attribution: Software citations should facilitate giving scholarly credit and normative, legal attribution to all contributors to the software, recognizing that a single style or mechanism of attribution may not be applicable to all software.
3. Unique identification: A software citation should include a method for identification that is machine actionable, globally unique, interoperable, and recognized by at least a community of the corresponding domain experts, and preferably by general public researchers.
4. Persistence: Unique identifiers and metadata describing the software and its disposition should persist even beyond the lifespan of the software they describe.
5. Accessibility: Software citations should facilitate access to the software itself and to its associated metadata, documentation, data, and other materials necessary for both humans and machines to make informed use of the referenced software.
6. Specificity: Software citations should facilitate identification of, and access to, the specific version of software that was used. Software identification should be as specific as necessary, such as using version numbers, revision numbers, or variants such as platforms.

MOTIVATION

As the process of research[1] has become increasingly digital, research outputs and products have grown beyond simply papers and books to include software, data, and other electronic components such as presentation slides, posters, (interactive) graphs, maps, websites (e.g., blogs and forums), and multimedia (e.g., audio and video lectures). Research knowledge is embedded in these components. Papers and books themselves are also becoming increasingly digital, allowing them to become executable and reproducible. As we move towards this future where research is performed in and recorded as a variety of linked digital products, the characteristics and properties that developed for books and papers need to be applied to, and possibly adjusted for, all digital products. Here, we are concerned specifically with the citation of software products. The challenge is not just the textual citation of software in a paper, but the more general identification of software used within the research process. This work focuses on making software a citable entity in the scholarly ecosystem. While software products represent a small fraction of the sum total of research output, this work together with other efforts such as the FORCE11 Data Citation Principles (*Data Citation Synthesis Group, 2014*; *Starr et al., 2015*) collectively represent an effort to better describe (and cite) all outputs of research.

Software and other digital resources currently appear in publications in very inconsistent ways. For example, a random sample of 90 articles in the biology literature found seven different ways that software was mentioned, including simple names in the full-text, URLs in footnotes, and different kinds of mentions in reference lists: project names or websites, user manuals or publications that describe or introduce the software (*Howison & Bullard, 2015*). Table 1 shows examples of these varied forms of software mentions and the frequency with which they were encountered. Many of these kinds of mentions fail to perform the functions needed

[1] We use the term research in this document to include work intended to increase human knowledge and benefit society, in science, engineering, humanities, and other areas.

Table 1 Varieties of software mentions in publications, from *Howison & Bullard (2015)*.

Mention type	Count (n = 286)	Percentage (%)
Cite to publication	105	37
Cite to user s manual	6	2
Cite to name or website	15	5
Instrument-like	53	19
URL in text	13	5
In-text name only	90	31
Not even name	4	1

of citations, and their very diversity and frequent informality undermine the integration of software work into bibliometrics and other analyses. Studies on data and facility citation have shown similar results (*Huang, Rose & Hsu, 2015*; *Mayernik, Maull & Hart, 2015*; *Parsons, Duerr & Minster, 2010*).

There are many reasons why this lack of both software citations in general and standard practices for software citation are of concern:

- Understanding research fields: Software is a product of research, and by not citing it we leave holes in the record of research of progress in those fields.
- Credit: Academic researchers at all levels, including students, postdocs, faculty, and staff, should be credited for the software products they develop and contribute to, particularly when those products enable or further research done by others.[2] Non-academic researchers should also be credited for their software work, though the specific forms of credit are different than for academic researchers.
- Discovering software: Citations enable the specific software used in a research product to be found. Additional researchers can then use the same software for different purposes, leading to credit for those responsible for the software.
- Reproducibility: Citation of specific software used is necessary for reproducibility, although not sufficient. Additional information such as configurations and platform issues are also needed.

[2] Providing recognition of software can have tremendous economic impact as demonstrated by the role of Text REtrieval Conference (TREC) in information retrieval (*Rowe et al., 2010*).

PROCESS OF CREATING PRINCIPLES

The FORCE11 Software Citation Working Group was created in April 2015 with the following mission statement:

> *The software citation working group is a cross-team committee leveraging the perspectives from a variety of existing initiatives working on software citation to produce a consolidated set of citation principles in order to encourage broad adoption of a consistent policy for software citation across disciplines and venues. The working group will review existing efforts and make a set of recommendations. These recommendations will be put off for endorsement by the organizations represented by this group and others that play an important role in the community.*

The group will produce a set of principles, illustrated with working examples, and a plan for dissemination and distribution. This group will not be producing detailed specifications for implementation although it may review and discuss possible technical solutions.

The group gathered members (see Appendix A) in April and May 2015, and then began work in June. This materialized as a number of meetings and offl work by group members to document existing practices in member disciplines; gather materials from workshops and other reports; review those materials, identifying overlaps and differences; create a list of use cases related to software citation, recorded in Appendix B; and subsequently draft an initial version of this document. The draft Software Citation Principles document was discussed in a day-long workshop and presented at the FORCE2016 Conference in April 2016 (https://www.force11.org/meetings/force2016). Members of the workshop and greater FORCE11 community gave feedback, which we recorded here in Appendix C. This discussion led to some changes in the use cases and discussion, although the principles themselves were not modifi We also plan to initiate a follow-on implementation working group that will work with stakeholders to ensure that these principles impact the research process.

The process of creating the software citation principles began by adapting the FORCE11 Data Citation Principles (*Data Citation Synthesis Group, 2014*). These were then modified based on discussions of the FORCE11 Software Citation Working Group (see Appendix A for members), information from the use cases in section Use Cases, and the related work in section Related Work.

We made the adaptations because software, while similar to data in terms of not traditionally having been cited in publications, is also different than data. In the context of research (e.g., in science), the term data usually refers to electronic records of observations made in the course of a research study (raw data) or to information derived from such observations by some form of processing (processed data), as well as the output of simulation or modeling software (simulated data). Some confusion about the distinction between software and data comes in part from the much wider scope of the term data in computing and information science, where it refers to anything that can be processed by a computer. In that sense, software is just a special kind of data. Because of this, citing software is not the same as citing data. A more general discussion about these distinctions is currently underway (https://github.com/danielskatz/software-vs-data).

The principles in this document should guide further development of software citation mechanisms and systems, and the reader should be able to look at any particular example of software citation to see if it meets the principles. While we strive to offer practical guidelines that acknowledge the current incentive system of academic citation, a more modern system of assigning credit is sorely needed. It is not that academic software needs a separate credit system from that of academic papers, but that the need for credit for research software underscores the need to overhaul the system of credit for all research products. One possible solution for a more complete

description of the citations and associated credit is the transitive credit proposed by *Katz (2014)* and *Katz & Smith (2015)*.

USE CASES

We documented and analyzed a set of use cases related to software citation in FORCE11 Software Citation Working Group (https://docs.google.com/document/d/1dS0SqGoBIFwLB5G3HiLLEOSAAgMdo8QPEpjYUaWCvIU) (recorded in Appendix B for completeness). Table 2 summarizes these use cases and makes clear what the requirements are for software citation in each case. Each example represents a particular stakeholder performing an activity related to citing software, with the given metadata as information needed to do that. In that table, we use the following definitions:

- Researcher includes both academic researchers (e.g., postdoc, tenure-track faculty member) and research software engineers.
- Publisher includes both traditional publishers that publish text and/or software papers as well as archives such as Zenodo that directly publish software.
- Funder is a group that funds software or research using software.
- Indexer examples include Scopus, Web of Science, Google Scholar, and Microsoft Academic Search.
- Domain group/library/archive includes the Astronomy Source Code Library (ASCL; http://ascl.net); biomedical and healthCAre Data Discovery Index Ecosystem (bioCADDIE; https://biocaddie.org); Computational Infrastructure for Geodynamics (CIG; https://geodynamics.org), libraries, institutional archives, etc.
- Repository refers to public software repositories such as GitHub, Netlib, Comprehensive R Archive Network (CRAN), and institutional repositories.
- Unique identifier refers to unique, persistent, and machine-actionable identifiers such as a DOI, ARK, or PURL.
- Description refers to some description of the software such as an abstract, README, or other text description.
- Keywords refers to keywords or tags used to categorize the software.
- Reproduce can mean actions focused on reproduction, replication, verification, validation, repeatability, and/or utility.
- Citation manager refers to people and organizations that create scholarly reference management software and websites including Zotero, Mendeley, EndNote, RefWorks, BibDesk, etc., that manage citation information and semi-automatically insert those citations into research products.

All use cases assume the existence of a citable software object, typically created by the authors/developers of the software. Developers can achieve this by, e.g., uploading a software release to figshare (https://figshare.com/) or Zenodo (*GitHub, 2014*) to obtain a DOI. Necessary metadata should then be included in a CITATION file (*Wilson, 2013*) or machine-readable CITATION.jsonld file (*Katz & Smith, 2015*). When software is not

Table 2 Use cases and basic metadata requirements for software citation, adapted from FORCE11 Software Citation Working Group. Solid circles (•) indicate that the use case depends on that metadata, while plus signs (+) indicate that the use case would benefit from that metadata if available.

Use case	Basic requirements											Example stakeholder(s)
	Unique identifier	Software name	Author(s)	Contributor role	Version number	Release date	Location/ repository	Indexed citations	Software license	Description	Keywords	
1. Use software for a paper	•	•	•		•	•	•		+	+		Researcher
in/with new software												software engineer
3. Contribute to software	•	•	•	+	•	•	•		+	+		Researcher, software engineer
4. Determine use/ citations of software	•	•						•				Researcher, software engineer
5. Get credit for software development	•	•	•	+		•	•	+				Researcher, software engineer
6. Reproduce analysis	•	•			•	•	•		+	+		Researcher
7. Find software to implement task	•	•	•				•	•	+	+	+	Researcher, software engineer
8. Publish software paper	•	•	•		•	•	•					Publisher
9. Publish papers that cite software	•	•	•		•	•	•	•				Publisher
10. Build catalog of software	•	•	•		•	•	•	•	+	+	+	Indexer
11. Build software catalog/registry	•	•	•				•			+	+	Domain group, library, archive
12. Show scientific impact of holdings	•	•						•				Repository
13. Show how funded software has been used	•	•						•				Funder, policy maker
14. Evaluate contributions of researcher	•		•	+		•		•				Evaluator, funder
15. Store software entry	•	•	•		•	•	•			+	+	Citation manager
16. Publish mixed data/software packages	•	•	•		•	•	•		+	+	+	Repository, library, archive

freely available (e.g., commercial software) or when there is no clear identifier to use, alternative means may be used to create citable objects as discussed in section Access to Software.

In some cases, if particular metadata are not available, alternatives may be provided. For example, if the version number and release date are not available, the download date can be used. Similarly, the contact name/email is an alternative to the location/repository.

RELATED WORK

With approximately 50 working group participants (see Appendix A) representing a range of research domains, the working group was tasked to document existing practices in their respective communities. A total of 47 documents were submitted by working group participants, with the life sciences, astrophysics, and geosciences being particularly well-represented in the submitted resources.

General community/non domain-specific activities

Some of the most actionable work has come from the UK Software Sustainability Institute (SSI) in the form of blog posts written by their community fellows. For example, in a blog post from 2012, *Jackson (2012)* discusses some of the pitfalls of trying to cite software in publications. He includes useful guidance for when to consider citing software as well as some ways to help convince journal editors to allow the inclusion of software citations.

Wilson (2013) suggests that software authors include a CITATION file that documents exactly how the authors of the software would like to be cited by others. While this is not a formal metadata specification (e.g., it is not machine readable) this does offer a solution for authors wishing to give explicit instructions to potential citing authors and, as noted in the motivation section (see Motivation), there is evidence that authors follow instructions if they exist (*Huang, Rose & Hsu, 2015*).

In a later post on the SSI blog, Jackson gives a good overview of some of the approaches package authors have taken to automate the generation of citation entities such as BibTeX entries (*Jackson, 2014*), and *Knepley et al. (2013)* do similarly.

While not usually expressed as software citation principles, a number of groups have developed community guidelines around software and data citation. *Van de Sompel et al. (2004)* argue for registration of all units of scholarly communication, including software. In Publish or be damned? An alternative impact manifesto for research software, *Chue Hong (2011)* lists nine principles as part of The Research Software Impact Manifesto. In the Science Code Manifesto (*Barnes et al., 2016*), the founding signatories cite five core principles (Code, Copyright, Citation, Credit, Curation) for scientific software.

Perhaps in light of the broad range of research domains struggling with the challenge of better recognizing the role of software, funders and agencies in both the US (e.g., NSF, NIH, Alfred P. Sloan Foundation) and UK (e.g., SFTC, JISC, Wellcome Trust) have sponsored or hosted a number of workshops with participants from across a range of

disciplines, specifically aimed at discussing issues around software citation (*Sufi et al., 2014*; *Ahalt et al., 2015*; *Software Credit Workshop, 2015*; *Norén, 2015*; *Software Attribution for Geoscience Applications, 2015*; *Allen et al., 2015*). In many cases these workshops produced strong recommendations for their respective communities on how best to proceed. In addition, a number of common themes arose in these workshops, including (1) the critical need for making software more citable (and therefore actions authors and publishers should take to improve the status quo), (2) how to better measure the impact of software (and therefore attract appropriate funding), and (3) how to properly archive software (where, how, and how often) and how this affects what to cite and when.

Most notable of the community efforts are those of WSSSPE Workshops (http://wssspe.researchcomputing.org.uk/) and SSI Workshops (http://www.software.ac.uk/community/workshops), who between them have run a series of workshops aimed at gathering together community members with an interest in (1) defining the set of problems related to the role of software and associated people in research settings, particularly academia, (2) discussing potential solutions to those problems, (3) beginning to work on implementing some of those solutions. In each of the three years that WSSSPE workshops have run thus far, the participants have produced a report (*Katz et al., 2014*; *Katz et al., 2016a*; *Katz et al., 2016b*) documenting the topics covered. Section 5.8 and Appendix J in the WSSSPE3 report (*Katz et al., 2016b*) has some preliminary work and discussion particularly relevant to this working group. In addition, a number of academic publishers such as APA (*McAdoo, 2015*) have recommendations for submitting authors on how to cite software, and journals such as ***F1000Research*** (http://f1000research.com/for-authors/article-guidelines/software-tool-articles), ***SoftwareX*** (http://www.journals.elsevier.com/softwarex/), ***Open Research Computation*** (http://www.openresearchcomputation.com) and the ***Journal of Open Research Software*** (http://openresearchsoftware.metajnl.com) allow for submissions entirely focused on research software.

Domain-specific community activities

One approach to increasing software citability is to encourage the submission of papers in standard journals describing a piece of research software, often known as software papers (see Software Papers). While some journals (e.g., Transactions on Mathematical Software (TOMS), Bioinformatics, Computer Physics Communications, F1000Research, Seismological Research Letters, Electronic Seismologist) have traditionally accepted software submissions, the American Astronomical Society (AAS) has recently announced they will accept software papers in their journals (*AAS Editorial Board, 2016*). Professional societies are in a good position to change their respective communities, as the publishers of journals and conveners of domain-specific conferences; as publishers they can change editorial policies (as AAS has done) and conferences are an opportunity to communicate and discuss these changes with Astrophysics Source Code Library their communities.

In astronomy and astrophysics: The Astrophysics Source Code Library (ASCL; http://ASCL.net) is a website dedicated to the curation and indexing of software used in the astronomy-based literature. In 2015, the AAS and GitHub co-hosted a workshop

(*Norén, 2015*) dedicated to software citation, indexing, and discoverability in astrophysics. More recently, a Birds of a Feather session was held at the Astronomical Data Analysis Software and Systems (ADASS) XXV conference (*Allen et al., 2015*) that included discussion of software citation.

In the life sciences: In May 2014, the NIH held a workshop aimed at helping the biomedical community discover, cite, and reuse software written by their peers. The primary outcome of this workshop was the Software Discovery Index Meeting Report (*White et al., 2014*) which was shared with the community for public comment and feedback. The authors of the report discuss what framework would be required for supporting a Software Discovery Index including the need for unique identifiers, how citations to these would be handled by publishers, and the critical need for metadata to describe software packages.

In the geosciences: The Ontosoft (*Gil, Ratnakar & Garijo, 2015*) project describes itself as A Community Software Commons for the Geosciences. Much attention was given to the metadata required to describe, discover, and execute research software. The NSF-sponsored Geo-Data Workshop 2011 (*Fox & Signell, 2011*) revolved around data lifecycle, management, and citation. The workshop report includes many recommendations for data citation.

Existing efforts around metadata standards

Producing detailed specifications and recommendations for possible metadata standards to support software citation was not within the scope of this working group. However some discussion on the topic did occur and there was significant interest in the wider community to produce standards for describing research software metadata.

Content specifications for software metadata vary across communities, and include DOAP (https://github.com/edumbill/doap/), an early metadata term set used by the Open Source Community, as well as more recent community efforts like Research Objects (*Bechhofer et al., 2013*), The Software Ontology (*Malone et al., 2014*), EDAM Ontology (*Ison et al., 2013*), Project CRediT (*CRediT, 2016*), the OpenRIF Contribution Role Ontology (*Gutzman et al., 2016*), Ontosoft (*Gil, Ratnakar & Garijo, 2015*), RRR/JISC guidelines (*Gent, Jones & Matthews, 2015*), or the terms and classes defined at schema.org related to the https://schema.org/SoftwareApplication class. In addition, language-specific software metadata schemes are in widespread use, including the Debian package format (*Jackson & Schwarz, 2016*), Python package descriptions (*Ward & Baxter, 2016*), and R package descriptions (*Wickham, 2015*), but these are typically conceived for software build, packaging, and distribution rather than citation. CodeMeta (*Jones et al., 2014*) has created a crosswalk among these software metadata schemes and an exchange format that allows software repositories to effectively interoperate.

DISCUSSION

In this section we discuss some the issues and concerns related to the principles stated in section Software Citation Principles.

What software to cite

The software citation principles do not define what software should be cited, but rather how software should be cited. What software should be cited is the decision of the author(s) of the research work in the context of community norms and practices, and in most research communities, these are currently in flux. In general, ***we believe that software should be cited on the same basis as any other research product such as a paper or book***; that is, authors should cite the appropriate set of software products just as they cite the appropriate set of papers, perhaps following the FORCE11 Data Citation Working Group principles, which state, In scholarly literature, whenever and wherever a claim relies upon data, the corresponding data should be cited (*Data Citation Synthesis Group, 2014*).

Some software which is, or could be, captured as part of data provenance may not be cited. Citation is partly a record of software important to a research outcome[3], where provenance is a record of all steps (including software) used to generated particular data within the research process. Research results, including data, increasingly depend on software (*Hannay et al., 2009*), and thus may depend on the specific version used (*Sandve et al., 2013*; *Wilson et al., 2014*). Furthermore, errors in software or environment variations can affect results (*Morin et al., 2012*; *Soergel, 2015*). This implies that for a data research product, provenance data will include some of the cited software. Similarly, the software metadata recorded as part of data provenance will overlap the metadata recorded as part of software citation for the software that was used in the work. The data recorded for reproducibility should also overlap the metadata recorded as part of software citation. In general, we intend the software citation principles to cover the minimum of what is necessary for software citation for the purpose of software identification. Some use cases related to citation (e.g., provenance, reproducibility) might have additional requirements beyond the basic metadata needed for citation, as Table 2 shows.

[3] Citation can be used for many purposes, including for software: which software has been used in the work, which software has influenced the work, which software is the work superseding, which software is the work disproving, etc.

Software papers

Currently, and for the foreseeable future, software papers are being published and cited, in addition to software itself being published and cited, as many community norms and practices are oriented towards citation of papers. As discussed in the Importance principle (1) and the discussion above, ***the software itself should be cited on the same basis as any other research product; authors should cite the appropriate set of software products.*** If a software paper exists and it contains results (performance, validation, etc.) that are important to the work, then the software paper should also be cited. We believe that a request from the software authors to cite a paper should typically be respected, and the paper cited ***in addition to*** the software.

Derived software

The goals of software citation include the linked ideas of crediting those responsible for software and understanding the dependencies of research products on specific software. In the Importance principle (1), we state that software should be cited on the same basis as any other research product such as a paper or a book; that is, authors should cite

the appropriate set of software products just as they cite the appropriate set of papers. In the case of one code that is derived from another code, citing the derived software may appear to not credit those responsible for the original software, nor recognize its role in the work that used the derived software. However, this is really analogous to how any research builds on other research, where each research product just cites those products that it directly builds on, not those that it indirectly builds on. Understanding these chains of knowledge and credit have been part of the history of science field for some time, though more recent work suggests more nuanced evaluation of the credit chains (*CRediT, 2016*; *Katz & Smith, 2015*).

Software peer review

Adherence to the software citation principles enables better peer review through improved reproducibility. However, since the primary goal of software citation is to identify the software that has been used in a scholarly product, the peer review of software itself is mostly out of scope in the context of software citation principles. For instance, when identifying a particular software artifact that has been used in a scholarly product, whether or not that software has been peer-reviewed is irrelevant. One possible exception would be if the peer-review status of the software should be part of the metadata, but the working group does not believe this to be part of the minimal metadata needed to identify the software.

Citation format in reference list

Citations in references in the scholarly literature are formatted according to the citation style (e.g., AMS, APA, Chicago, MLA) used by that publication. (Examples illustrating these styles have been published by *Lipson (2011)*; the follow-on Software Citation Implementation Group will provide suggested examples.) As these citations are typically sent to publishers as text formatted in that citation style, not as structured metadata, and because the citation style dictates how the human reader sees the software citation, *we recommend that all text citation styles support the following: a) a label indicating that this is software, e.g., [Software], potentially with more information such as [Software: Source Code], [Software: Executable], or [Software: Container], and b) support for version information, e.g., Version 1.8.7.*

Citations limits

This set of software citation principles, if followed, will cause the number of software citations in scholarly products to increase, thus causing the number of overall citations to increase. Some scholarly products, such as journal articles, may have strict limits on the number of citations they permit, or page limits that include reference sections. Such limits are counter to our recommendation, and *we recommend that publishers using strict limits for the number of citations add specifi instructions regarding software citations to their author guidelines to not disincentivize software citation.* Similarly, *publishers should not include references in the content counted against page limits.*

Unique identification

The Unique Identification principle (3) calls for a method for identification that is machine actionable, globally unique, interoperable, and recognized by a community. What this means for data is discussed in detail in the Unique Identification section of a report by the FORCE11 Data Citation Implementation Group (*Starr et al., 2015*), which calls for unique identification in a manner that is machine-resolvable on the Web and demonstrates a long-term commitment to persistence. This report also lists examples of identifiers that match these criteria including DOIs, PURLs, Handles, ARKS, and NBNs. For software, ***we recommend the use of DOIs as the unique identifier due to their common usage and acceptance, particularly as they are the standard for other digital products such as publications.***

While we believe there is value in including the explicit version (e.g., Git SHA1 hash, Subversion revision number) of the software in any software citation, there are a number of reasons that a commit reference together with a repository URL is not recommended for the purposes of software citation:

1. Version numbers/commit references are not guaranteed to be permanent. Projects can be migrated to new version control systems (e.g., SVN to Git). In addition, it is possible to overwrite/clobber a particular version (e.g., force-pushing in the case of Git).
2. A repository address and version number does not guarantee that the software is available at a particular (resolvable) URL, especially as it is possible for authors to remove their content from, e.g., GitHub.
3. A particular version number/commit reference may not represent a preferred point at which to cite the software from the perspective of the package authors.

We recognize that there are certain situations where it may not be possible to follow the recommended best-practice. For example, if (1) the software authors did not register a DOI and/or release a specific version, or (2) the version of the software used does not match what is available to cite. In those cases, falling back on a combination of the repository URL and version number/commit hash would be an appropriate way to cite the software used.

Note that the unique in a UID means that it points to a unique, specific software version. However, multiple UIDs might point to the same software. This is not recommended, but is possible. ***We strongly recommend that if there is already a UID for a version of software, no additional UID should be created.*** Multiple UIDs can lead to split credit, which goes against the Credit and Attribution principle (2).

Software versions and identifiers

There are at least three different potential relationships between identifiers and versions of software:

1. An identifier can point to a specific version of a piece of software.
2. An identifier can point to the piece of software, effectively all versions of the software.
3. An identifier can point to the latest version of a piece of software.

It is possible that a given piece of software may have identifiers of all three types. In addition, there may be one or more software papers, each with an identifier.

While we often need to cite a specific version of software, we may also need a way to cite the software in general and to link multiple releases together, perhaps for the purpose of understanding citations to the software. The principles in section Software Citation Principles are intended to be applicable at all levels, and to all types of identifiers, such as DOIs, RRIDs, etc., though we again recommend when possible the use of DOIs that identify specific versions of source code. We note that RRIDs were developed by the FORCE11 Resource Identification Initiative (https://www.force11.org/group/resource-identification-initiative) and have been discussed for use to identify software packages (not specific versions), though the FORCE11 Resource Identification Technical Specifications Working Group (https://www.force11.org/group/resource-identification-technical-specifications-working-group) says Information resources like software are better suited to the Software Citation WG. There is currently a lack of consensus on the use of RRIDs for software.

Types of software

The principles and discussion in this document have generally been written to focus on software as source code. However, we recognize that some software is only available as an executable, a container, or a virtual machine image, while other software may be available as a service. We believe the principles apply to all of these forms of software, though the implementation of them will certainly differ based on software type. ***When software is accessible as both source code and another type, we recommend that the source code be cited.***

Access to software

The Accessibility principle (5) states that software citations should permit and facilitate access to the software itself. This does not mean that the software must be freely available. Rather, the metadata should provide enough information that the software can be accessed. If the software is free, the metadata will likely provide an identifier that can be resolved to a URL pointing to the specific version of the software being cited. For commercial software, the metadata should still provide information on how to access the specific software, but this may be a company s product number or a link to a website that allows the software be purchased. As stated in the Persistence principle (4), we recognize that the software version may no longer be available, but it still should be cited along with information about how it was accessed.

What an identifier should resolve to

While citing an identifier that points to, e.g., a GitHub repository can satisfy the principles of Unique Identification (3), Accessibility (5), and Specificity (6), such a repository cannot guarantee Persistence (4). ***Therefore, we recommend that the software identifier should resolve to a persistent landing page that contains metadata and a link to the software itself, rather than directly to the source code files, repository, or executable.*** This ensures

longevity of the software metadata even perhaps beyond the lifespan of the software they describe. This is currently offered by services such as figshare and Zenodo (*GitHub, 2014*), which both generate persistent DataCite DOIs for submitted software. In addition, such landing pages can contain both human-readable metadata (e.g., the types shown by Table 2) as well as content-negotiable formats such as RDF or DOAP (https://github.com/edumbill/doap/).

Updates to these principles

As this set of software citation principles has been created by the FORCE11 Software Citation Working Group (https://www.force11.org/group/software-citation-working-group), which will cease work and dissolve after publication of these principles, any updates will require a different FORCE11 working group to make them. As mentioned in section Future Work, we expect a follow-on working group to be established to promote the implementation of these principles, and it is possible that this group might find items that need correction or addition in these principles. ***We recommend that this Software Citation Implementation Working Group be charged, in part, with updating these principles during its lifetime, and that FORCE11 should listen to community requests for later updates and respond by creating a new working group.***

FUTURE WORK

Software citation principles without clear worked-through examples are of limited value to potential implementers, and so in addition to this principles document, the final deliverable of this working group will be an implementation paper outlining working examples for each of the use cases listed in section Use Cases.

Following these efforts, we expect that FORCE11 will start a new working group with the goals of supporting potential implementers of the software citation principles and concurrently developing potential metadata standards, loosely following the model of the FORCE11 Data Citation Working Group. Beyond the efforts of this new working group, additional effort should be focused on updating the overall academic credit/citation system.

APPENDIX A

Working group membership

Alberto Accomazzi, Harvard-Smithsonian CfA

Alice Allen, Astrophysics Source Code Library

Micah Altman, MIT

Jay Jay Billings, Oak Ridge National Laboratory

Carl Boettiger, University of California, Berkeley

Jed Brown, University of Colorado Boulder

Sou-Cheng T. Choi, NORC at the University of Chicago & Illinois Institute of Technology

Neil Chue Hong, Software Sustainability Institute

Tom Crick, Cardiff Metropolitan University
Mercè Crosas, IQSS, Harvard University
Scott Edmunds, GigaScience, BGI Hong Kong
Christopher Erdmann, Harvard-Smithsonian CfA
Martin Fenner, DataCite
Darel Finkbeiner, OSTI
Ian Gent, University of St Andrews, recomputation.org
Carole Goble, The University of Manchester, Software Sustainability Institute
Paul Groth, Elsevier Labs
Melissa Haendel, Oregon Health and Science University
Stephanie Hagstrom, FORCE11
Robert Hanisch, National Institute of Standards and Technology, One Degree Imager
Edwin Henneken, Harvard-Smithsonian CfA
Ivan Herman, World Wide Web Consortium (W3C)
James Howison, University of Texas
Lorraine Hwang, University of California, Davis
Thomas Ingraham, F1000Research
Matthew B. Jones, NCEAS, University of California, Santa Barbara
Catherine Jones, Science and Technology Facilities Council
Daniel S. Katz, University of Illinois (co-chair)
Alexander Konovalov, University of St Andrews
John Kratz, California Digital Library
Jennifer Lin, Public Library of Science
Frank Löffler, Louisiana State University
Brian Matthews, Science and Technology Facilities Council
Abigail Cabunoc Mayes, Mozilla Science Lab
Daniel Mietchen, National Institutes of Health
Bill Mills, TRIUMF
Evan Misshula, CUNY Graduate Center
August Muench, American Astronomical Society
Fiona Murphy, Independent Researcher
Lars Holm Nielsen, CERN
Kyle E. Niemeyer, Oregon State University (co-chair)
Karthik Ram, University of California, Berkeley
Fernando Rios, Johns Hopkins University
Ashley Sands, University of California, Los Angeles
Soren Scott, Independent Researcher

Frank J. Seinstra, Netherlands eScience Center
Arfon Smith, GitHub (co-chair)
Kaitlin Thaney, Mozilla Science Lab
Ilian Todorov, Science and Technology Facilities Council
Matt Turk, University of Illinois
Miguel de Val-Borro, Princeton University
Daan Van Hauwermeiren, Ghent University
Stijn Van Hoey, Ghent University
Belinda Weaver, The University of Queensland
Nic Weber, University of Washington iSchool

APPENDIX B

Software citation use cases

This appendix records an edited, extended description of the use cases discussed in section Use Cases, originally found in FORCE11 Software Citation Working Group. This discussion is not fully complete, and in some cases, it may not be fully self-consistent, but it is part of this paper as a record of one of the inputs to the principles. We expect that the follow-on Software Citation Implementation Group will further develop these use cases, including explaining in more detail how the software citation principles can be applied to each as part of working with the stakeholders to persuade them to actually implement the principles in their standard workflows.

Researcher who uses someone else's software for a paper

One of the most common use cases may be researchers who use someone else s software and want to cite it in a technical paper. This will be similar to existing practices for citing research artifacts in papers.

Requirements for researcher:

- Name of software
- Names of software authors/contributors
- Software version number and release date, or download date
- Location/repository, or contact name/email (if not publicly available)
- Citable DOI of software
- Format for citing software in text and in bibliography

Possible steps:

1. Software developers create CITATION file and associate with source code release/ repository.
2. Researcher finds and uses software for research paper.
3. Researcher identifies citation metadata file (e.g., CITATION file) associated with downloaded/installed software source code or in online repository/published location.

CITATION file includes necessary citation metadata. CITATION file may include BibTeX entry, suggested citation format.

4. Researcher cites software appropriately, e.g., in methodology section; reference included in bibliography.

Researcher who uses someone else's software for new software

In this case, a researcher develops new software that incorporates or depends on existing software. In order to credit the developer(s), the researcher will include citations in his/her source code, documentation, or other metadata in a similar manner to papers.

Requirements for researcher:

- Name of software
- Names of software authors/contributors
- Software version number and release date
- Location/repository
- Citable DOI of software
- Format for citing software in source code, documentation, or citation metadata file

Possible steps:

1. Assume that software developers have created a CITATION file and associated with the source code release/repository.
2. Researcher finds and uses software in the development of new software.
3. Researcher identifies citation metadata file (e.g., CITATION file) associated with downloaded/installed software source code or in online repository/published location. CITATION file includes necessary citation metadata. CITATION file may include BibTeX entry, suggested citation format.
4. Researcher cites software in source code, documentation, or other metadata-containing file.

Researcher who contributes to someone else's software (open source project)

A researcher wants to contribute to someone else s software in the manner in which their contributions will be accepted and recognized.

Possible steps:

1. Researcher finds information about the software, and how contributors will be recognized
2. Researcher possibly submit a Contributor License Agreement (CLA) or Copyright Assignment Agreement (CAA) to allow the contributed content to be distributed with the software being contributed to
3. Researcher contributes to the software
4. Software maintainers accept contribution, recognize researcher s contribution, and update the software metadata as appropriate

Researcher who wants to know who uses the researcher's software

This case is similar to a researcher who wants to fi other papers/publications that cite a particular paper. A researcher wants to gauge the usage of her software within or across communities and measure its impact on research for both credit and funding.

Requirements:

- Uniquely identify software
- Indexed citations of software
- Indexed papers that use software

Steps:

1. Researcher finds software official name or unique DOI in metadata associated with downloaded/installed source code or in online repository/published location.
2. Researcher searches for software, may use online indexer (e.g., Scopus, Web of Science, Google Scholar) using software name or DOI.
3. Online indexer presents entry for software with list of citations, if any. Ideally, entry will also include metadata contained in software CITATION file and citation example.

Researcher gets credit for software development at the academic/governmental institution, in professional career, etc

This case describes the need for a researcher who has contributed to software (by design, software engineering, development, testing, patching, documentation, training, evangelizing, etc.) to have their software work recognized by their employer or colleagues for the purpose of career advancement and increased professional reputation.

Requirements for researcher:

- Name of software
- Names of software authors/contributors
- Location/repository
- Citable DOI of software
- Format for citing software in an official CV, in a departmental/institutional review report, etc.
- Role in the software creation, that is linked to version or component
- Role in contributing to the software as a package (not just lines of code) development of benchmarks, testing, documentation, tutorials etc.

Researcher who wants to "reproduce" another person/group's analysis

When a researcher wants to understand or verify a research results from another researcher, they would like to use the same software. Note that accessing the exact same software is necessary but not sufficient for reproducibility.

Requirements for researcher:

- Name of software
- Location/repository for the exact release that was used
- DOI or other persistent handle for that specific release
- Release has all components necessary for reproducing the work (Note: this ideally also means sample inputs and outputs)

Researcher who wants to find a piece of software to implement a task

This is the case where a research is looking for software to use but wants to understand whether it is being used in a scholarly fashion. For example, a researcher searches through a software repository and finds a package that might be useful. They look to find whether it has been used by others in the scientific literature.

Requirements:

- Either the software documentation page has a reference to existing literature that makes use of it.
- There is a mechanism to look it up.

Publisher wants to publish a software paper

This case asks what information regarding software is needed for a publisher who wants to publish a paper describing that software.

Requirements:

- Name of software
- Names of software authors/contributors
- Location/repository
- Citable DOI of software
- Format for citing software in JATS, for example, as well as references in the text itself

Publisher who wants to publish papers that cite software

This case asks what information regarding software is needed for a publisher who wants to publish papers that cite that software.

Requirements for publisher:

- Name of software
- Names of software authors/contributors
- Location/repository
- Citable DOI of software
- Format for citing software in, e.g., JATS, as well as references in the text itself

Indexer (e.g., Scopus, WoS, Scholar, MS Academic Search) who wants to build a catalog of software

Provide an index over the software that is used within the research domain. Track how that software is being used by different groups of researchers and to what ends.

Requirements:

- Uniquely identify pieces of software used by the research literature
- Connect authors and organizations to that software
- Connect various software versions together

Domain group (e.g., ASCL, bioCADDIE), Libraries, and Archives (e.g., University library, laboratory archive, etc.) wants to build a catalog/registry of institutional or domain software

There are two different examples here: One is building a catalog/archive of software produced by those affiliated with the institution. The other is along the lines of Sayeed Choudhury s note that data are the new special collections. An institution may choose to build a catalog/archive of many things within a single topic or subject in order to secure all the software on a certain topic or build a collection that may draw users to their establishment, much like special collections now do for university libraries and archives.

Repository showing scientific impact of holdings

A repository that archives and/or maintains a collection of software. The repository would like to address usage and impact of software in its holding. Usage would aid potential users whether the software is being actively maintained or developed or has been superseded. Both would help repository know how to direct resources, e.g., maintenance, training etc. This is similar to the case of a funder wanting to know the impact of funded work.

Requirements:

- Code name, or a unique identifier
- Relationships to previous versions
- Connect to repository
- Connect to research

Funder who wants to know how software they funded has been used

This use case is similar to Repository showing scientifi impact of holdings , where a funder wants to find out the use and impact and software that they supported. It is also similar to Researcher who wants to know who uses the researcher s software.

Evaluator or funder wants to evaluate contributions of a researcher

In this use case, an evaluator (e.g., academic administrator) or funder wants to evaluate the contributions of a researcher who develops software. This case is related to those

where researchers want to get credit for software development, or where organizations want to evaluate the impact of software itself.

Reference management system used by researchers to author a manuscript

Reference management systems may need to be updated to internally understand that their is a software reference type, and to be able to output references to software in common formats.

Requirements for reference manager:

- Names of software authors/contributors
- Software version number and release date
- Location/repository
- Citable DOI of software or paper recommended for citation
- Format for citing software in citation metadata file
- Citation metadata tags embedded in DOI landing page/software project page for easy ingest

Possible steps:

1. Reference management system such as EndNote, Mendeley, Zotero, etc. builds affordances for software references.
2. Researcher finds software citation and adds it to their reference manager library, by (a) importing from the CITATION file (e.g., BibTeX, RIS), or (b) clicking on, e.g., an add to Zotero library widget in web browser.
3. Researcher writes a paper and uses the reference manager to generate citations or bibliography.

Repository wants to publish mixed data/software packages

Domain and institutional data repositories have both data and software artifacts, and want to link these together in a provenance trace that can be cited. Sometimes the software is a separately identified artifact, but at other times software is included inside of data packages, and the researcher wants to cite the combined product.

Use cases not adopted in the table

Researcher who benchmarks someone else's software with or without modification on one or many hardware platforms for publication

This case describes the need for a researcher who has contributed to software (by design, software engineering, development, testing, patching, documentation, training, evangelizing, etc.) to have their software work recognized by their employer or colleagues for the purpose of career advancement and increased professional reputation.

Requirements for researcher:

- Name of software
- Names of software authors/contributors
- Software version number and release date
- Location/repository
- Citable DOI of software or paper recommended for citation
- Format for citing software in source code or citation metadata file

Possible steps:

1. Software developers create CITATION file and associate with source code release/repository.
2. Researcher finds and uses software in the development of new software.
3. Researcher identifies citation metadata file (e.g., CITATION file) associated with downloaded/installed software source code or in online repository/published location. CITATION file includes necessary citation metadata. CITATION file may include BibTeX entry, suggested citation format.
4. Researcher cites software in source code, documentation, or other metadata-containing file.

After review of this use case, we decided that based on the title this falls under use case 1, where a researcher uses someone else s software for a paper. Unlike use case 1, which is general in terms of the use of software, here the use leads to a benchmarking study but the outcome in both cases is a paper that needs to cite the software.

Researcher who wants to publish about a piece of software

The research wants to publish about a version of software they have produced. A key part of this use case is to be able to connect the given narrative to a specific version of the software in questions and connect that in large story.

Requirements:

- Name of software
- Names of software authors/contributors
- Location/repository
- Citable DOI of Software
- Links to older versions of software

This is similar to use case 1, other than the fact that the software developer(s) and paper author(s) will likely be the same person/people here.

Researcher wants to record the software that generated some data

This is the case where a researcher is using some software to perform an analysis, either of a physical sample or of data. The researcher needs to know which version was used, for

example in case a bug was fixed. Note that knowing the software and its version is not sufficient to determine the conditions of the analysis, but they are essential.

Requirement: The analysis, or the generated data, has information about the software used.

This is also similar to use case 1, except in that case the research output is a paper, while here the output is a dataset.

Researcher who wants to reproduce experience of use of a particular software implementation in context

Researcher is engaged in historical/cultural research, e.g., a study of video games as cultural artifacts.

Requirements:

- Name of software
- Software version number
- Documentation of the execution environment/context
- Location/repository for virtual machine (or equivalent) comprising both software and execution environment/context
- Persistent identifier associated with virtual machine instance (or equivalent) comprising both software and execution environment/context

Possible steps:

1. Researcher obtains persistent ID from citation
2. Research uses a persistent ID resolution service to resolve ID to a location of an executable VM instance in a repository
3. Researcher obtains VM in the repository, executes it, and interacts with software

This overlaps use case 6 (reproducing analysis), and so we decided not to include this as a distinct use case.

APPENDIX C

Feedback following FORCE2016

This appendix contains a record of comments made by the FORCE11 community on the draft Software Citation Principles, either directly via Hypothesis on the draft document (https://www.force11.org/softwarecitation-principles) posted following the FORCE2016 conference (https://www.force11.org/meetings/force2016) or via GitHub issues (https://github.com/force11/force11-scwg/issues), and the responses to these comments.

On unique identification

I know this suggestion of a single unique identifier comes from the DOI perspective where it works pretty well, but I m wondering if something different in the way of identification should be used for software. For creative works generally there is the FRBR model

(https://en.wikipedia.org/wiki/Functional_Requirements_for_Bibliographic_Records) which defines several levels for a creative entity work, expression, manifestation, and item. I think something along these lines are particularly relevant for software it is useful to be able to locate all uses of a particular piece of software no matter what version (the work level software identified by a particular name and purpose over a period of time), but it is also important to specify the particular version used in any given work (expression the source code at the time of use) and in some cases also the platform (manifestation the compiled bytes including libraries, for example a docker image). Item probably isn t relevant for software. That is, I think a software citation perhaps could use THREE distinct unique identifiers, one for the work itself, one for the specific version (source code), and possibly an additional one for the actual downloadable binary image that can be run. Rather than leave it implicit I think recognizing the different levels of citable record would be helpful here. #F11SC

Reply: I interpret the requirement for global uniqueness as referring to the identifier itself. Two different people can have the same name (not globally unique) but cannot share a single ORCID (globally unique). Global uniqueness of the identifier does not preclude multiple identifiers pointing to the same person. I think the suggestion of differentiating between different software expressions/manifestations/items is a reasonable one, but I don t think it relaxes the requirement for identifiers to be globally unique.

Our response: We agree that there are valid points here, but on balance we don t feel that the rewards from implementing this outweigh the practical challenges.

On accessibility

Should this document address this in further detail? For example, permit and facilitate access could be explored further. Should this be done through open access licensing? repositories? Who s responsible for providing this access?

I am also wondering if this is a separate issue since citing traditionally pointed to publications but did not necessarily address access. DOI, for example is stated, but doesn t guarantee access, so does this simply restating point 3, or should it provide something new?

Our response: We agree that accessibility should receive further attention, which the follow-on group focusing on implementation will provide. However, this is out of scope for the document outlining the principles.

To the second point, accessibility provides information about access, but does not guarantee access itself (e.g., paywalled article).

On specificity

I am wondering if this should be folded into number 3 Unique Identification. Both seem to deal with the issue of identification and access.

Our response: A unique software identifier can point to the specific version/variant of software, but it can also identify other things (collection of versions, repository, etc.),

while this principle deals with the need to identify the specific version of software used (via citation).

On academic credit

A lot of software that were developed by non-academic engineers also contribute to academic research indirectly. Their names and contributions should also be credited. So removing Academic makes more sense?

Reply: This is a good point, though I think academic and non-academic credit are different, so perhaps we can add to this regarding non-academic credit, rather than removing academic.

Reply: I agree with Daniel on this. Keep Academic and add non-academic.

Our response: We ve made the bullet more general, just about credit, discussing academic credit and adding a sentence about non-academic credit as well.

On citations in text

Although the focus here is on citations in the references, as a publisher, our experience is that most common practice of citation of data and software for authors is typically in the main body of the text. In order to encourage software to be treated and valued as a fi research object, it is important that citations to it be positioned in the references as citations to articles and books are. However, it would be a missed opportunity if we did not leverage current practices of authors. This will also likely arise during implementation, as it has for the Data Citation Implementation Publisher Early Adopters Pilot. This could be addressed in future work on implementation.

Our response: In the principles, we propose that software should be cited in the references list, to recognize the primary role of software in research. However, this practice is not mutually exclusive with also referencing/citing software in the main body of a paper as long as the software is cited in the references.

On unique identification

Clearer instructions will be needed for authors on which version to cite. For BioMed Central journals, we ask authors to cite two versions of the software, an archived version (e.g., on Zenodo) as well as the current version (e.g., on GitHub). This is to ensure accessibility. However, if repositories and archives were to include a persistent link to the current version of the software, publishers could then instruct authors to cite only software with a UID, which wouldn t point to a current version, but would point to the version(s) used and would be a more accurate version of scientific record. Related to this point is the idea of group object identifiers. A need for group identifiers has been identified in the area of data (e.g., in the case of meta-analyses), and one could also identify a use case for these in the case of software, collecting metadata around all versions of a given software package. See blog here (https://blog.datacite.org/to-better-understand-research-communication-we-need-a-groid-group-object-identifier/).

Our response: We recommend citing the specific version of the software that was used. We expect that the unique identifier (e.g., DOI) will point to a landing page that directs to the repository/current version. However, this is more of a convenience issue that the software developers should address, rather than the author citing the software they used.

On future work

For implementation we would recommend both consulting with adopters as well as developing metadata standards simultaneously rather than developing metadata standards and then pursuing early adopters implementation. The work early adopters are doing now for data citation will be able to be leveraged for software citation and the changes needed to do so could happen now. There is no need to wait on approval of new tagging for a specific metadata standard. Many publishers will have their own preferred metadata standards and so implementation could begin now with publishers, as long as we know what we want to capture. Future implementation groups might also consider levels of contribution. This is particularly relevant for software. Who is considered an author? For example, to what extent should authors of pull requests receive attribution? This might be considered in an FAQs group, or possibly an early adopters group.

Our response: We agree that metadata standards should be developed with the input of adopters, and have updated this text accordingly.

Additional thoughts (not sure what section this applies to)

The principles do not address virtual machines. As these are becoming more common and relevant when addressing the reproducibility of research, it is important this form of software is acknowledged. The question remains in which cases should authors cite the current version, which the static archived version, and in which the virtual machine? In this way software is very much a unique evolving research object and might not fit perfectly into the same citation practices and structure as other research objects. In addition, software citation could possibly occur within the virtual machine. This could be added as a use case.

Our response: We feel this has been addressed in Section 5.8, with the explicit addition of virtual machines in addition to executables and containers. This is also an issue that should be addressed further by the follow-on implementation working group.

On persistence of identifier vs. persistence of software

The persistence principle outlined in (4) is a key element in making software citeable. Where software has become part of the record of science not only the identifier and metadata of the software should be persistent, it should also be the goal to keep a persistent copy of the source code, where applicable. This links with the accessibility principle (5).

There are still many open questions about how to resolve package dependencies in the long term, therefore I would not make the persistent access to code a hard

requirement but may add something more specific towards preserving the record of science.

Our response: Our goal is for software citations to point to (persistent) archived source code, but we are not nor could we require this.

Granularity of the citation

One of the key issues with any citation, whether document, individual, or software is the specificity of what is being cited. In the case of publications, there is almost zero specificity most of the time.

It s very easy to cite an entire package even though one function was used. Part of this problem is being solved in the Python world through this project (https://github.com/duecredit/duecredit).

Any citation should have the ability to specify more than just the obvious, but even the obvious would be a good starting point.

The citation/url should therefore allow for greater specificity within a code base. In general though, a provenance record of the workflow would be significantly more useful than a citation from a research perspective.

Our response: We agree that greater specificity is desirable in some cases, but we do not believe this rises to the level of what should be specified or discussed in the principles at this time.

"Software citations should permit ::: access to the software itself"

Under the Access header, the data declaration states that:

Data citations should facilitate access to the data themselves.

Under the same header, the software declaration states:

Software citations should permit and facilitate access to the software itself.

The addition of permit suggests that software citations should also grant the user with permission to access the software. Is this intentional?

It doesn t seem like a good idea to make access a requirement for discovery, so permit might not be helpful in this sentence.

Our response: To avoid confusion, we removed permit and from the accessibility principle.

Access to software: free vs commercial

The section talks about software that is free as well as commercial software. I am not sure whether this is about free as in freedom (or just gratis or freely available), since it is compared with commercial software, which is unrelated in general, see http://www.gnu.org/philosophy/words-to-avoid.html#Commercial.

I suppose that free should be replaced by gratis and commercial be replaced by non-free in that section.

Our response: We think this is sufficiently clear as written.

ACKNOWLEDGEMENTS

While D. S. Katz prepared this material while employed at the NSF, any opinion, finding, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

Work by D. S. Katz was supported by the National Science Foundation (NSF) while working at the Foundation. Work by K. E. Niemeyer was supported in part by the NSF under grant ACI-1535065. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
NSF: ACI-1535065.

Competing Interests

Arfon M. Smith is an employee of GitHub, Inc., San Francisco, California.

Author Contributions

- Arfon M. Smith wrote the paper, prepared figures and/or tables, reviewed drafts of the paper.
- Daniel S. Katz wrote the paper, prepared figures and/or tables, reviewed drafts of the paper.
- Kyle E. Niemeyer wrote the paper, prepared figures and/or tables, reviewed drafts of the paper.

Data Deposition

The following information was supplied regarding data availability:
The research in this article did not generate, collect or analyse any raw data or code.

REFERENCES

AAS Editorial Board. 2016. Policy statement on software. ***Available at*** *http://journals.aas.org/policy/software.html* (accessed 17 February 2016).

Ahalt S, Carsey T, Couch A, Hooper R, Ibanez L, Idaszak R, Jones MB, Lin J, Robinson E. 2015. NSF workshop on supporting scientific discovery through norms and practices for software and data citation and attribution. ***Technical Report***. Arlington: National Science Foundation. ***Available at*** *http://dl.acm.org/citation.cfm?id=2795624*.

Allen A, Berriman GB, DuPrie K, Mink J, Nemiroff R, Robitaille T, Shamir L, Shortridge K, Taylor M, Teuben P, Wallin J. 2015. Improving software citation and credit. ***Technical Report***. ***Available at*** *http://arxiv.org/abs/1512.07919* [cs.DL].

Barnes N, Jones D, Norvig P, Neylon C, Pollock R, Jackson J, Stodden V, Suber P. 2016. Science code manifesto. ***Available at*** *http://sciencecodemanifesto.org* (accessed 18 April 2016).

Bechhofer S, Buchan I, Roure DD, Missier P, Ainsworth J, Bhagat J, Couch P, Cruickshank D, Delderfield M, Dunlop I, Gamble M, Michaelides D, Owen S, Newman D, Sufi S, Goble C. 2013. Why linked data is not enough for scientists. *Future Generation Computer Systems* 29(2):599 611 DOI 10.1016/j.future.2011.08.004.

Chue Hong N. 2011. Publish or be damned? An alternative impact manifesto for research software. *Available at http://www.software.ac.uk/blog/2011-05-02-publish-or-be-damned-alternative-impact-manifesto-research-software* (accessed 17 February 2016).

CRediT. 2016. Consortia Advancing Standards in Research Administration Information. *Available at http://casrai.org/CRediT* (accessed 17 February 2016).

Data Citation Synthesis Group. 2014. Joint declaration of data citation principles. Martone M. (ed.). *Final Document*. San Diego: FORCE11. *Available at https://www.force11.org/group/joint-declaration-data-citation-principles-final*.

Fox P, Signell R. 2011. NSF geo-data informatics: exploring the life cycle, citation and integration of geo-data workshop report. *Final Document*. Troy: Rensselaer Polytechnic Institute. *Availableathttp://tw.rpi.edu/web/workshop/community/GeoData2011*.

Gent I, Jones C, Matthews B. 2015. Guidelines for persistently identifying software using DataCite. *A JISC Research Data Spring Project*. *Available at http://rrr.cs.st-andrews.ac.uk/wp-content/uploads/2015/10/guidelines-software-identification.pdf* (accessed 25 April 2016).

Gil Y, Ratnakar V, Garijo D. 2015. OntoSoft: capturing scientific software metadata. In: *Proceedings of the Eighth ACM International Conference on Knowledge Capture (K-CAP)*. New York: ACM.

GitHub. 2014. Making your code citable with GitHub & Zenodo. *Available at https://guides.github.com/activities/citable-code/* (accessed 10 March 2016).

Gutzman K, Konkiel S, White M, Brush M, Ilik V, Conlon M, Haendel M, Holmes K. 2016. Attribution of work in the scholarly ecosystem. *fi are* DOI 10.6084/m9.fi e.3175198.v1.

Hannay JE, Langtangen HP, MacLeod C, Pfahl D, Singer J, Wilson G. 2009. How do scientists develop and use scientific software? In: *Proceedings 2009 ICSE Workshop on Software Engineering for Computational Science and Engineering, SECSE*. Piscataway: IEEE, 1 8 DOI 10.1109/SECSE.2009.5069155.

Howison J, Bullard J. 2015. Software in the scientific literature: problems with seeing, finding, and using software mentioned in the biology literature. *Journal of the Association for Information Science and Technology* 67(9):2137 2155 DOI 10.1002/asi.23538.

Huang Y-H, Rose PW, Hsu C-N. 2015. Citing a data repository: a case study of the protein data bank. *PLoS ONE* 10(8):e136631 DOI 10.1371/journal.pone.0136631.

Ison J, Kalas˘ M, Jonassen I, Bolser D, Uludag M, McWilliam H, Malone J, Lopez R, Pettifer S, Rice P. 2013. EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats. *Bioinformatics* 29(10):1325 1332 DOI 10.1093/bioinformatics/btt113.

Jackson M. 2012. How to cite and describe software. *Available at http://www.software.ac.uk/how-cite-and-describe-software* (accessed 17 February 2016).

Jackson M. 2014. Oh research software, how shalt I cite thee? *Available at http://www.software.ac.uk/blog/2014-07-30-oh-research-software-how-shalt-i-cite-thee* (accessed 17 February 2016).

Jackson I, Schwarz C. 2016. Debian policy manual. Version 3.9.8.0. *Available at https://www.debian.org/doc/debian-policy/ch-controlfields.html* (accessed 17 April 2016).

Jones MB, Smith AM, Cabunoc Mayes A, Boettiger C. 2014. Minimal metadata schemas for science software and code, in JSON and XML. *Available at https://github.com/codemeta/codemeta* (accessed 25 March 2016).

Katz DS. 2014. Transitive credit as a means to address social and technological concerns stemming from citation and attribution of digital products. *Journal of Open Research Software* 2(1):e20 DOI 10.5334/jors.be.

Katz DS, Choi S-CT, Lapp H, Maheshwari K, Löffler F, Turk M, Hanwell M, Wilkins-Diehr N, Hetherington J, Howison J, Swenson S, Allen G, Elster A, Berriman B, Venters C. 2014. Summary of the first workshop on sustainable software for science: practice and experiences (WSSSPE1). *Journal of Open Research Software* 2(1):e6 DOI 10.5334/jors.an.

Katz DS, Choi S-CT, Wilkins-Diehr N, Chue Hong N, Venters CC, Howison J, Seinstra FJ, Jones M, Cranston K, Clune TL, de Val-Borro M, Littauer R. 2016a. Report on the second workshop on sustainable software for science: practice and experiences (WSSSPE2). *Journal of Open Research Software* 4(1):e7 DOI 10.5334/jors.85.

Katz DS, Choi S-CT, Niemeyer KE, Hetherington J, Löffler F, Gunter D, Idaszak R, Brandt SR, Miller MA, Gesing S, Jones ND, Weber N, Marru S, Allen G, Penzenstadler B, Venters CC, Davis E, Hwang L, Todorov I, Patra A, de Val-Borro M. 2016b. Report on the third workshop on sustainable software for science: practice and experiences (WSSSPE3). *Technical Report. Available at http://arxiv.org/abs/arXiv:1602.02296* [cs.SE].

Katz DS, Smith AM. 2015. Implementing transitive credit with JSON-LD. *Journal of Open Research Software* 3:e7 DOI 10.5334/jors.by.

Knepley MG, Brown J, McInnes LC, Smith BF. 2013. Accurately citing software and algorithms used in publications. *figshare* DOI 10.6084/m9.figshare.785731.v1.

Lipson C. 2011. *Cite Right, Second Edition: A Quick Guide to Citation Styles–MLA, APA, Chicago, the Sciences, Professions, and More, Chicago Guide to Writing, Editing, and Publishing.* Chicago: University of Chicago Press.

Malone J, Brown A, Lister AL, Ison J, Hull D, Parkinson H, Stevens R. 2014. The Software Ontology (SWO): a resource for reproducibility in biomedical data analysis, curation and digital preservation. *Journal of Biomedical Semantics* 5(1):1–13 DOI 10.1186/2041-1480-5-25.

Mayernik M, Maull K, Hart D. 2015. Tracing the use of research resources using persistent citable identifiers. *Poster presented at NSF SI2 PI Meeting, Arlington, VA. Available at https://share.renci.org/SI2PI2015/2015_SI2PI_Posters/mayernik_SI2poster_Feb2015.pdf* (accessed 3 March 2016).

McAdoo T. 2015. How to Cite Software in APA Style. *Available at http://blog.apastyle.org/apastyle/2015/01/how-to-cite-software-in-apa-style.html*.

Morin A, Urban J, Adams PD, Foster I, Sali A, Baker D, Sliz P. 2012. Shining light into black boxes. *Science* 336(6078):159–160 DOI 10.1126/science.1218263.

Norén L. 2015. Invitation to comment on a proposal for a cohesive research software citation-enabling platform. *Available at http://astronomy-software-index.github.io/2015-workshop/* (accessed 17 February 2016).

Parsons MA, Duerr R, Minster J-B. 2010. Data citation and peer review. *Eos, Transactions American Geophysical Union* 91(34):297–298 DOI 10.1029/2010EO340001.

Rowe BR, Wood DW, Link AN, Simoni DA. 2010. Economic impact assessment of NIST s Text REtrieval Conference (TREC) program. *Final Report*. Research Triangle Park: RTI International. *Available at http://trec.nist.gov/pubs/2010.economic.impact.pdf* (accessed 17 April 2016).

Software Attribution for Geoscience Applications (SAGA). 2015. Software for science: getting credit for code. *Available at https://geodynamics.org/cig/projects/saga/* (accessed 6 April 2016).

Sandve GK, Nekrutenko A, Taylor J, Hovig E. 2013. Ten simple rules for reproducible computational research. *PLoS Computational Biology* 9(10):e1003285 DOI 10.1371/journal.pcbi.1003285.

Soergel DAW. 2015. Rampant software errors may undermine scientific results [version 2; referees: 2 approved]. *F1000Research* 3:303 DOI 10.12688/f1000research.5930.2.

Software Credit Workshop. 2015. Software Credit Workshop. *Available at http://www.software.ac.uk/software-credit* (accessed 6 April 2016).

Starr J, Castro E, Crosas M, Dumontier M, Downs RR, Duerr R, Haak L, Haendel M, Herman I, Hodson S, Hourclé J, Kratz JE, Lin J, Nielsen LH, Nurnberger A, Proell S, Rauber A, Sacchi S, Smith A, Taylor M, Clark T. 2015. Achieving human and machine accessibility of cited data in scholarly publications. *PeerJ Computer Science* 1:e1 DOI 10.7717/peerj-cs.1.

Sufi S, Chue Hong NP, Hettrick S, Antonioletti M, Crouch S, Hay A, Inupakutika D, Jackson M, Pawlik A, Peru G, Robinson J, Carr L, De Roure D, Goble C, Parsons M. 2014. Software in reproducible research: advice and best practice collected from experiences at the collaborations workshop. In: *Proceedings 1st ACM SIGPLAN Workshop on Reproducible Research Methodologies and New Publication Models in Computer Engineering (TRUST '14)*. Edinburgh: ACM, 2:1 2:4 DOI 10.1145/2618137.2618140.

Van de Sompel H, Payette S, Erickson J, Lagoze C, Warner S. 2004. Rethinking scholarly communication: building the system that scholars deserve. *D-Lib Magazine* 10:9. *Available at http//www.dlib.org/dlib/september04/vandesompel/09vandesompel.html*.

Ward G, Baxter A. 2016. Distributing Python Modules. *Available at https://docs.python.org/3.6/distutils/setupscript.html#additional-meta-data* (accessed 31 August 2016).

White O, Dhar A, Bonazzi V, Couch J, Wellington C. 2014. NIH Software Discovery Index Meeting Report. NIH. *Available at http://www.softwarediscoveryindex.org/* & https://gist.github.com/mhucka/44921ea1e9a01697dbd0591d872b7b22.

Wickham H. 2015. *R Packages*. First Edition. Sebastopol: O Reilly Media.

Wilson G, Aruliah DA, Brown CT, Chue Hong NP, Davis M, Guy RT, Haddock SHD, Huff KD, Mitchell IM, Plumbley MD, Waugh B, White EP, Wilson P. 2014. Best practices for scientific computing. *PLoS Biology* 12(1):e1001745 DOI 10.1371/journal.pbio.1001745.

Wilson R. 2013. Encouraging citation of software introducing CITATION files. *Available at http://www.software.ac.uk/blog/2013-09-02-encouraging-citation-software-introducing-citation-files* (accessed 17 February 2016).

Submission Date

12/18/2016

Submitter Name

Anton Popov

Name of Organization

National Technical University of Ukraine "Kyiv Polytechnic Institute"

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

biomedical signal processing

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

raw records of the vital signs + verified labels of clinical events (types of interventions and/or changes in the patient state)

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

as long-term records as possible, to be able to develop the algorithms for events forecasting and prevention

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

The absence of the public available labeled data

4. Any other relevant issues respondents recognize as important for NIH to consider

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

b. Inclusion of a link to the data/software resource with the citation in the report

c. Identification of the authors of the Data/Software products

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

4. Any other relevant issues respondents recognize as important for NIH to consider

Additional Comments

Submission Date

12/20/2016

Submitter Name

Jonathan Petters

Name of Organization

Virginia Polytechnic Institute and State University

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

Virginia Tech conducts research in a wide variety of topics

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

I have two recommendations: 1.) Data and code that underlie peer-reviewed publications must be made publicly accessible via a trustworthy repository in open machine-readable formats. Ultimately I would like to see a publication author make their entire digital research workflow accessible and repeatable, but recommend beginning with this requirement only for data displayed in publication tables and figures. Researchers could be encouraged to share more, however. 2.) Communities of practice funded by NIH (perhaps segmented by NIH funded program?) should be requested to nominate a few datasets that would be of value to share broadly for re-use. NIH can then decide to invest (or not) in the resources necessary to share these few datasets from each program.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

For the purposes of scientific integrity we can argue that it data and code underlying publications (as specified above) should be accessible for as long as the peer-reviewed publication version of record is accessible. They could be stored either via institutional repositories or with the publication version of record. For other datasets I am inclined to go with the Uniform Guidance requirement for research records (e.g. 3 years after the research was completed), and could be stored in trustworthy repositories as deemed appropriate by NIH researchers and NLM in coordination with other repository experts. Sustaining access to these data comes at a cost, and as this infrastructure is built we may need to accept that less peer-reviewed publications will be generated through NIH funding in the short-term. However, I expect conclusions in publications where data and code are shared will eventually be more sound, and that eventually the sharing of datasets and code will enable new (and speedier) research pathways.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

Because data stewardship and sharing is neither rewarded nor mandated, NIH must provide both incentives and requirements to get improved data stewardship and sharing off the ground. Project CRediT (http://docs.casrai.org/CRediT) and the FORCE11 guidance on software (https://www.force11.org/software-citation-principles) and data citation (https://www.force11.org/group/joint-declaration-data-citation-principles-final) are great starts in providing a framework for incentives and credit. However messy it might be in the beginning, I do not believe much progress will be made until NIH begins asking and requiring these changes in research behavior (in coordination with other funders and with publishers).

4. Any other relevant issues respondents recognize as important for NIH to consider

(Copied from I.4. above) Within academic institutions there are research data management consultants and librarians like myself who are already helping researchers improve their data stewardship and sharing, and to comply with funder data sharing policies. This includes helping researchers consider how to share data resulting from human subjects research while minimizing the chances of confidentiality breaches. When NIH does update their policies regarding

research data sharing and stewardship, we stand ready to help researchers at our respective institutions comply with those policies. NIH might also be interested in examining data sharing policies from other federal agencies like NOAA and USGS from which they may learn. I participated in a project with the Scholarly Publishing and Research Coalition (SPARC) to extract key information from federal public access plans and data sharing policies. The resulting resource (http://datasharing.sparcopen.org/data) may be useful in learning what other agencies are doing in relatively short order.

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

A persistent unique identifier is a critical piece of information to be included within a data citation; without it another researcher would be unable to identify the particular dataset with some veracity. DOIs are also precursors to being able to locate and access a particular dataset; they can be pointed at a dataset as it the inevitably moves from one server to the other. I am aware that data experts have their quibbles about the limitations of DOIs. However, I find it more important that researchers are becoming more comfortable with the appending of a DOI to peer-reviewed publications, and thus their appending to datasets will not be too jarring of an additional practice.

b. Inclusion of a link to the data/software resource with the citation in the report

Yes, this is also important, especially for data and code associated with the publication. The F in FAIR means findable!

c. Identification of the authors of the Data/Software products

See my reference to Project CRediT (http://docs.casrai.org/CRediT) above; its contribution model incorporates creation of datasets and code. Elevating the importance of dataset and code creators is vital if we wish to elevate the importance of these research products to the same level as peer-reviewed publications.

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

For data associated with a peer-reviewed publication, the most pragmatic advice is to have one DOI that points as the whole of the associated data. For other datasets we can argue (and have argued) about DOI granularity for a long, long time. From a pragmatic view each repository (like ours) has an underlying data model upon which there are rules regarding data organization and DOIs. For now we do our best to explain this underlying data model to the data depositor and then allow them to organize their data for reusability and citability as they wish.

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

Digital repositories can move from server to server and change URLs just like any other digital object. As an example of how to address this issue, see the Registry of Research Data Repositories (http://www.re3data.org/). For each repository in their registry they provide a persistent identifier that could be pointed at the repository as it moves. One could argue that this identifier should be a DOI too, but it is a good start.

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

• Provide for the development of resources and tools (e.g. infrastructure) to ease research data sharing. This includes continued development of the NIH Commons Cloud Credit Model. I have heard it discussed a few times and am optimistic about its potential for success. • Exhibiting the usefulness of data sharing from a few well-planned pilot projects, where small communities of researchers share data to answer research questions they otherwise could not

answer, could go a long way towards demonstrating its potential to skeptical researchers. • Provide incentives/prizes for best documented and best cited datasets. • Work with librarians, data management consultants, and others at academic institutions to help provide practical, effective guidance to both young and established researchers in improving their data management, sharing and security habits.

4. Any other relevant issues respondents recognize as important for NIH to consider

(Copied from I.4. above) Within academic institutions there are research data management consultants and librarians like myself who are already helping researchers improve their data stewardship and sharing, and to comply with funder data sharing policies. This includes helping researchers consider how to share data resulting from human subjects research while minimizing the chances of confidentiality breaches. When NIH does update their policies regarding research data sharing and stewardship, we stand ready to help researchers at our respective institutions comply with those policies. NIH might also be interested in examining data sharing policies from other federal agencies like NOAA and USGS from which they may learn. I participated in a project with the Scholarly Publishing and Research Coalition (SPARC) to extract key information from federal public access plans and data sharing policies. The resulting resource (http://datasharing.sparcopen.org/data) may be useful in learning what other agencies are doing in relatively short order.

Additional Comments

Submission Date

12/21/2016

Submitter Name

David Hansen

Name of Organization

Duke University Libraries

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

All

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

See attachment

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

See attachment.

4. Any other relevant issues respondents recognize as important for NIH to consider

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

b. Inclusion of a link to the data/software resource with the citation in the report

c. Identification of the authors of the Data/Software products

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

4. Any other relevant issues respondents recognize as important for NIH to consider

Additional Comments

NIH letter - 12 21 2016_0.pdf (280 KB)

December 21, 2016

Tim McGeary, Associate University Librarian for Information Technology Services
tim.mcgeary@duke.edu

Joel Herndon, Head of Data and Visualization Services
Joel.herndon@duke.edu

David Hansen, Director of Copyright and Scholarly Communications
david.hansen@duke.edu

Duke University Libraries
411 Chapel Drive
Durham, NC 27708

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

VIA WEB SUBMISSION FORM

Re: Notice Number NOT-OD-17-015, Request for Information: Strategies for NIH Data Management, Sharing, and Citation

This response is from members of the Duke University Libraries who are charged with supporting researchers at Duke in managing and sharing their data. Our comments primarily focus on the questions NIH poses about data sharing strategy development in Section I of its Request for Information.

We are extraordinarily supportive of the overall approach outlined by OSTP and the plan developed by NIH in response regarding increasing access to scientific publications and data. We have seen firsthand that data stewardship and sharing are critical for enhancing the reach and impact of scholarship. We also know that data stewardship and sharing are expensive, and so we write to highlight two major resource barriers that NIH should seek to resolve as it implements its plan.

First, we see resources for education and training as a major barrier. At Duke, we have invested significantly in services to help our researchers understand research data sharing requirements, good data management practices, and how to develop and execute a research data management plan. Duke currently provides a consultation service, web based data management guidance, and training sessions to help researchers address data management concerns. Duke Libraries staff devote thousands of hours annually to these issues. This year alone we were compelled to add four new dedicated data management positions whose work will be entirely focused on researcher concerns with data management planning and workflows. These staff will guide researchers in identifying, structuring, and curating their data for sharing and long term preservation in compliance with university, publisher, and funding agency data management requirements.

One way we believe NIH should address this resource need is by developing and maintaining updated, clear, succinct guidelines and educational tools for researchers on what exactly NIH requires for data stewardship and sharing. That guidance would help both researchers directly and academic and administrative units, such as libraries, who support those researchers in creating and executing research data management plans. NIH should also consider how it supports those who work directly with faculty in support roles, for example, by funding workshops and training for assisting researchers in meeting NIH data sharing requirements.

Second, the long-term resource implications of data stewardship and sharing are a special concern for us. We have significant experience developing and supporting infrastructure to preserve and share scholarly articles, data, digital objects from library collections, and a variety of other items for which long-term stewardship and sharing are considerations. Through that experience, we have learned several things that affect how we determine what resources are needed. For one, the time period for which we must maintain objects is a major planning consideration. Indefinite preservation and access is difficult to plan and budget for. That problem is compounded when, as with research data, the storage needs can vary greatly and in many cases are much larger than needs for storage of research articles or similar research outputs. In our experience, a commitment to preserve and share data for a period of five to seven years seems about right for supporting the needs of researchers while facilitating the planning for a definable investment of financial resources. The recognition by the NIH that this is a new area requiring funding through indirect costs will be significant for the ability of universities and libraries to fulfill the commitments of data retention.

The time commitment also raises other resource challenges related to supporting systems for sharing data. NIH minimum requirements for discoverability will affect how much we invest in enhancing data by developing and maintaining metadata to aid researchers in finding research data. Similarly, requirements for assigning persistent identifiers of data such as DOIs create additional expense. In some circumstances, such as with final datasets, those kinds of identifiers are appropriate.

But for other types—such as citation of in-progress data sets in RPPRs—other less permanent and less costly options such as Archival Resource Keys may be more appropriate.

Thank you for conducting this inquiry. If you have any questions, please feel free to contact us.

Sincerely,

David Hansen
Director, Copyright and Scholarly Communications
Duke University Library

Signing for

Tim McGeary, Associate University Librarian for Information Technology Services
Joel Herndon, Head of Data and Visualization Services

Submission Date

12/21/2016

Submitter Name

Robert Thomas

Name of Organization

Beth Israel Deaconess Medical Center

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

sleep science and disorders

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

Genomics and elated Physiological signals including sleep

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

These data should be available indefinitely, is NIH supported data repositories and sharing sites.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

Federal funding and fee for use, industry use fes

4. Any other relevant issues respondents recognize as important for NIH to consider

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

Credit for publications downstream of data sharing

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

Agree

b. Inclusion of a link to the data/software resource with the citation in the report

Agree

c. Identification of the authors of the Data/Software products

Yes

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

Yes

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

Yes

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

Academic credit can enhance sharing ideals

4. Any other relevant issues respondents recognize as important for NIH to consider

Additional Comments

Submission Date

12/23/2016

Submitter Name

Kevin Read

Name of Organization

NYU Health Sciences Library

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

Epidemiology

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

We recommend a data catalog model as a means to assess the value of many types of datasets for which curation is not standard practice. While a repository requires the resource-intensive curation and stewardship of data, the data catalog model is a low cost, low barrier approach -- requiring only detailed metadata from researchers, rather than the data itself. By empowering researchers to identify and request access to datasets of interest, important data of interest to research communities can be organically identified and selected for future preservation. Researchers who receive requests for their datasets will be motivated to properly curate them. This model prevents the allocation of resources to curation of datasets that do not hold interest for the research community. The data catalog model will also help the NIH identify the datasets and more broadly the data types of highest-priority for resource allocation. Beyond the catalog model, certain data types clearly warrant the allocation of resources needed for curation and sharing of data. De-identified clinical trial data should be prioritized. As discussed in the ICMJE proposal for clinical trial data sharing in NEJM, sharing data will increase the confidence in the conclusions drawn from trials, reduce the potential for study repetition, increase the development of potential new areas of study, and make the data collected from study participants readily available and transparent. See attached report for full response...

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

It should be required as part of funding applications that researchers specify, and provide a justification for, the length of time their data should remain available for secondary research purposes. Separate from the grant review process, domain experts should be engaged by the NIH to conduct regular reviews across data types within their area of expertise to provide input as to their continuing value for secondary use. We recommend that these review boards are held at two-year intervals to assess changes in the landscape within their domain that affect the value of different data types.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

The lack of standards in data collection and dissemination is a considerable barrier for data stewardship and sharing. To the fullest extent possible, standards should be recommended by the NIH for data collection based on the applicant's area of research. Experts in each field (e.g. review panel) should designate certain standards as mandatory for data collection, which can then serve as a baseline for data collection for each research discipline (e.g. CDASH for clinical studies). Stronger educational initiatives should support the broad use of standards. We recommend in-depth training of librarians in the current landscape around and use of standards. This could provide a broad means of dissemination of standards, which has begun to be the case for research data management more generally. We believe the NLM should develop an educational program for in-depth training in standards that will provide librarians with the ability to work across a broad range of research domains to facilitate the adoption of standards within their institutions. See attached report for full response...

4. Any other relevant issues respondents recognize as important for NIH to consider

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

A publication that analyzes a dataset created by another set of researchers is directly, and strongly, reliant on the work of those researchers; the citation of that dataset is an indication of a much larger contribution to the results of that paper than would be the citation of a publication. A metric for data citations should be developed that takes this into account. We suggest a metric that tracks not only the number of papers citing that dataset, but also the number of citations received by each of those publications.

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

A method should be created for assigning DOIs to data that are not accessible via a repository due to the collection of PHI or for repositories/warehouses that store and mediate requests for data with PHI. Because data of this kind is common and relevant for discovery, it is important that they are not neglected. Journal data sharing policies (e.g. PLOS) require that authors make sensitive or identifiable data available through a third party provider, and NIH should have a mechanism for tracking and identifying data of this type.

b. Inclusion of a link to the data/software resource with the citation in the report

Links to data and the software resource should be mandatory. Links could include the DOI for data, links to software available on GitHub through the online repository Zenodo (which assigns DOIs to GitHub software products by proxy), and a researcher's laboratory website where they may be hosting data on a server. The URL for any and all data/software resources should be included within a citation to that dataset/code.

c. Identification of the authors of the Data/Software products

An ORCID identifier should be required for each NIH funded researcher to track the creation or data and software products supported by NIH funding, just as the publisher John Wiley & Sons has made ORCID ID's a requirement for authors looking to publish in Wiley journals. This method would allow the NIH to utilize the ORCID API to pull individual author information and their research outputs into any future infrastructure that may describe research data, publications or software such as the NIH Commons, or NIH BioCADDIE's DataMed.

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

The FORCE11 Joint Declaration of Data Citation Principles (https://www.force11.org/group/joint-declaration-data-citation-principles-final) state that, regarding the granularity of citation, "Data citations should facilitate identification of, access to, and verification of the specific data that support a claim. Citations or citation metadata should include information about provenance and fixity sufficient to facilitate verfiying that the specific timeslice, version and/or granular portion of data retrieved subsequently is the same as was originally cited." The issue with too great a degree of granularity in defining and indexing datasets is that the data from a given study may become fragmented, making discoverability difficult or more cumbersome. In citing a publication, the citation itself refers to the entire document, and the surrounding text allows the reader to determine what aspect of the publication is being referred to in the citation. We suggest a parallel approach for data citation. The unit of the citation should be at the highest level at which the data included can be considered to belong to a single study. The surrounding text and, in particular, the description of the data use in the methods section, should provide the information that would allow a reader to determine precisely which data was used in the secondary analysis.

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

The NIH should collaborate with re3data to ensure that the complete listing of NIH data repositories are kept up to date and identifiable. Digital repositories should not be cited on their own, and instead be included as one component of a citation pointing to specific research data.

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

The NIH should make data sharing in some capacity -- whether using a repository to store data or a data catalog to describe data -- a scoreable element on grants. Within the application, a researcher should clearly state what will be shared, how it will be shared, and how the data or software will be managed and curated so that the data is meaningfully shared and preserved for the length of time it is made available. The white paper "Plan for Increasing Access to Scientific Publications and Digital Scientific Data from NIH Funded Scientific Research" released by the NIH in February 2015 can serve as a good basis from which to begin the process of identifying what elements should be scored in terms of managing, sharing, and preserving research data.

4. Any other relevant issues respondents recognize as important for NIH to consider

Additional Comments

NYUHSL_RDM_RFI_Response_20161223.pdf (88 KB)

NLM RFI Response: NYU Health Sciences Library

Contributors
Kevin Read MLIS, MAS
Assistant Curator
kevin.read@med.nyu.edu

Alisa Surkis PhD, MLS
Head of Data Services / Translational Science Librarian
alisa.surkis@med.nyu.edu

Fred LaPolla MLS
Knowledge Management Librarian
fred.lapolla@med.nyu.edu

Nicole Contaxis MLS
Data Catalog Coordinator
nicole.contaxis@med.nyu.edu

Neil Rambo
Director, NYU Health Sciences Library
neil.rambo@med.nyu.edu

The highest-priority types of data to be shared and value in sharing such data;

We recommend a data catalog model as a means to assess the value of many types of datasets for which curation is not standard practice. While a repository requires the resource-intensive curation and stewardship of data, the data catalog model is a low cost, low barrier approach -- requiring only detailed metadata from researchers, rather than the data itself. By empowering researchers to identify and request access to datasets of interest, important data of interest to research communities can be organically identified and selected for future preservation. Researchers who receive requests for their datasets will be motivated to properly curate them. This model prevents the allocation of resources to curation of datasets that do not hold interest for the research community. The data catalog model will also help the NIH identify the datasets and more broadly the data types of highest-priority for resource allocation.

Beyond the catalog model, certain data types clearly warrant the allocation of resources needed for curation and sharing of data. De-identified clinical trial data should be prioritized. As discussed in the ICMJE proposal for clinical trial data sharing in NEJM, sharing data will increase the confidence in the conclusions drawn from trials, reduce the potential for study

repetition, increase the development of potential new areas of study, and make the data collected from study participants readily available and transparent.

Sharing of observational or qualitative data that is collected from unique study subjects at a single point in time should be prioritized since it cannot be reproduced. This category of data could also be extended to study data where the reproduction of the experiment would be difficult or costly. This approach would make novel and unique data more readily available, and help eliminate the reproduction of similar experiments that may be financially or temporally costly.

Finally, data should be considered high priority if it is collected from humans or animals where they are subject to harm or risk. Sharing data of this type could reduce the potential for the unnecessary replication of these experiments. This approach would not only benefit the subjects of study, but potentially reduce the operating costs devoted to the replication of human and animal experiments.

The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications;

It should be required as part of funding applications that researchers specify, and provide a justification for, the length of time their data should remain available for secondary research purposes. Separate from the grant review process, domain experts should be engaged by the NIH to conduct regular reviews across data types within their area of expertise to provide input as to their continuing value for secondary use. We recommend that these review boards are held at two-year intervals to assess changes in the landscape within their domain that affect the value of different data types.

Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers; and

The lack of standards in data collection and dissemination is a considerable barrier for data stewardship and sharing. To the fullest extent possible, standards should be recommended by the NIH for data collection based on the applicant's area of research. Experts in each field (e.g. review panel) should designate certain standards as mandatory for data collection, which can then serve as a baseline for data collection for each research discipline (e.g. CDASH for clinical studies).

Stronger educational initiatives should support the broad use of standards. We recommend in-depth training of librarians in the current landscape around and use of standards. This could provide a broad means of dissemination of standards, which has begun to be the case for research data management more generally. We believe the NLM should develop an educational program for in-depth training in standards that will provide librarians with the ability to work

across a broad range of research domains to facilitate the adoption of standards within their institutions.

Finally, there should be a level of cost allocation for data management and curation via NIH grant funding. An initial pilot test of data management and curation requirements within a particular field will help identify what resources (personnel, tools, time) are needed and can serve as a proof of concept that can then be more broadly adopted. A focus on training researchers in specific fields on best practices in data management and curation would also serve to improve the quality of data stewardship needed to prepare data for sharing.

The NIH seeks comment on any or all of the following topics:

The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing;

A publication that analyzes a dataset created by another set of researchers is directly, and strongly, reliant on the work of those researchers; the citation of that dataset is an indication of a much larger contribution to the results of that paper than would be the citation of a publication. A metric for data citations should be developed that takes this into account. We suggest a metric that tracks not only the number of papers citing that dataset, but also the number of citations received by each of those publications.

Important features of technical guidance for data and software citation in reports to NIH, which may include:

Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI) (https://www.iso.org/obp/ui/#iso:std:iso:26324:ed-1:v1:en)

A method should be created for assigning DOIs to data that are not accessible via a repository due to the collection of PHI or for repositories/warehouses that store and mediate requests for data with PHI. Because data of this kind is common and relevant for discovery, it is important that they are not neglected. Journal data sharing policies (e.g. PLOS) require that authors make sensitive or identifiable data available through a third party provider, and NIH should have a mechanism for tracking and identifying data of this type.

Inclusion of a link to the data/software resource with the citation in the report

Links to data and the software resource should be mandatory. Links could include the DOI for data, links to software available on GitHub through the online repository Zenodo (which assigns DOIs to GitHub software products by proxy), and a researcher's laboratory website where they

may be hosting data on a server. The URL for any and all data/software resources should be included within a citation to that dataset/code.

Identification of the authors of the data/software products

An ORCID identifier should be required for each NIH funded researcher to track the creation or data and software products supported by NIH funding, just as the publisher John Wiley & Sons has made ORCID ID's a requirement for authors looking to publish in Wiley journals. This method would allow the NIH to utilize the ORCID API to pull individual author information and their research outputs into any future infrastructure that may describe research data, publications or software such as the NIH Commons, or NIH BioCADDIE's DataMed.

Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

The FORCE11 Joint Declaration of Data Citation Principles (https://www.force11.org/group/joint-declaration-data-citation-principles-final) state that, regarding the granularity of citation, "Data citations should facilitate identification of, access to, and verification of the specific data that support a claim. Citations or citation metadata should include information about provenance and fixity sufficient to facilitate verfiying that the specific timeslice, version and/or granular portion of data retrieved subsequently is the same as was originally cited."

The issue with too great a degree of granularity in defining and indexing datasets is that the data from a given study may become fragmented, making discoverability difficult or more cumbersome. In citing a publication, the citation itself refers to the entire document, and the surrounding text allows the reader to determine what aspect of the publication is being referred to in the citation. We suggest a parallel approach for data citation. The unit of the citation should be at the highest level at which the data included can be considered to belong to a single study. The surrounding text and, in particular, the description of the data use in the methods section, should provide the information that would allow a reader to determine precisely which data was used in the secondary analysis.

Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed;

The NIH should collaborate with re3data to ensure that the complete listing of NIH data repositories are kept up to date and identifiable. Digital repositories should not be cited on their own, and instead be included as one component of a citation pointing to specific research data.

Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications;

The NIH should make data sharing in some capacity -- whether using a repository to store data or a data catalog to describe data -- a scoreable element on grants. Within the application, a researcher should clearly state what will be shared, how it will be shared, and how the data or software will be managed and curated so that the data is meaningfully shared and preserved for the length of time it is made available. The white paper “Plan for Increasing Access to Scientific Publications and Digital Scientific Data from NIH Funded Scientific Research” released by the NIH in February 2015 can serve as a good basis from which to begin the process of identifying what elements should be scored in terms of managing, sharing, and preserving research data.

Submission Date

12/24/2016

Submitter Name

John Conway

Name of Organization

Global Director R&D Strategy and Solutions

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

Clinical

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

LabAnswer has strong opinions regarding scientific data management and seems to be in alignment with the FAIR principles. We are approaching and solving these aligning principles for industry, government and academia with the end in mind; model quality or ready data. In order to achieve this a scientific data strategy must be put in place and adopted. This includes (governance, stewardship and ontology/taxonomy management). Discovery research is the ultimate data provisioning team sport. Being able to share discovery results (HTE screening) as well as omics data with the right level of context is high on the list. Sharing measured properties of different modalities is also very important. LabAnswer is responsible building collaborative scientific tools and have witnessed researchers collaborating with data. Model data (predictive/suggestive) backed up or reinforced with measured results makes a for a very collaborative decision support outcome. Another set of data types come from the analytical/characterization groups. The ability to share quality data is proving its worth in many pharma companies when it comes to separation sciences and approaches. A biopharmaceutical discovery research group benefits from data sharing starting in target identification, from hit to lead and optimization and finally preclinical. The omics and targeted knowledge bases point to feasible drug targets. High throughput experimentation and targeted data results assist in the formation of screening campaigns and approaches. Finally, animal data and additional model quality data provide a potential reduction in laboratory work and faster cycle times.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

The most important aspect of the data is not time but relevance. In the right environment not only is data quality, confidence and relevance measured but usage is tracked so that the proper metrics can determine shelf life. Data provenance is also a critical part of this and will prove to be important in data used for building predictive and suggestive models. These variables will drive data reduction and focus in on relevant quality data. The resource implications can probably be modeled today by overlaying this approach on current publically available research data.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

It's LabAnswer’s experience and opinion that some of the biggest barriers to change are culture and the lack of Organizational Change Management (OCM). An OCM strategy will have to be put in place in order to drive adoption. LabAnswer has many referenceable case studies that highlight successful OCM in scientific data management projects. A scientific data management strategy that focuses on a true data provisioning approach will actually reduce the cost and burdens of trying to share data and information without the proper framework. This framework builds a foundation that will eventually provide the auto aggregation of data and information thus reducing the downstream handling and manipulation that so many researchers spend 40-60 percent of their time on.

4. Any other relevant issues respondents recognize as important for NIH to consider

It will be important to understand how the NIH and its data fits into the ensuing ecosystem as producers and consumers. How will it monitor and use metrics for continual improvement and relevance? Lastly, how will address data versioning and the feedback of a true collaborative environment to drive discovery and information and knowledge sharing.

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

This goes back to section 1 question 2 answer when we describe data relevance and the metrics needed to drive. The semi-automated and automated tracking of shared data used in publications drives the relevance which drives the culture of data sharing and collaboration. Imagine a searchable audit trail that allows journal reviewers to interrogate authors and coauthors of shared data usage and referencing within paper or government organization. The use of criteria based unique identifiers will help match and follow the provenance of data.

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

This is exactly what we pointed to in the previous answer. A DOI can be used to link the cited data back to the source. This now allows for automated monitoring and the associated metrics discussed previously. In fact, a similar less stringent approach to mining data out of journal articles has been used successfully for competitive intelligence by companies and governments for years.

b. Inclusion of a link to the data/software resource with the citation in the report

Links are great but eventually break or erode. A smarter linking system could be created that ensures persistence to the data/software resource. There is opportunity here to create smart links that have a persistence.

c. Identification of the authors of the Data/Software products

This again would be part of the criteria of a DOI or other unique identifier. The model must be simple and redundant.

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

This should be part of the citing business rules and governed. The stewardship surrounding this can evolve but will have to well thought out so that business rules are followed and exceptions are minimized and or nonexistent.

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

The governance board should consider this and the previous example and actually work through the use cases to establish the most practical and meaningful business rules.

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

Our recommendation on this would to encourage local and global models built upon appropriate data sets.

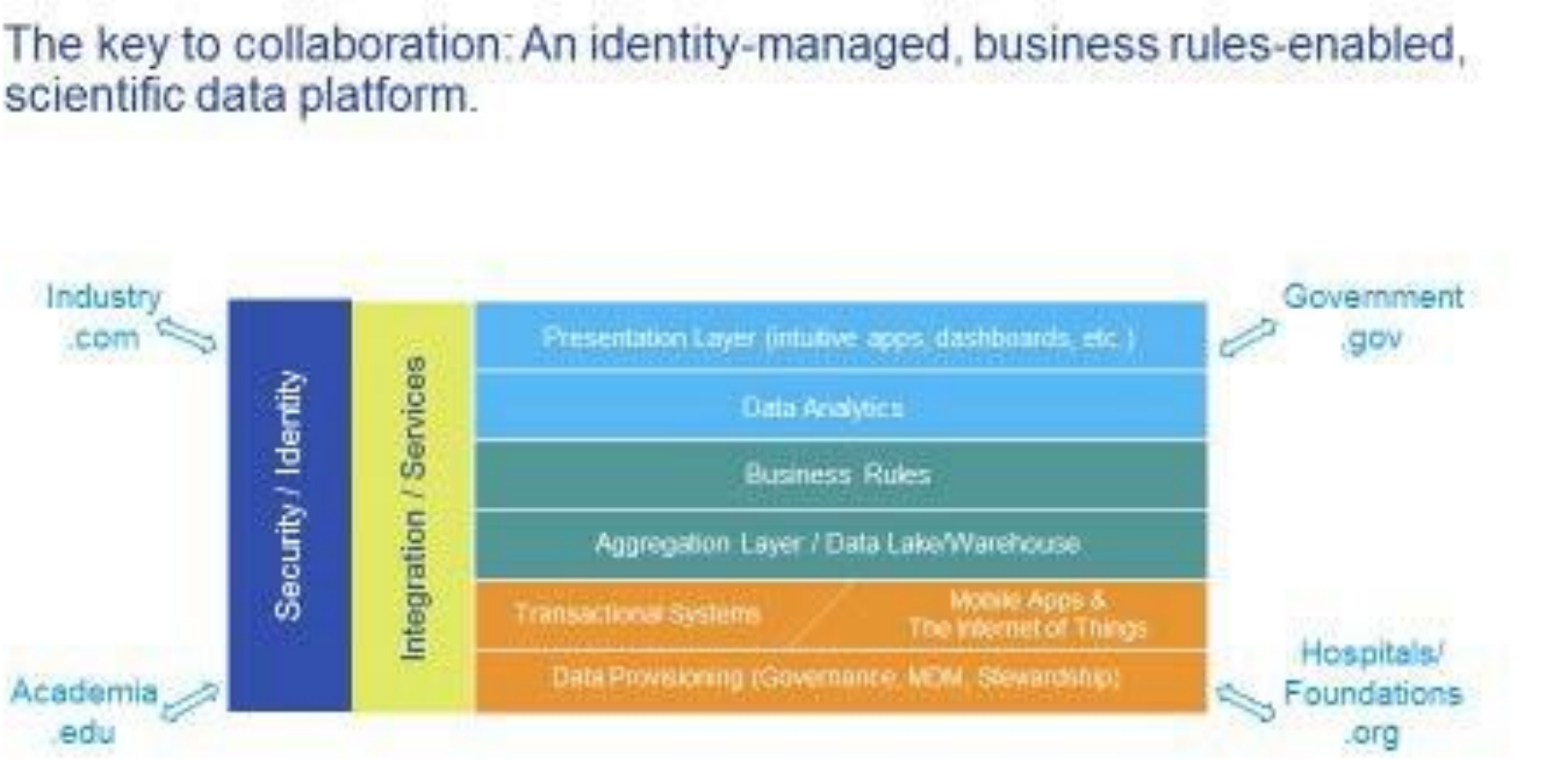
4. Any other relevant issues respondents recognize as important for NIH to consider

It will be important to understand how the NIH and its data fits into the ensuing ecosystem as producers and consumers. How will it monitor and use metrics for continual improvement and relevance? Lastly, how will address data versioning and the feedback of a true collaborative environment to drive discovery and information and knowledge sharing.

Additional Comments

Other Relevant Issue - Collaboration Graphic LA.doc (49 KB)

LabAnswer is working several initiatives in the industry and is members of several community groups that are driving standards and best practices. We have also partnered with Accenture in the building of the Research Life Sciences Cloud offering. This platform approach has garnered the interest of many of the top pharma and biotech companies and the premise is to solve the type of issues outlined in the FAIR paper. We have interviewed 32 companies and built a proof of concept and are planning on development starting Q1 2017. If you see the image below you will see we are taking a holistic approach and know the resulting ecosystem with transform the discovery research environment.



© 2016 LabAnswer all rights reserved. Confidential.

Submission Date

12/27/2016

Submitter Name

Michael Rueschman

Name of Organization

Brigham and Women's Hospital

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

Sleep disorder epidemiology

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

High-resolution physiological and genomic data should be a high priority, given that bandwidth is increasing our ability (everyday) of transferring and analyzing/mining these types of data for useful scientific knowledge.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

Forever. Base the infrastructure for data sharing resources on open source frameworks / packages. Perhaps partner with a major player in the web services world (e.g. Amazon, Google) to host data in the long-term.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

This depends on the type of resources you put into data stewardship and sharing. Cost of IT infrastructure (e.g. web servers, data transfer) is a big one. Beyond that you may want to retain staff time to answer questions from inquiring users and to collate a basic set of documentation about the data being shared.

4. Any other relevant issues respondents recognize as important for NIH to consider

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

b. Inclusion of a link to the data/software resource with the citation in the report

c. Identification of the authors of the Data/Software products

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

4. Any other relevant issues respondents recognize as important for NIH to consider

Submission Date

12/27/2016

Submitter Name

Rich Platt & Adrian Hernandez

Name of Organization

On behalf of the NIH Health Care Systems Research Collaboratory

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

Clinical Research and Public Health

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

We attach the highest priority for sharing to datasets created by conventional randomized clinical trials, for which the participant provides explicit consent for specific kinds of secondary use. In this case, the ownership and permissions can be clearly established. Additional concerns arise regarding reuse of data collected for clinical care or administrative use, especially when research subjects do not give consent. There are new concerns about risk of re-identification of research subjects, and there are additional concerns about the need for the providers and health systems to consent to disclosures about their performance that cannot be de-identified. These data have the capacity to do harm if taken out of context, used inappropriately or for comparative purposes, or to single out an individual, provider or institution. Healthcare systems voluntarily participate in embedded research. Some have indicated that they will not participate in subsequent research if there is a requirement for public disclosure of certain elements of their data. We should consider best practices for inducing participation when the health system is the trial participant and adopt policies, processes, and technical capabilities for data sharing that accommodate health systems' needs. Different solutions will be required for sharing data that supports embedded research compared to those developed for conventional randomized trials. Critical issues are who decides which uses are allowable, who controls access to the data, and how curation, annotation, and access will be supported financially.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

We encourage NIH to assess the experience of its own repositories in terms of the frequency of secondary use over time. Our expectation is that most uses occur within one or two years of completion of the original research, but this should be tested empirically. The NIH Health Care Systems Research Collaboratory trials intend to use each of the following methods, with varying degrees of restriction and cost. • A public archive. In every case, the investigator will either omit some variables, such as clinic site, or group some information. This is necessary to protect the privacy of health systems or providers. • A private archive, such as the NIDDK central repository (which requires a DUA). In these cases, data are aggregate and de-identified by provider and facility. The availability of more repositories for data sharing will help future investigators more effectively and efficiently share data. • A public enclave allows any user to conduct research on any topic, using a protected research environment. The investigator does not take possession of the analyzable data – only the results. • A private enclave, like Yale University Open Data Access (YODA), in which an honest broker or the original owners (like the NIH ABATE study) will determine appropriate use. Private enclaves may establish their own rules regarding users and uses of their data. Decision making can rest with an honest broker or with the original data owner. Private enclaves are the most expensive and restrictive option.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

Data from embedded research may contain substantial information about clinicians' or delivery systems' care practices or business processes. Some uses of those data could create significant risks for clinicians or care systems. The emerging

policies and procedures for data sharing arise from experience with conventional, individually randomized trials and do not consider these risks—and failure to do so may dissuade healthcare organizations from participating in embedded research. To motivate organizations to opt in to embedded research, it will be important to develop a framework that delineates data that can automatically be shared from data that requires review and approval (with approval rights either maintained by the original data provider or delegated). It is also important to recognize the growing amount of multi-center research conducted using distributed methods in which each organization with data creates its own private enclave and executes programs sent by the investigators. They then return results to the investigator. In this situation, the investigator never has possession of a sharable dataset, and it would be necessary to maintain the original distributed private enclave structure and the technical infrastructure to perform the analyses.

4. Any other relevant issues respondents recognize as important for NIH to consider

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

b. Inclusion of a link to the data/software resource with the citation in the report

c. Identification of the authors of the Data/Software products

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

4. Any other relevant issues respondents recognize as important for NIH to consider

Additional Comments

Submission Date

12/28/2016

Submitter Name

Sara Mariani

Name of Organization

Brigham and Women's Hospital

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

Sleep Medicine

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

From the point of view of my research, which consists in the quantitative analysis of biomedical signals (EEG, ECG, respiratory signals), any signals and imaging from specific populations (e.g. premature children, elderly patients affected by cardiovascular disorders, psychiatric patients...) with related clinical outcomes, or annotations for events of interest (e.g. cardiac events, sleep apneas, seizures) are of top priority. In human studies there can be a paucity of data, which limits the statistical power of analyses linking features extracted from these signals to outcomes, or the size of the training and test sets for developing algorithms for automatic signal analysis. Sharing data allows more meaningful analyses, development of accurate algorithms, and the tailoring of outcomes to a more comprehensive variety of individuals.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

The data should, ideally, be made available for as long as they are useful, e.g., until the sensor system, imaging technology, file format, etc. that generated them become obsolete. Of course, there is a trade off between benefit deriving from giving more researchers access to the data over time and the cost and resources needed to maintain the data available, which requires skilled data managers and signal analysis experts to provide assistance with proper usage of the data. In my opinion, the benefits generally outweigh the cost, both for the general progress of science, but also in terms of efficiency (sharing data leads to saving funds for new patient enrollment and data acquisition, especially with rare or difficult to enroll populations).

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

I think there are two main barriers: one related to costs and resources and one related to academic needs in a time of high competition in science. Sharing data requires a robust infrastructure in terms of storage and computational power for upload, download and data management, plus constant expertise ranging from the computer science to the medical level. In addition, research groups may not feel comfortable sharing their data, which is typically the result of lengthy and costly studies, and thus sharing the potential to derive groundbreaking results, with the consequent rewards, with others. Funding processes should reward and encourage the public sharing of data collected in a project, while promoting collaboration, rather than competition, between the research groups who generated the data and other interested researchers, and envision in their budget sufficient resources to cover the maintenance/sharing costs.

4. Any other relevant issues respondents recognize as important for NIH to consider

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

The reporting of data and software sharing in progress reports and grant applications definitely favors proper evaluation of the outreach and value of the funded research. The number of downloads of datasets and software by the author/research team can also be reported.

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

b. Inclusion of a link to the data/software resource with the citation in the report

c. Identification of the authors of the Data/Software products

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

Possibilities could be: grant announcements especially requiring data sharing as one of the output of the project, a dedicated section in the NIH biosketch to describe created open-source software, and a PubMed tool allowing to add links to open-source software to a researcher's bibliography.

4. Any other relevant issues respondents recognize as important for NIH to consider

Additional Comments

Submission Date

12/29/2016

Submitter Name

Chris Bourg, Director

Name of

Organization MIT

Libraries

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

Basic and Applied Bioscience that included, but is not limited to: Bioinformatics, Brain and Cogniti

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

We believe the following are high priority types of data to share. The value in sharing data is that it facilities reproducibility of research, re-use, and discovery. ● Data covered by the White House Directive for federally funded research. ● Data associated with registered clinical trials ● Researcher contributed genomic data to NCBI, including but not limited to GenBank ● Other HHS agency generated data (like the CDC or AHRQ)

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

● These data should be made available for secondary research purposes for a minimum of 10 years recognizing that the data may contain a subset that merits long term preservation and availability. ● The researcher should not be responsible for hosting, maintaining, or sustaining their shared data. ● Funding long term access should not be the responsibility of the researcher who generated the data, but could be supported by the funding agency, institutional repositories, or discipline repositories.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

Barriers to data stewardship and sharing include: ● Lack of awareness and well established workflows for data management best practices ● Lack of tools to facilitate the sharing of data from active phase to storage and sharing phases of research ● There are many choices for where to share/deposit research data. Some communities are well served but researchers may find the variety of choices difficult to evaluate and select from. Other communities need better choices for their research data than those that are currently available to them. ● Need to develop professional norms and incentives that sustain sharing behaviors beyond mandates from external bodies (funders, publishers, institutions) Mechanisms to overcome barriers include: ● Provide support for multiple models of data sharing and storing to identify effective and successful approaches ● Provide support for research and education in data management best practices ● Encourage inclusion of data sharing and storing infrastructures (as such infrastructure emerges) in the research proposal development process ● Transferring the responsibility for sharing and sustaining research data from the researcher to an entity dedicated to that purpose (eg. repository).

4. Any other relevant issues respondents recognize as important for NIH to consider

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

Currently, links to shared data often go through a link in a supporting publication listed in the RPPRs or competing grant

applications. Instead, providing direct links to shared data is efficient and reliable and will increase the use of that data and enhance the researcher reputation. This also promises more accuracy in pointing users to the correct version of the data or software.

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

NIH should be actively engaged in current efforts to develop these systems to ensure the solutions are appropriate for the health and biomedical community; and that the approaches used by the health community are integrated and interoperable with the approaches that are becoming standardized in other scientific fields. In particular, use of persistent unique identifiers should be consistent with the widely endorsed joint declaration of data citation principles (https://www.force11.org/group/joint-declaration-data-citation-principles-final) and software citation principles (https://peerj.com/articles/cs-86/). Persistent identfiers used in citation should be integrated/interoperable with the widely adopted DataCite (https://www.datacite.org/) infrastructure for data, and the ACM publishing policies for software.

b. Inclusion of a link to the data/software resource with the citation in the report

Consistent with the emerging standards referenced in (2a): Wherever a claim in a report relies upon data, the corresponding data should be cited. Similarly software should be cited when the claim made relies on the application of software (often to data) as part of the production of evidence for that claim. (These citations should thus be associated with claims within a report, not with the report as a whole.) Citations need not contain links (URL's), but should contain persistent identifiers that are resolvable via a widely recognized mechanism to URL's.

c. Identification of the authors of the Data/Software products

Data/software citations should facilitate giving scholarly credit and normative and legal attribution to all contributors to the data. However, no single style or mechanism of storing or presenting contributorship is applicable to all use cases. Nor is it necessary for all contributors to appear in a "printed" citation form -- especially where there are many contributors, or contributorship relationships are complex. Authorship information not directly presented in reported citations should be stored in metadata that is associated to the persistent identifier associated with the data (e.g. metadata added to a DOI); or within metadata embedded within the dataset/software distribution package. Identification of authors in citation presentation and/or metadata should enable unambiguous attribution of authors, and of contributorship roles. Consistent with the emerging standards references in (2a) ORCID researcher identifiers should be included for disambiguation of authors; and CASRAI Credit (http://docs.casrai.org/CRediT) metadata should be included to disambiguate contributor roles.

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

Citations should facilitate identification of, access to, and verification of the specific data/software that support a claim. Citations or citation metadata should include information about provenance and fixity sufficient to facilitate verifying that the specific timeslice, version and/or granular portion of the data/software retrieved subsequently is the same as was originally cited. However, no single style or mechanism of storing or presenting contributorship is applicable to all use cases, so long as the combined machine-actionable metadata in the citation, citation metdata (e.g. DOI metadata), and metdata embedded in the resolved object are sufficient for establishing specificity and fixity, for example: ○ Derived/aggregated data sets may be cited directly if they contain internal machine-actional metadata that refers to the individual data sets from which they were derived. ○ Subsets of data used for a figure, table, or analysis need not be separately archived, if a more comprehensive data set is cited, and sufficient metadata is included to automatically identify the subset of data they rely on. ○ Citations should include reference to the version/timestamp of the data or software, however, separate DOI's may or may not be used for separate versions; as long as there is a well-understood machine actionable mechanism for retrieving a specified version based on the combination of persistent identifier and version identifier.

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

Data citations should include persistent identifiers that are resolveable. At least two forms of resolution should be supported -- resolution to a human-readable “landing page” presenting metadata describing the object; and resolution to the instance of the data/software itself (see CODATA/ITSCI Task Force on Data Citation, 2013. “Out of cite, out of mind: The Current State of Practice, Policy and Technology for Data Citation”. Data Science Journal 12: 1-75., .The endpoint of the resolution should be a repository that has an institutional commitment to persistence. The repository may limit direct acces to an instance of the object itself to authorized machine clients.

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

To support community-based transparency and collegial incentives for sharing NIH should publish the data-, software-, and research- plan section of awarded proposals at the same time as, and in conjunction with, publishing the award abstracts. This would enable members of the relevant scientific communities to anticipate what products will be shared and plan accordingly; to establish stable norms of sharing among research communities by aligning expectations, public commitments and monitoring; and enable reviewers to more readily evaluate past work.

4. Any other relevant issues respondents recognize as important for NIH to consider

Additional Comments

Submission Date

12/30/2016

Submitter Name

José Luis Carrillo Alduenda

Name of Organization

Academia Mexicana de Medicina del Dormir

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

Breathing Sleep Disorders

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

For us clinical and translational research data are the most important because they offer usable and directly applicable data to our patients. Sharing these data makes the information become global, generalizable and therefore applicable.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

I think a year is a reasonable period for conducting a secondary investigation with public data. In relation to the means to sustain the data, perhaps a central regulator and controller is most appropriate; in addition, I believe that a public database must recognize its initial generators as authors even though a secondary research was done by another group of researchers.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

I believe that the most important barrier to overcome is the human barrier, researchers must overcome selfishness to share our data, overcome this is not easy but the Spanish Sleep Investigation Network achieved it through coexistence.

4. Any other relevant issues respondents recognize as important for NIH to consider

No comment.

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

No comment.

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

I create a unique identifier as the DOI is appropriate.

b. Inclusion of a link to the data/software resource with the citation in the report

I believe a link is appropriate.

c. Identification of the authors of the Data/Software products

No comment.

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

No comment.

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

Citing a digital repository is difficult, to do so this should be registered in a centralized information core.

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

I believe that encouraging collaboration with Latin American centers is easy, even without subsidies, maintaining direct communication with the most important centers can be a good start.

4. Any other relevant issues respondents recognize as important for NIH to consider

No comment.

Additional Comments

Submission Date

12/31/2016

Submitter Name

Mara Mather

Name of Organization

University of Southern California

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

cognitive neuroscience

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

In terms of cost/benefit analysis, all cognitive behavioral data should be shared, as it can be easily aggregated in text files and thus requires relatively little storage space to be shared. Functional neuroimaging data should be shared because it is such high cost data to collect and can be used by other researchers for other purposes than initially collected for. For instance, the http://fcon_1000.projects.nitrc.org/ efforts have been highly fruitful. My lab and many others have published from these publicly shared data.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

These data should be made available for secondary research purposes for as long as possible, or for at least five years post publication. A one-time fee akin to a open-access journal fee could be collected by sites such as openfmri or openscience framework (osf.io), with that fee covering the cost of hosting the data for the next five years. This fee could be charged to the grant funding the research.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

The biggest barrier is the time and effort it takes for the researcher to make the data accessible. This can be overcome by making it a requirement of any federally funded grant to share data unless an exception is granted. If it were a requirement of funding (as is uploading the publication to the PubMed repository), this would make researchers do it.

4. Any other relevant issues respondents recognize as important for NIH to consider

I think that the public investment of money into NIH makes it imperative that researchers make their projects as generative as possible. Sharing of all data from NIH funded projects that is feasible to share would immensely increase the return on investment on these funds. Yes, it will take some additional researcher time and grant funds to make this possible, but these costs will yield significant benefits to the field that exceed the investment. Furthermore, it will go a long way to reducing fraud and errors in research. Without making it a requirement of funding to share data, most researchers simply will not do it. Ideally, it should be linked to the reporting of publications. Upon publication, data associated with the publication should be released to the public. RPPRs should be required to include information on how the data are shared for every publication listed in the report.

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

I think it needs to be a requirement to report not only all publications resulting from a funded project annually, but

for each publication, to indicate where the data can be downloaded by other researchers.

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

b. Inclusion of a link to the data/software resource with the citation in the report

c. Identification of the authors of the Data/Software products

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

Requiring the data sharing link to be reported in all RPPRs and Competitive Grant Renewals whenever a publication is reported, unless an exception was granted at time of application. If it is not required, most researchers will feel they are too busy to deal with it. If it is required and it truly is too onerous to complete it for some datasets, that case can be made in the application and considered by the review panel.

4. Any other relevant issues respondents recognize as important for NIH to consider

I think that the public investment of money into NIH makes it imperative that researchers make their projects as generative as possible. Sharing of all data from NIH funded projects that is feasible to share would immensely increase the return on investment on these funds. Yes, it will take some additional researcher time and grant funds to make this possible, but these costs will yield significant benefits to the field that exceed the investment. Furthermore, it will go a long way to reducing fraud and errors in research. Without making it a requirement of funding to share data, most researchers simply will not do it. Ideally, it should be linked to the reporting of publications. Upon publication, data associated with the publication should be released to the public. RPPRs should be required to include information on how the data are shared for every publication listed in the report.

Additional Comments

Submission Date

01/05/2017

Submitter Name

Shawn Murphy

Name of Organization

Partners Heathcare

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

Informatics

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

4. Any other relevant issues respondents recognize as important for NIH to consider

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

b. Inclusion of a link to the data/software resource with the citation in the report

c. Identification of the authors of the Data/Software products

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

4. Any other relevant issues respondents recognize as important for NIH to consider

Additional Comments

Partners Healthcare Response to RFI NOT-OD-17-015.docx (17 KB)

Partners Healthcare response to NIH RFI NOT-OD-17-015 [http://grants.nih.gov/grants/guide/notice-files/NOT-OD-17-015.html] - NIH Request for Information (RFI): Strategies for NIH Data Management, Sharing, and Citation

Purpose

The development of a data sharing strategy must deal with at least four issues; 1) what data to focus upon sharing, 2) how to place boundaries on the way data are shared, 3) how to maintain the cost of sharing data, and 4) how to provide adequate motivation for people to support the sharing of data.

Background

Ultimately, the question of what data to focus upon is intimately tied to use cases, and should be directed by a needs assessment for popular use cases. Although this seems obvious, it is surprising how often this question is left open when building a network, which is often assumed to need to encompass a very large number of use cases; an approach that leaves many unmet needs for specific use cases. For example, in PCORNet, the majority of needs were found to be for specific data such as cancer registries, cardiology data, and laboratory testing outside of the initially collected datasets. It would be most helpful to build an infrastructure that can accommodate many forms of data, accommodating multiple networks formed around specific data needs. Each data sharing project can then invest in acquiring the specific data types that are needed for the use case.

Data sharing can focus on sharing at different levels of granularity. For some use cases one wishes to share a “block” of self-contained data such as the complete data collected as part of a previous clinical trial, for others it is continuously flowing data transactions such as the constantly accumulating healthcare data of an Entity. Semantically defined chunks can allow various levels of granularity to be defined and thus allows accommodation to various use cases.

Regarding boundaries for sharing healthcare data, there are consented and unconsented data to consider. Defining semantic data types for sharing is important because consents often specify what semantic type of data can be shared (such as “no HIV labs”, etc.). Thus blocks of data that include “all hospital laboratories” and do not have specific types of semantically-defined data are unsuitable for sharing because they may contain data not covered or prohibited by the consent.

Regarding unconsented data, data obfuscation is important so as not to to include highly specific data in some “de-identified” datasets that are actually easily assigned to a single patient. This, combined with poorly managed chains of custody, can make a “de-identified” data set carry considerable risk for re-identification.

The cost of sharing data involves preparation, legal input, consensus, and software management. The cost of sharing data is especially high in cases of continuously flowing data because many micro-transactions need to be managed, requiring a sophisticated system for accepting the updating transactions, reconciliation, and failure recovery. Well established data sharing applications (github,

dropbox) are not generally geared towards the management of continuously flowing data. Overall, continuously flowing data is easier to use with a "query on demand" approach through web services.

Several proposals have been made regarding attribution for data sharing, including the FAIR (Findability, Accessibility, Interoperability, and Reusability) paradigm. The FAIR paradigm has been formulated for machines to automatically find and use data, in addition to supporting its reuse by individuals. However, motivation to publish data has largely focused on forcing individuals to comply though various mechanisms. Because proper reuse of data often relies heavily upon understanding the nuances of the data, this is unlikely to successfully create better use of available data, although it may increase the gross amount of data available. Journals have been resistant to suggesting that authorship on papers be appropriate for those who curate and provide data, but the effort involved in curating most data sets is often the most important of all tasks for potential authors for a paper. The argument against authorship is that when the data is reused many times it would unfairly skew authorship to data providers. However, this is not usually the case and compared to bench science, for example, when an antibody is developed and reused for many different inventions, typically the provider of the antibody will receive authorship.

Response

Based on the observations above, two suggestions follow:

1) **A permissive, ontology driven data representation strategy – for example, the common STAR type schema for generally representing data, adapted to healthcare.**
2) **Use of web services to wrap the data rather than sending around large "blocks" of self-contained data.**

Many of the above factors would benefit from a permissive data modelling structure, preferably dynamically expandable using an ontology driven approach. This can be linked to indexing and data retrieval paradigms though web services. Advantages are:

With a dynamically expandable data model new types of data for new projects could be easily accommodated. Groups could accommodate specific use cases by creating specialized data sets.

Different data standards could be accommodated as agreed upon by specialized groups. The mappings of one standard to another would be possible with an ontology driven approach to data management.

Various levels of data granularity could be accommodated and indexed, from the distribution of large "blocks" of data, to segmented, semantically specified data on specific patients and variants.

The data presentation can more precisely adhere to privacy needs to allow a more targeted answer for what we "need to know", and accountability for that need, rather that distributing large data sets that have a greater potential to be re-identified.

The ability to lose the chain of custody of data is more limited. This enables attribution to data holders.

New data can be added to the infrastructure without needing to agree upon many of the logistics of sending and updating data.

Submission Date

01/06/2017

Submitter Name

Daureen Nesdill

Name of

Organization

University of Utah Libraries

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

As librarians we support all domains

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

De-identified patient data (in its rawest form), observational data that can't be collected again, long-term or longitudinal research, expensive research paid for with public money.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

Ten years as a minimum guideline and reassess its value at ten years. Ideally, researchers with librarians reassess the value of the data. The long-term resource implication is who pays for storing, archiving, and preserving data. One problem is faculty creating personal project repositories rather than shared solutions.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

A prominent barrier is that long-term responsibility to data is not clear. Universities claim ownership but don't always provide the means to archive data long-term. Many funding agencies are not providing clear guidance on where and how to store data. Faculty don't want to use research funding for sharing; they think data storage should be provided by the agencies or part of the university infrastructure. Mechanisms for overcoming the barrier are increasing communication from NIH with university administrations, clarifying responsibilities, and providing guidance and support for creating data infrastructure and services.

4. Any other relevant issues respondents recognize as important for NIH to consider

NIH needs to provide standards and training in de-identification and documentation of datasets.

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

We support increased reporting of data and software sharing in RPPRs, e.g.what has been shared and how, any changes in selection of repository, etc. . We support using data management and sharing practices appropriately as part of post-award assessment and for new grant applications.

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

At what level does a dataset get a DOI? At the project level? The dataset level? An information bundle with datasets and

documentation? A few vendors (Figshare, LabArchives) are allowing researchers to mint DOIs at will. We need standards. Also, how can we guarantee that each dataset gets only one persistent identifier and who maintains those identifiers?

b. Inclusion of a link to the data/software resource with the citation in the report

Yes, in the final report as long as the DOI is the final DOI for the data. In-progress datasets don't usually have DOIs.

c. Identification of the authors of the Data/Software products

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

Ideally, a researcher should be able to cite the data that was used rather than a large study where it might be difficult for a secondary user to pinpoint which data was used. However, there should also be a citation for the full project. For example, with conference proceedings, you can cite the entire proceedings or individual papers within the proceedings. Parallel to that, you should be able to cite data for each article and/or for the entire project. We really like the "Granularity" section on this DCC page: http://www.dcc.ac.uk/resources/how-guides/cite-datasets#sec:elements

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

Research is moving towards using electronic lab notebooks for the research which store the data in secure clouds. Security includes identifying individuals by one logon one person, therefore this will become decreasingly feasible.

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

Training grant reviewers in assessing DMPs and be specific about what information should be included. Develop a rubric they can follow.

4. Any other relevant issues respondents recognize as important for NIH to consider

NIH needs to provide standards and training in de-identification and documentation of datasets.

Additional Comments

Submission Date

01/06/2017

Submitter Name

Madhvi Upender

Name of Organization

Awarables

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

Sleep and related research

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

Raw data files from sleep studies (including PSG, home sleep monitors, actigraphy)

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

4. Any other relevant issues respondents recognize as important for NIH to consider

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

b. Inclusion of a link to the data/software resource with the citation in the report

It would be extremely beneficial if publications include a link to the raw data that is used in the research study. This allows for cross validation and potential advancements.

c. Identification of the authors of the Data/Software products

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

It may be of value to allow access/interaction of the study researcher to interested parties who can utilize their data for further development or application.

4. Any other relevant issues respondents recognize as important for NIH to consider

Additional Comments

Submission Date

01/07/2017

Submitter Name

Livia Dinu

Name of Organization

Engineering Custom Solutions, Inc.

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

Cognitive Neuroscience

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

Summary of the subject protected data to be shared and presented at the highest scientific level of comprehension. Set Authority level clearance as Citation Strategy for access. The value is not only awareness and recognition rewards, most importantly the limitless value of streamlining this data into results processing with relevant application technology already existing. Develop Transactional process.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

Most likely a 12 months availability, then a sustainability cost to be applied for availability up to 10 years, if applicable.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

Changing the status quo is the most significant burden impacting time and cost in an undesired fashion. Data stewardship is a complex and significantly challenging journey, considering the new Digital MarketPlace demand for smart relevant technologies and automatic Applications results such as documents, reports, graphs (non-human, however verifiable).

4. Any other relevant issues respondents recognize as important for NIH to consider

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

b. Inclusion of a link to the data/software resource with the citation in the report

c. Identification of the authors of the Data/Software products

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

Incentivizing data sharing without a transactional value process in place is at odds with competition and innovation efforts.

4. Any other relevant issues respondents recognize as important for NIH to consider

Additional Comments

Capability Statement_ ECS,Inc. HHS.pdf (193 KB)

TechSolve4U

ECS,Inc.

Information Systems and Engineering
Data Technology Solutions

WE help solve Tech FAST and Act SIMPLE

CAPABILITY STATEMENT

www.techsolve4u.com

ENGINEERING CUSTOM SOLUTIONS,Inc. is a Management Information Systems and Engineering Data Technology Consulting Firm, founded in 2012 and servicing a highly strategic and specialized niche.
We have implemented cost savings solutions and executed strategic action plans for compliance and industry standards upgrades for manufacturing in the highly regulated Healthcare, Medical Equipment, Chemical and Consumer Products industries.
Developing the Digital MarketPlace with relevant new technology and author of the best CRM application and data collection solution for any business.

CORE COMPETENCIES

Critical / Innovative Thinking

Vision/Strategic Planning
Root Cause Assessment / Analysis
Team Development / Facilitation
LEAN Operations / General Management
Organizational Development /Transformation
DecisionMaking / Proactive
Corporate Culture / Process Change
Relationship Building
Turnaround Strategies
High Emotional Intelligence Leverage

BUSINESS SNAPSHOT

CAGE CODE: 7RAD2

SBA

DUNS Number: 085576664

GOVERNMENT BUSINESS POC:
Livia Dinu
Email: techsolve4u@gmail.com
Phone: 440-623-2209
Address: 22517 Christopher Ct.
Strongsville, OH 44149

PAST PERFORMANCE

SLY Incorporated

* We delivered for Sly Inc. a compettitve edge solution; a novel tool and creative work to upgrade existing quoting process to a digital technology application Web program,including beta testing, UI process creation and training manual.
Contact: Mick Ruggiero, Inside Sales Manager; (440)891-3200

Linde

Services

**** We had delivered various project management and engineering services for Linde Gases NA, Medical Oxygen cylinder filling operations division, at 40 plants.
Data collection, analysis, assessment of needs for compliance and action plan implementation was executed for each location.
Managed design, planning, budgeting, scheduling and execution of projects like: project initiation, scope definition, turnkey systems design, plumbing, medical gas and fire protection systems design, acquisition contract award and commissioning.
Contact: Anne Ghorashi, Head Applications Equipment Engineering; (440)525-3704; anne.ghorashi@linde.com.

SOCIO - ECONOMIC STATUS

SBA **Women-Owned Small Business**

NAICS CODES :

541330 – Engineering

541512 - Computer System DesignServ.

541690 – Other Scientific and Technical Consulting Servs.
541380 – Product & AnalyticalTesting

DIFFERENTIATORS

-Solving industry wide market pain in terms of time and data management, with enterprise value-added Web application.
-Provided input on NIH RFI on Strategies for Data Management,

Sharing and Citation ; Patent Quality Conference

PROSPECT PARTNERS

NIH USPTO NIST

IdeaScale

Hyland Software **Linde**

Submission Date

01/10/2017

Submitter Name

Steven Ruggles

Name of Organization

University of Minnesota

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

Population

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

All data used in research publications should be shared. If it can't be shared, it should not be used in peer-reviewed research, because then the cannot be verified or replicated.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

With the exceptions of data that can be easily replicated (e.g. from simulations or easily replicable experiments) or data that has little value, most research data should be maintained over multiple decades. Generally the costs of archiving are declining; whenever the costs of archiving are significantly lower than the costs of reproducing the data, preservation should be a priority. For data that are impossible to replicate--such as surveys or administrative records pertaining to a particular time and place--preservation should be the highest priority.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

Much data used for health research has inadequate metadata. Producing the metadata needed for archiving can be expensive.

4. Any other relevant issues respondents recognize as important for NIH to consider

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

NIH should coordinate with NSF and other federal funding agencies to develop a common standard for data citation, probably involving DOIs.

b. Inclusion of a link to the data/software resource with the citation in the report

If there is a DOI, that should be sufficient. URLs are not persistent and should not be used for data citation.

c. Identification of the authors of the Data/Software products

Data are intellectual products just as much as publications, and the human authors--not just institutions--should always be identified.

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

This is tricky, and I don't think it is possible to develop a common standard that will work in all instances.

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

I believe the DOI should cover this.

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

Require data citations in publications that appear in PMC.

4. Any other relevant issues respondents recognize as important for NIH to consider

Additional Comments

Submission Date

01/11/2017

Submitter Name

James Poterba/Jonathan Skinner

Name of Organization

National Bureau of Economic Research

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

Health Economics, Determinants of Health, Health Trends

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

The highest priority data products to promote within the research community are: (i) New data sets that could be used by the broad research community. These may have been created by a researcher or secured by a researcher through collaboration with data providers such as government agencies or private firms to secure data access. (ii) Derivative data sets, drawn from large public-use data files, that include new data products that are based on the existing data files. Examples of such data products would be an index of health status from a collection of data elements in a household survey such as the Health and Retirement Study, or a set of variables from another data set that have been linked through the use of a synthetic or actual data-linking procedure. (iii) Software code that other data users can use to create sub-samples of a public data set or to merge one data set with another. An example might be code that links location-specific indicators of health service availability to individual level data on health care outcomes. Making code available for the construction of new variables is important, because that enables the new variables to be verified and replicated by follow-on users.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

Most of the cost of making data available for public use is incurred when creating the accessible data product and depositing this product in the archive. Long lifetimes for public access are therefore reasonable, and permanent access is not unreasonable. In some cases, however, if the data product must be accessed through a web-based interface for example and there are updates to the software platform that support the interface, there may be ongoing costs associated with maintaining data access. In the case of such data products, NIH grantees should be encouraged to apply for research funds that will support long-term access. If the NIH mandates that long-term access is required, then budgets for new grants should recognize that there are in some cases costs to such access. It might be useful to specify a horizon over which access must be maintained with updated software (5 years?), and to provide funds for that time. It is difficult to envision very long-term access support from research grants, and it is important that NIH avoid standards that would impose long-lived, unfunded responsibilities on PIs who are creating data.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

There are well developed archiving standards and database repositories for new data collections, addressing issues of confidentiality and the protection of sensitive individual information. These standards respect the constraints imposed by HIPAA and by other related regulations. Crafting protocols for addressing privacy and related issues can be a time-consuming task for the PI and associated research team on a project, and this can draw resources from other aspects of the research. If the NIH could specify acceptable common protocols, and offer a “safe harbor” to researchers who used any of these protocols, that would reduce the burden of creating archival data sets. In many cases, health economists use administrative data that can only be accessed under highly restrictive data use protocols and data sharing restrictions. A good example is the Medicare claims data that are provided by the Centers for Medicare and Medicaid

Services (CMS). An NBER-affiliated researcher working with the CMS data may not make that data publicly available. A document describing how data access was obtained, and any code that was used to carry out the research, could however be made available in the public domain. Such information could be archived in a central place such as ICPSR.

4. Any other relevant issues respondents recognize as important for NIH to consider

Outreach to the research community about the importance of data access, code access, and data archiving would improve the diffusion of best practices. NIH could develop training modules, like those about responsible conduct of research and human subject protection, to inform researchers about archiving principles and options, and to highlight ways to create reproducible data sets and to maintain a code archive. Presentations could be included in professional meetings to further raise the visibility of these issues.

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

NBER health economists have directed substantial and growing attention to issues associated with the sharing of new databases, database enhancements, software code, and other research products. These achievements can and should be the subject of reporting in grant progress reports and final reports. The reporting could also include information, where possible, on the cost of developing these archived data products.

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

b. Inclusion of a link to the data/software resource with the citation in the report

Progress reports should be specific about where data has been, or is going to be, archived. They can provide links to the archive location.

c. Identification of the authors of the Data/Software products

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

Within health economics research, the typical study would rely on a small number of data sets so a link to those data sets, located through the links provided in working papers or published research products, would be an appropriate level of aggregation.

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

Some flexibility in the archiving of data and research products is valuable in light of the wide variation in the types of data and code sharing that are applicable in different research settings. NIH-based advisory guidelines would however aid researchers in designing dissemination plans and in reporting on these plans. It would also be helpful to disseminate information about new public resources such as new data repositories. The NIH could also require researchers to include information on data and related research products within scientific publications and other research reports on NIH-funded grants, just as researchers include references to the grant funding itself. NIH could play an important role in developing “safe harbors” for researchers and thereby offer some certainty about whether particular data access plans meet all necessary requirements. These safe harbors could include data citing and formatting standards and a list of acceptable repositories.

4. Any other relevant issues respondents recognize as important for NIH to consider

Outreach to the research community about the importance of data access, code access, and data archiving would improve the diffusion of best practices. NIH could develop training modules, like those about responsible conduct of research and human subject protection, to inform researchers about archiving principles and options, and to highlight ways to create reproducible data sets and to maintain a code archive. Presentations could be included in professional meetings to further raise the visibility of these issues.

Additional Comments

Submission Date

01/13/2017

Submitter Name

Matthew Dougherty

Organization Name

National Center for Macromolecular Imaging, Baylor College of Medicine

Additional Comments

Data management and sharing activities
Interoperability and archiving of imaging data

Imagery is foundational in scientific and medical research. It is frequently at the center of scientific conclusions and its dissemination.

Imagery represents the bulk of collected empirical data, and it is also the product of post processing and analytical pipelines. It is used ubiquitously throughout science, engineering and medicine.

Images come in many dimensions and modalities, such as 1D (e.g., base-four DNA sequences), 2D (e.g., gray scale electron micrographs), 3D (e.g., tensor MRI), and n-dimensional multi-modal (e.g., multiple detectors+time PET-CT).

Within many research communities there are semi-duplicated non-interoperable imaging formats, making image utilization and archiving difficult. To large extent this cannot be avoided, metadata definitions have uniquely evolved through specialized sub-communities researching and competing for domain-relevant specifications. For them to be Findable, Accessible, Interoperable, and Reusable, the best that might be hoped for that is that they are self-describing formats providing sufficient detail for the archiving, operational management, and utility of imagery. Having a standard image infrastructure to draw upon would provide the means to assimilate image metadata in a uniform way in regard to describing the metadata and provide uniform versioning of these legacy formats.

But the essence of images are pixels, quanta constituting the vast majority of bytes that compose an image. It is the organization of pixels, the multi-modal n-dimensional array, that is most amendable to interoperability, compression, performance, parallelization and generic standardization. Targeting pixels and amalgamating community metadata, including adjunct datasets and files, is needed for a comprehensive generic research image standard.

Some researchers can sufficiently rely on consumer based image standards (e.g., TIFF, PNG and MPEG) as their primary scientific image format. But when image complexity or performance demands are great there is not one generic advanced image standard to baseline a design from, requiring researchers to create their image designs from scratch with inherent limitations of time, skill sets and funding, often because good data management is not a goal in itself.

If such a standard generic, n-dimensional image format existed presumably it would have advanced computational infrastructure, such as self-describing extensibility, pixel chunking, compression, mipmaps/pyramids, RDF definitions, archival tools, and the ability to assimilate non-image research metadata, including encapsulating whole files such as PDF publications and spreadsheets.

Two integrated digital instruments are strategically needed for good data management:

1) Standardization of a *multi-dimensional multi-modal image format* that is open, free, and universally implementable. This would provide an extensible baseline for research communities to devise their own community specific image formats through a self-describing infrastructure. Such an instrument would remove tremendous obstacles to image accessibility and interoperability, increasing the expectation for reusability and archiving.

2) A scientific data container capable of being mounted as a filesystem, supporting this image standard. This advanced image filesystem should be able to encapsulate files such as legacy image formats and adjunct metadata. This data container would provide computational mechanisms for interoperable conversions between legacy image formats and the *standardized multi-dimensional multi-modal image format*. Consider filesystems have traditionally been under the domain of computer operating systems and media manufacturers, whose corporate interests and goals may not align with science data objectives such as interoperability and archiving (i.e., open, free, and universally implementable). Also, translating files and data across different operating systems is not an obligation of any one computer manufacturer; further it is a major nuisance for researchers to copy data into various file systems formats to obtain operating system interoperability.

Both of these instruments can be developed through the Hierarchical Data Format. HDF was originally created by the National Center for Supercomputer Applications (NCSA) twenty-five years ago precisely for managing and curating scientific research data. HDF is the fundamental data unit for NASA's Earth Observation System and for NOAA research data; therefore, it is expected that HDF will be the long-term data infrastructure for the stewardship of this research data. HDF may not be suitable for all biomedical or scientific research, but it has demonstrated its suitability for **many** scientific and biomedical data formats. They key import is that there is no other established data infrastructure that can be drawn upon which is comparable in terms of design, performance, reliability, an established track record by its institutional management, and its anticipated longevity.

It is proposed:

1) The standardization of a *generic multi-dimensional multi-modal image format* either through ISO or ITU. This would provide at least one advanced scientific image standard that domain-communities could baseline their format designs for their image data needs.

2) The image format should have a filesystem container mechanism that allows the amalgamation of non-image datasets such as PDF files, spreadsheet files, XML metadata, etc.

3) The use of HDF as the underlying construct for a *generic multi-dimensional multi-modal image format* and filesystem.

4) The creation of intrinsic software infrastructure that includes error correction, hash functions, crypto, RDF definitions, globally unique and persistent identifiers, compression, pixel chunking, versioning methods, mipmaps/pyramids, and efficient dimensional pixel extraction.

5) Development within the filesystem a mechanism to transparently convert legacy image files into the *standardized multi-dimensional multi-modal image format*. Mounted through a filesystem, legacy image files would perform normally as expected in legacy software; and through the HDF API the pixels and metadata would be accessible as an advanced image format.

6) NIH should co-develop with NIST, NSF, and other governmental agencies through NITRD, in the development of this image standard in order to sufficiently fund, engage wide consensus, and maximize success.

7) Fiscally support the development, demonstration, and publication of best practices as a necessary task in achieving optimal image format designs by different domain communities.

Publication would provide detailed explanations of prior image formats, thus assisting research communities in planning and implementing new image formats.

Having a standardized scientific image format that is deliberatively defined, which research communities can optionally draw upon as boiler-plate, will provide a useful pathway to achieve FAIR objectives.

SECTION I. Data Sharing Strategy Development

NIH recognizes that many factors must be considered when determining what, when, and how data should be managed and shared. These factors include, for example, the purpose for sharing, supporting data re-use and reproducibility, maturity of the science, the infrastructure uniqueness of the data, and ethical considerations.

The NIH seeks comment on any or all of the following topics to help formulate strategic approaches to prioritizing its data management and sharing activities:

1. The highest-priority types of data to be shared and value in sharing such data (Maximum: 250 words)

Imagery of various dimensions and modalities are coin of the realm. See attachment.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications (Maximum: 250 words)

The length of maintaining imagery is highly variable, from milliseconds to decades. It would be desirable to have a standardized generic, self-describing, multi-dimensional multi-modal image format with metadata that includes the description of provenance, verification of integrity, and other fundamental archival information; at the same time having high operational computational performance.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers (Maximum: 250 words)

The lack of a standard image format having an advanced infrastructure, part and parcel to its design, is a major barrier to achieving FAIR. Such infrastructure would include efficient dimensional pixel subset extraction, mipmaps/pyramids, RDF definitions, lossy/lossless pixel compression, and a data container filesystem capable of non-image data encapsulation.

There exists no standardized generic multi-modal multi-dimensional image format, therefore researchers must choose to either utilize consumer image formats or they must design their own unique formats. Either way, these formats frequently lack critical strategies and infrastructure needed for research interoperability, performance, and archiving.

The path that can overcome these barriers is standardization of a generic scientific image format optimized for research. Such a standard could be used as a reliable baseline by research communities for the design of their specific image needs through extensible methods within the standard.

4. Any other relevant issues respondents recognize as important for NIH to consider (Maximum words: 250)

NIH should co-develop with NIST, NSF, and other governmental agencies through NITRD, in the development of this image standard in order to sufficiently fund, engage wide consensus, and maximize success.

http://osp.od.nih.gov/content/nih-request-information-strategies-nih-data-management-sharing-and-citation

Submission Date

01/16/2017

Submitter Name

Holly Falk-Krzesinski

Name of Organization

Elsevier

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

All areas

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

The definition of research data differs from field to field, but broadly speaking it refers to the result of observations or experimentations that validate research findings and can include, but are not limited to: raw data, processed data, software, algorithms, protocols, methods, materials and methods descriptions as well as lab notebook entries. Research data do not include text in manuscript or final published article form, nor do they include other/supplementary materials submitted and published as part of a journal article. In principle, all data should be stored, with an emphasis on two points: 1) When it is harder, more costly, or even impossible to reproduce a dataset, data storage is essential, for example: - Human-subject studies, where patients or other participants have spent time and effort to make themselves available and samples have been gathered; - Data where animals have been sacrificed to enable research; - Non-replicable data such as environmental observation studies, where the conditions of study cannot physically be reproduced because they present a view on a moment in time. 2) Where possible, the rawest form of data should be stored, as well as any software, scripts and methods to interpret or reformat this data. For example, in the case of questionnaires the original answers as well as any software to process these must be preserved. This approach enables a replication of the analysis work, which can support rigor; secondly, it allows other researchers to reuse the raw data, which supports data reuse.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

Obviously, the longer data can be stored, the better, but also obviously, there are costs involved. In general it seems important that raw data is stored well beyond the period in which the data might be reexamined: in most domains this will mean a period of 10 years or more, though in the case of human study or observations of natural phenomena (e.g. in ecology, epidemiology, etc) it would be worth looking at much longer time spans in the order of 10 – 50 years. Where it is not possible to preserve data in perpetuity, an indexing and abstracting service such a Scopus could preserve metadata associated to a specific dataset, and allow for citations and permanent reference. In Elsevier's Mendeley Data repository, for example, data is preserved in perpetuity via an agreement with DANS (Data Archiving and Networked Services), whereby DANS archives every dataset posted to Mendeley Data which passes the internal review process : refer to https://www.knaw.nl/en/news/news/collaboration-dans-and-mendeley-on-archiving-datasets. Mendeley Data can therefore guarantee that any data deposited will always be available at the DOI provided.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

A barrier to data stewardship is uncertainty of long-term funding of data repositories: the fact that many repositories are judged and funded in competition with research leads to great uncertainty: refer to https://www.rd-alliance.org/sites/default/files/case_statement/RDA_WDS_IG_Publishing_Costs.pdf. The fact that many repositories are funded via many different routes further enhances the burdens to seek funding sources by the repository directors, who should be focusing their efforts on providing the best possible data curation support. In other cases, repositories are

informed that their grants will not be renewed and are encouraged to seek alternate funding models, without being given the time or resources to procure those funding sources. It's critical the NIH work cooperatively with other funders globally to develop models for long-term data infrastructure support and develop clear guidelines within and between agencies and divisions as to the type of work that will be supported. Additionally, career opportunities and the acknowledgement and promotion trajectories of data stewards/curators are currently under-supported. Lastly, we point to the barriers mentioned in http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0118053#pone.0118053.ref059, specifically, a lack of clarity regarding rights and privacy issues concerning human data. A clear legal understanding of the rights of use of research data are needed, especially in medicine and in the social sciences. Funding agencies could play an important role here, and educate researchers on the copyright and need for anonymization of human subjects data they collect.

4. Any other relevant issues respondents recognize as important for NIH to consider

Elsevier is eager and enthusiastic to remain involved in the next stages of discussion on this important topic, and are looking forward to continuing and expanding our current engagement and collaboration related to research data with the NIH. In addition to the current response, Elsevier has submitted responses to all research data-related NIH RFIs in the last two years, including: • 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) --> Refer to Comment 5 only • NOT-AI-15-045, Input on NIAID Data Sharing Repository, Immunology Database and Analysis Portal (ImmPort), and Services • NOT-ES-15-011, Input on Sustaining Biomedical Data Repositories • NOT-OD-16-133, Metrics to Assess Value of Biomedical Digital Repositories Copies of these other related RFI responses are appended here as an attachment for reference.

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

Reporting the software and data researchers' create can provide additional evidence of usefulness of the products of funded research and may help enforce data sharing mandates. Elsevier, the NIH, and other stakeholders can work together to create a coherent ecosystem that allows many different paths (dependent on domain, role, and personal preference) for scientists/scholars to identify, report, and track data sharing and reuse practices. Funding agencies' data sharing policies are named as a key factor to encourage academic data sharing (e.g. http://dx.doi.org/10.1371/journal.pone.0118053 and doi: 10.1111/j.1755-263X.2012.00259.x). However, funding policies still show varying degrees of enforcement: achieving clarity and correspondence between funding programs (within/between funding agencies) is a key factor to encourage compliance with data sharing mandates: http://journals.sagepub.com/doi/abs/10.1177/1745691613491579. Elsevier is a leader in highlighting the association between articles and data, and including data as an output associated with a specific author/institution. Using quality filters similar to that Scopus uses to index articles, we enable data and software citations for evaluating the scientific output of a single researcher/institution.

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

Regarding citation of software and data, Elsevier is an active supporter of the Force11 Data Citation group, https://www.force11.org/group/dcip, as shown by the recent implementation of these standards in our over 1,800 journals, https://www.elsevier.com/about/press-releases/science-and-technology/elsevier-implements-data-citation-standards-to-encourage-authors-to-share-research-data. Within the Force11 DCIP Publisher Early Adopters group , we are coleading efforts to develop a joint Force11 DCIP Data Citation Roadmap for science publishers, due for publication shortly. Regarding the specific use of persistent identifiers, we fully support the recommendations provided by Force11 DCIP Repositories Early adopters group, who prepublished their Roadmap, https://doi.org/10.1101/097196, which explicitly states: • All datasets intended for citation must have a globally unique persistent identifier that can be expressed as unambiguous URL. • Persistent identifiers for datasets must support multiple levels of granularity, where

appropriate. • This persistent identifier expressed as URL must resolve to a landing page specific for that dataset. Within Elsevier's data repository, Mendeley Data, we enable unique identification of data versions, as well. When a published dataset is edited, the last digits of the data DOI will change to reflect a new version of the dataset. Regarding software citations, we support the principles published by the Force11 Software Citation Working Group, https://www.force11.org/software-citation-principles, on Unique Identification, which states: "A software citation should include a method for identification that is machine actionable, globally unique, interoperable, and recognized by at least a community of the corresponding domain experts, and preferably by general public researchers."

b. Inclusion of a link to the data/software resource with the citation in the report

With regards to links between publications and data, within the aegis of the RDA Data Publishing Group, https://rd-alliance.org/groups/rdawds-publishing-data-services-wg.html, we have helped lead the development of a Linked data demonstrator and set of guidelines, the Scholix Initiative, http://www.scholix.org/. Scholix and the accompanying DLI aggregation service offers a high level interoperability framework for exchanging information about the links between scholarly literature and data, https://www.icsu-wds.org/news/news-archive/rda-and-icsu-wds-announce-the-scholix-framework-for-linking-data-and-literature. With regards to software access, we again concur with the Software Citation principles on accessibility: "Software citations should facilitate access to the software itself and to its associated metadata, documentation, data, and other materials necessary for both humans and machines to make informed use of the referenced software"

c. Identification of the authors of the Data/Software products

To unambiguously assign credit it is highly recommended that authors use a Unique Identifier, such as their ORCID/Scopus/Mendeley profile ID. Next to this, it is advisable that authors include a unique identifier of the grant number which was used to collect and analyze the data included. This assumes such a grant ID is unique and readily accessible. Specifically, we are interested in matching up our unique author IDs with those in the funder's information systems, so we can support institutions and individual researchers in developing reporting systems that correctly identify individuals. To that end, it would be useful to be able to have access to the NIH's systems of identification of individuals, institutions and departments. For software citations, we again concur with the Software Citation Principles: Credit and Attribution: Software citations should facilitate giving scholarly credit and normative, legal attribution to all contributors to the software, recognizing that a single style or mechanism of attribution may not be applicable to all software.

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

On Mendeley Data, citations currently point to either datasets or to data files. In future citations will be possible to collections of datasets. For articles, editors and reviewers are rejecting articles that don't contain sufficient novelty; maybe there should be some sort of responsibility for repositories on how to aggregate data. NIH is doing it in one of the most important dataset ever collected: www.cdc.gov/nchs/nhanes/nhanes_citation.htm For software citations, we again concur with the Software Citation Principles on Specificity: "Software citations should facilitate identification of, and access to, the specific version of software that was used. Software identification should be as specific as necessary, such as using version numbers, revision numbers, or variants such as platforms."

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

In accordance with the DCIP report mentioned earlier, we support the unambiguous identification and creation of a Landing Page containing a PID for each dataset. We have contributed to and are in support of the example set by the Force11 Resource Identifier Initiative, https://www.force11.org/group/resource-identification-initiative, to provide an unambiguous identifier to any electronic resource utilized in a research report. The Scholix project, mentioned above, also supports the creation of Linked Data Systems to enable unambiguous data citation and identification.

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

We would like to suggest that that researchers should not only be encouraged to document and report on how they share data and software but also how they use and contribute to existing data sets and software outputs. This would encourage community cooperation around common data sets and software, and reward the creators of the original data and software. Proper authorship for data and software allows attribution and credit to both the author and their institution. Tools such as Scopus can provide metrics and analytics around high quality scientific output of any form, software and data as well as articles and books. Using metrics and quality assessments based on Scopus data, especially if it would include data, can provide a valuable tool to give credit to researchers and evaluate the impact of their output.

4. Any other relevant issues respondents recognize as important for NIH to consider

Elsevier is eager and enthusiastic to remain involved in the next stages of discussion on this important topic, and are looking forward to continuing and expanding our current engagement and collaboration related to research data with the NIH. In addition to the current response, Elsevier has submitted responses to all research data-related NIH RFIs in the last two years, including: • 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) --> Refer to Comment 5 only • NOT-AI-15-045, Input on NIAID Data Sharing Repository, Immunology Database and Analysis Portal (ImmPort), and Services • NOT-ES-15-011, Input on Sustaining Biomedical Data Repositories • NOT-OD-16-133, Metrics to Assess Value of Biomedical Digital Repositories Copies of these other related RFI responses are appended here as an attachment for reference.

Additional Comments

Elsevier Previous Research Data RFI Responses.pdf (781 KB)

Request for Information (RFI): Request for Information (RFI): Soliciting Input into the Deliberations of the Advisory Committee to the NIH Director (ACD) Working Group on the National Library of Medicine (NLM)

Thank You - Your Comments Have Been Received. You may want to print this page with your comments for your records.

03/09/2015 at 06:04:49:363 PM

Name:

Holly J Falk-Krzesinski, PhD

Email Address:

h.falk-krzesinski@elsevier.com

Name of Organization:

Elsevier

City and State:

New York, NY, USA

Comment 1:

Current NLM elements that are of the most, or least, value to the research community (including biomedical, clinical, behavioral, health services, public health, and historical researchers) and future capabilities that will be needed to support evolving scientific and technological activities and needs.

Responses in Comment 4 and Comment 5

Comment 2:

Current NLM elements that are of the most, or least, value to health professionals (e.g., those working in health care, emergency response, toxicology, environmental health, and public health) and future capabilities that will be needed to enable health professionals to integrate data and knowledge from biomedical research into effective practice.

Responses in Comment 4 and Comment 5

Comment 3:

Current NLM elements that are of most, or least, value to patients and the public (including students, teachers, and the media) and future capabilities that will be needed to ensure a trusted source for rapid dissemination of health knowledge into the public domain.

Responses in Comment 4 and Comment 5

Comment 4: 141

Current NLM elements that are of most, or least, value to other libraries, publishers, organizations, companies, and individuals who use NLM data, software tools, and systems in developing and providing value-added or complementary services and products and future capabilities that would facilitate the development of products and services that make use of NLM resources.

Elsevier values its multi-faceted and synergistic relationship with the National Library of Medicine (NLM) and is appreciative for the opportunity to provide a response to NOT-OD-15-067, a Request for Information (RFI) Soliciting Input into the Deliberations of the Advisory Committee to the NIH Director (ACD) Working Group on the NLM. TAXONOMIES/THESAURI/DATABASES: UMLS provides a wide range of medical vocabularies. These by themselves are valuable for determining names of medical concepts and alternative names for the same concepts. More importantly, UMLS maps equivalent notions from different vocabularies. Those notions are classified into a reasonable number of semantic groups, which is helpful for us as Elsevier processes our content and looks for relations between things such as classes of drugs and types of diseases. The UMLS browser is helpful for quick lookups of vocabulary and relation data. NLM also provides tagging tools like MetaMap, useful in work on recognizing medial entity mentions. Elsevier's EMMeT Taxonomy uses UMLS as the primary source for the taxonomy. ClinicalKey licenses the PubMed taxonomy and proposes its content in the ClinicalKey suite of products. GoldStandard sends its drug data to RxNorm to get it coded. These three resources are very important contributors to our product offerings. In terms of vocabularies representation and alignment, MeSH and MedDRA are critical resources for our projects. What would be useful in the future would be a "graph of biomedical data" linking biomedical data across MeSH and MedDRA (and ideally all of UMLS) using Linked Data formats. The current work on representing MeSH in RDF is a very exciting step, but a SKOS/SKOS-XL representation would also have a lot of value and would make the integration with our own datasets easier. Elsevier is also interested in the multi-lingual aspect of some UMLS vocabularies, for building cross-language bridges; here again, MeSH and MedDRA are key. Our Natural Language Processing group is a user of both the MeSH thesaurus and its supplementals (mostly drugs and chemical compounds) and of UMLS. We follow the annual update cycle and are quite satisfied in doing so. MeSH is the de facto standard for general Life Sciences /Medical concept annotation. There are domain specific ontologies / thesauri but none beats MeSH in the 'general' area. Recently NLM took the initiative to put out MeSH in RDF format, to connect it to the world as 'linked data', starting from concepts are unique URI identifiers. This is an important initiative that the Elsevier Labs group is glad to be part of, particularly if this project continues to evolve so that the data is truly linked to other resources (PubMed at least) and easily accessible. Our Corporate R&D business unit utilizes and incorporates several NLM elements such as data streams of raw data, bibliographic information, and taxonomies, into several different products. The NLM resources are invaluable as they contain high quality standardized information which we leveraged together with Elsevier content to advance biomedical and health related science. The availability of this high quality and standardized information for researchers is extremely important, and should continue to be an important part of NLM's efforts. As technologies and platforms evolve, the demand for high throughput data retrieval and analysis workflows continues to increase, so it will be beneficial for researchers/corporations if NLM continues to develop its access mechanisms for NLM elements to meet this demand. Specifically, our products Text Mining and Pathway Studio leverage bibliographic, text, and taxonomic/vocabulary data, among other NLM elements. As we merge many data elements together for comprehensive solutions for our customers, we have identified some areas we hope NLM will consider for future development: 1) Convene stakeholder groups in standardizing structures of other biomedical research and health data elements. Similar to the development of the NISO JATS XML standard, NLM could work with stakeholder partners towards either extending this standard or developing new standards such that other data types/formats could also be captured and delivered in a standardized way e.g. electronic health records; 2) While NLM has been involved with ORCID and other unique author identifiers, which are gaining wider use, it would be good if the public could have a better understanding of how these elements are intended to be disseminated, i.e. as part of which data fields in particular record types; and, 3) Further map *and* provide mappings between taxonomies, e.g., UMLS–RxNorm-SNOMED are all very integrated but mapping files between them are complex and somewhat difficult to discern. A potential solution could be API for mapping translation. NLM seems to be keen to improve their services to the community, which we applaud. We'd be interested in a number of developments in this regard: 1) Linking MeSH to other resources that are in the linked-data sphere; provide equivalences (exactMatch, partMatch, etc.) between MeSH concepts and concepts in other taxonomies that are linked-data-enabled, such as NAL, DBPedia etc. 2) Make all NLM vocabularies available by API on a day-to-day basis. Getting access to MeSH is currently non-trivial and cumbersome. Elsevier would appreciate having a query API that allows us to receive updates on at least a weekly basis. PUBMED/MEDLINE: From the traffic we receive from PubMed to our Health & Life Sciences Content on ScienceDirect we can see how important it is as a discovery tool for researchers in these disciplines. We appreciate how our content is indexed for MEDLINE, especially the assignment of MeSH terms and making these terms available to other search and discovery services. This has a great contribution to the discoverability and dissemination of the content Elsevier publishes. Our general analytical services reporting (commercial and extensive pro bono activities) also benefits from PubMed/MEDLINE through Elsevier's Scopus because of the well-defined/assigned PMIDs. The PMID-DOI converter and API are especially useful. PUBMED CENTRAL/PUBLIC ACCESS POLICY: Elsevier welcomes the opportunity to enhance delivery of public access through collaboration and interoperability with NLM to avoid duplication and wasted resources. There are opportunities for NLM to collaborate more effectively with publishers in the context of PubMed Central (PMC) to avoid duplication of effort and cost and to minimize administrative costs to research institutions and burden to researchers. One of the significant collaboration opportunities in facilitating public access is via the CHORUS service (http://www.chorusaccess.org/). At Elsevier, we are concerned that the NIH is the only US federal funding agency that has not met directly with representatives of the CHORUS service, and has not considered how this new approach presents opportunities for cost-savings within the NIH budget and for institutions receiving NIH research support. NLM should actively seek opportunities to work with publishers, including integration with CHORUS, to develop and implement open access publication options that leverage existing

infrastructure, tools, and services that support sharing, access, discoverability, reporting, and preservation. It is also important for NLM to recognize that its public access policy's one-size-fits all 12-month embargo period is not suitable for all journals nor for all publishers, and to introduce a petition mechanism, as outlined in the OSTP memo, so publishers can signal these exceptional cases and provide supporting evidence. We would welcome greater sensitivity from NLM colleagues to more clearly distinguish approaches that are effective in the life and biomedical sciences from other disciplinary domains. Finally, while the NLM claims PMC to be a public-private partnership, in practice, the opportunities for collaboration with Elsevier and other publishers have been minimal. Collaboration is a recursive process that relies on continuous lines of open communication; with partners working together to develop and meet shared goals and involves shared governance and review procedures. Elsevier urges NLM to focus on engaging in more genuine collaboration around public access policy and policy implementation. Elsevier requests that NLM share COUNTER-compliant distributed usage statistics for manuscripts in PMC so that publishers can continue to report on impact and usage to authors and to their institutions that subscribe to these publications and pay their publication costs. It is also critical that NLM cease reformatting and enhancing manuscripts to make them appear more like, and substitute for, the final version-of-record of articles. Moreover, it is essential that PMC ensure readers are presented with the best version of the article available, which means that entitled users are transparently linked to the final published version. Finally we believe NLM needs to commit to taking concrete steps to prevent commercial re-use of manuscripts archived in PMC that is not authorized by the copyright holders of these works. PUBLISHING: As a health, medical, life, and biomedical sciences publisher and our involvement with the International Committee of Medical Journal Editors, Elsevier deeply values its collaboration with NLM in setting standards for journal articles and for developing and strengthening policies and practices in the field of publication ethics. NLM's leadership in publication standards makes it a unique participant in the national library space. In particular, we value NLM's commitment to PubMed and ClinicalTrials.gov. PubMed is the medical community standard reference point for article search and ClinicalTrials.gov is a vital mechanism for ensuring accountability, helping to deliver accurate published randomized trial reports and holding authors accountable not only for reporting standards but also for the timely release of their findings. There are areas we believe NLM can make further strides. We feel strongly that NLM should adopt a more global role in fulfilling its mission and responsibilities, with these specific recommendations: 1) Invest in advocacy and infrastructure to advance sustainable platforms for information access in low and middle income country settings to support the health dimensions of the Sustainable Development Goals, e.g., in library services, human resources, national leadership, in partnership with country health sectors; 2) Work to assist countries in developing their capacities for research information generation, publication, and implementation; 3) Partner with journals and publishers to advance these global goals; and 4) Make global equity in information access core to NLM's mission. Elsevier is a proud partner and promoter of the NLM's Emergency Access Initiative (EAI), through which we provide free access to our primary online clinical information and reference tool, ClinicalKey, and to a corpus of relevant literature on our ScienceDirect platform. As a member of the NLM-Publisher Panel, Elsevier is pleased to have a forum to discuss issues of common interest. Topics discussed at recent meetings include the 'Article of the Future' initiative; the MEDLINE submission and review process, including the Literature Selection Technical Review Committee; Emergency Access Initiative; improving access to publisher full-text content; and reproducibility and rigor of research findings. The Panel has provided essential collaboration on these and other initiatives. The Panel can continue to increase its usefulness by addressing additional matters of common interest, for example: 1) Increasing the acceptance rates of evaluated journals and book serials, which would lead to additional high-quality content being available via PubMed; 2) Indexing book content beyond serials as books offer a unique view into biomedical and health related information that is not mirrored in journals, providing an integration of research across time and subject areas, consolidating disparate literatures into one source, and synthesizing research advances and applications; 3) Linking of all information relating to clinical trials, including all articles published as a result of a trial; and, 4) Sharing of clinical trial data, including protocols for how to cite data, where to store data, and how to share data. Elsevier looks forward to our continued participation in the Panel and collaboration with the NLM and other publisher representatives.

142

Comment 5:

How NLM could be better positioned to help address the broader and growing challenges associated with:

- Biomedical informatics, "big data", and data science;
- Electronic health records;
- Digital publications; or
- Other emerging challenges/elements warranting special consideration.

RESEARCH DATA: Elsevier would like to see the NLM allow mining of all database content inside the suite of databases managed and curated by the NLM and provide actionable copyright metadata elements on all NLM content so we understand what we can mine/use for commercial and non-commercial purposes. Elsevier's research data policy (http://www.elsevier.com/about/research-data) commits us to encouraging and supporting researchers to making their research data freely available with minimal reuse restrictions wherever possible. Alongside our policy, we have developed a range of tools and services to support researchers to store, share, access, and preserve research data. These include our open data pilot, our database linking program, and our data journals, such as Genomics Data and Data in Brief. Collectively, Elsevier as partners with NLM, we should to be thinking about the big picture goal of enabling researchers to properly collect and annotate their research data in ways that lead to archiving, auditing, reproducibility, and interoperability. This might include making vocabularies and other data models available in the researchers' workflow (e.g., controlled vocabularies and

drop-downs in Electronic Lab Notebooks). This is especially for vocabularies, databases, and other data models that identify entities that define research data (anatomy, diseases, organisms, etc.). Making this available in formats that foster interoperability is a big part of this. This way, unique identifiers and codes are captured early on and can stay with the research data through its entire lifecycle (whether or not research ends up getting published). Research data adds huge value to the users of published research articles. An important focus is twofold: 1) Attach and make available publicly the methods and data underlying published research; and, 2) Develop standard markups (XML) to allow machine interpretation of the data (this is an area that Elsevier's Mendeley team is currently working on). It will be important for NLM to work in close partnership with a broad stakeholder group to consider the most effective approach to enforcing data transparency and developing a set of markup standards. Data fraud detection tools will need to be an important focal point for NLM. In recent scientific fraud causes, fraud was detected as data that was statistically, "too good to be true." Similarly, image manipulation for scientific articles has been observed and is being addressed by a number of publishers at high cost due to the manual labor involved. To avoid future problems and resulting distrust in our data-drive scientific approaches, NLM and publishers will need to work together to find efficient and effective ways to detect data fraud before data sharing and publication. Regarding research data repositories, we think it is most useful to think in terms of data management plans and data archives. Elsevier is supportive of mandates for data management plans where researchers have the flexibility to choose where to deposit their data and that data publication routes are not limited (e.g., linking data, data journals, interactive data plots, etc.). Importantly, as efforts on research data repositories advance, it will be essential for the NLM to seek out collaboration opportunities with a broad and diverse range of stakeholders across sectors to ensure that collective expertise and experience as leveraged, a duplication of effort and resources are minimized, and cost savings and administrative efficiency are maximized. There is a need for data standards, but it should also be recognized that such standards do develop continuously. So any standardization proposal should include a proposal for continuous maintenance and further development of the standard. It should also be noted that data standards have to be discipline, perhaps even subdiscipline, specific, and will always have some element of least common denominator as science, by definition, goes beyond what has been standardized. Tools for automatic mapping of data would indeed be extremely useful as they can provide the input for data search engines. Furthermore, such tools can help scientists to better comply with funder requirements to share data in a meaningful way, especially when such tools are combined with proper (provenance) annotation capabilities. Elsevier would be very interested in working with the NLM, other publishers, and data archive managers on mechanisms to connect articles and related datasets. It would be valuable for publishers to link plug-ins into their systems, such that authors could submit the data to the archive of their choice and simultaneously link this to an article. We also feel that it is important that the NLM work with stakeholders on developing capabilities (at a variety of levels) to validate data and mark it as "OK" following a certain hierarchy of quality, from data has been well-described to data that has been fully reproduced in a different environment by a different team. Elsevier's data articles and microarticles do provide one of the steps in this continuum of quality/integrity validation, but there are additional levels beyond peer-review that need to be considered and built into developing systems. With regards to the quality criteria and quality stamps for data archives, there has been considerable discussion in this space, especially in the EU, but it is essential that there be commonly shared view on what a data repositories should adhere to, e.g., the National Digital Stewardship Alliance (NDSA) levels of preservation do make a step in one dimension of data repositories (archives), but there are many more dimensions to consider. NOTE: Elsevier is also developing a separate and detailed response to the NIH RFI NOT-ES-15-011, Input on Sustaining Biomedical Data Repositories, which we will submit by the deadline of March 18, 2015. ELECTRONIC HEALTH/MEDICAL RECORDS: Since HIPAA issues make it near impossible to obtain actual health records, a test/gold set of anonymized Electronic Health Records would be a great resource to Elsevier to develop and test point of care applications we are currently developing. Also, a test bed EHR/EMR system would be incredibly valuable, where different content providers could plug in applications to show added value of relevant data at the point of care. Elsevier's Health Analytics group is especially interested in developments with regards to EHR/EMR. We are supportive of: 1) Central, anonymized linked patient databases (including detailed clinical encounters in primary and secondary care, medication, genetic data, etc.) for research; and, 2) Central patient records, or at least interoperability standards (including federated search or HIEs) as a method of improving care delivery to individual patients. We encourage the NLM to continue working in coordination with the Office of the National Coordinator for Health Information Technology to drive both of these initiatives. We also want to make sure that NLM is aware of our high-performance computing (HPC) capabilities to analyze data for patterns. Reed Elsevier, Elsevier's parent company, is one of the very few companies in the world that has analytical HPC capabilities and is expert and experienced in dealing with highly confidential and very private data. Regarding the linked patient databases, more (diverse) and bigger (simply more) is better. Broad accessibility (under appropriate safeguards) to the anonymized, longitudinally linked for-research data, including by industry, is desirable for Elsevier's Health Analytics. Industry finances applied research and product development that brings universities' basic research to the point of care and to actually benefit patients. Broad accessibility will also drive innovation from big data, which is currently hindered by selective access. Heath Analytics currently conducts substantial research projects granting us securely anonymized patient data access together with healthcare systems in Europe. Regarding central patient records, comprehensive (all individual patient encounters) and timely is better. As an example, Denmark has introduced a shared medication record. Physicians there can see their colleagues' prescriptions. This transparency among providers is dramatically transforming the Danish healthcare system, already one of the best in the world. Physicians now feel responsible for the full array of prescriptions, even those of their colleagues. Also patients can access and review their complete personal health record, which makes them a responsible partner in their health management. The networking of all players improves patient outcomes substantially.

145

Part I: Elsevier's Response to Request for Information NOT-AI-15-045: Input on NIAID Data Sharing Repository, Immunology Database and Analysis Portal (ImmPort), and Services

For NOT-AI-15-045, areas of possible comment include but are not limited to:

1. **Best practices in maintaining public data sharing repositories.***
2. Innovative bioinformatics or data analysis tools or methods for research data visualization that are currently missing from or need to be improved upon in ImmPort.
3. Metadata analysis tools and methodology for extracting new information and knowledge from studies in public data repositories that are currently missing from or need to be improved upon in ImmPort.
4. Existing barriers that prevent maximum utilization of ImmPort including specific obstacles related to accessibility, readability, or usability of data from ImmPort or to the data submission process.
5. Outcomes from utilizing the ImmPort dataset and tools including, but not limited to: new collaborations, manuscripts, grant proposals, research proposals, research funding, and consultations.
6. Ability to use ImmPort in conjunction with other databases and analytical tools.
7. **Other emerging technologies or research initiatives that may impact the future development of ImmPort.***
8. **Data model and data repository infrastructure that support efficient data collection, curation, annotation, integration, and public sharing.***
9. **Data standards and transformation methods for integrating disparate datasets.***
10. Suggestions for improving ImmPort.

Responses below are provided for the **BOLDED areas above*

Elsevier is appreciative for the opportunity to provide a response to NOT-AI-15-045, a Request for Information NOT-AI-15-045: Input on NIAID Data Sharing Repository, Immunology Database and Analysis Portal (ImmPort), and Services. Our response is split into two parts (this is Part I) and were submitted by Holly Falk-Krzesinski, PhD, Vice President, Strategic Alliances, Global Academic Relations, on behalf of Elsevier, July 30, 2015

1. BEST PRACTICES IN MAINTAINING PUBLIC DATA SHARING REPOSITORIES

Regarding research data repositories, we think it is most useful to think in terms of data management plans and preferably discipline-specific data repositories. Elsevier is supportive of mandates for data management plans where researchers/authors have the flexibility to choose where to deposit their data and that data sharing routes are not limited (e.g., linking data, data journals, interactive data plots, etc.). We also recognize that deposit into repositories is not an end in itself, the goal of depositing data should be on enabling reuse, thus it is essential to focus on making repositories and the data therein readily discoverable, e.g., through linking. Importantly, as efforts on research data repositories advance, it will be essential for the NIH to seek out collaboration opportunities with a broad and diverse range of stakeholders across sectors to ensure that collective expertise and experience are leveraged, a duplication of effort and resources are minimized, quality and trustworthy data is separated from other types of data, data discoverability across multiple repositories is guaranteed, and cost savings and administrative efficiency are maximized.

The new NIH's Plan for Increasing Access to Scientific Publications and Digital Scientific Data from NIH Funded Scientific Research (the Plan) indicates that, "the NIH will expect funded researchers to deposit data in 'appropriate, existing, publicly accessible repositories before considering other means of making data available,' but where needed, NIH will take steps to support the development of 'selected community-based data repositories

and standards.' To help researchers find an appropriate repository to deposit their data, NIH will expand its database of existing repositories and plans to develop guidance and criteria to aid researchers in identifying 'acceptable repositories' not funded by NIH." While we are assuredly in favor of establishing authentication methods for data repositories we contend that researchers/authors must have the flexibility to choose where to deposit their research data into repositories as they are most knowledgeable about determining the repository best suited to their data and research. This principle should be at the center of any criteria NIH seeks to develop, and the NIH criteria should not inadvertently limit data publication routes, such as linking data, data journals, interactive data plots, etc.

Rigid repository-prescribing funder-specific mandates might lead to direct depositing of research data to a limited number of more generic repositories, running the risk of losing discipline- and domain-specific repositories that add significant value for data reuse and reproducibility. Similarly, mandates that require depositing to a single funder's repository will lead to fragmentation on the basis of country, which is counterproductive to the ever-expanding global nature of (biomedical) science and creation and use of (biomedical) research data by international teams of researchers working across sectors. Research data should be created in formats that allow deposition in a multitude of repositories, and published or deposited in any repository that best suits the research and the discipline. It is also important for the NIH not to put a policy in place that requires undue burden on researchers. It should take special care to ensure that NIH-supported investigators working in international collaborations don't find that they are required to meet multiple—and especially not disparate—funder data posting mandates.

The NIH needs to be a strong partner in defining data repository quality requirements and ensuring that repositories are validated. This would offer the NIH the opportunity for a more flexible policy that allows research data to be stored at repositories that meet specific the quality levels; more flexibility will facilitate compliance on the part of researchers and their institutions. Moreover, quality of repositories must also relate to unfettered access and linking abilities by multiple stakeholders. Recognizing that quality of data repositories is critical, Elsevier encourages the development of data repository certification standards building on initiatives like the Data Seal of Approval, an effort by several data repositories (working in partnership with other research data community stakeholder groups) to ensure sustainable and trusted data repositories. Data validation and data publishing are areas in which Elsevier has deep expertise that we can lend to this effort. Elsevier's data articles and microarticles (see below) are part of the continuum of quality/integrity validation, but there are additional levels beyond peer-review that need to be considered and built into developing research data systems and repositories.

One element that Elsevier is interested in working with the NIH on is defining the difference between data posting and data publishing. When researchers *post* a description of their research on the web, it is not validated by peers. When the text describing research is *published*, then others know that the associated research is peer-reviewed and validated, and thus can be trusted. It is important to make a similar distinction between *data posting* and *data publishing*: validating and quality stamping the data is becoming an ever more important element of a data-driven research community. Elsevier has developed a hierarchy of trust levels of data, where all of these issues are being addressed in a step-wise manner (see Figure 1 below). We also developed best-practice solutions for pushing data up in this hierarchy (like data journals, data profiles, data citations. and data linking), and are continuing to develop others (data repositories, data management, and data search). We are furthermore interested in

collaborating with NIH and others to increase data trust through development of methods to identify data fabrication and data falsification.

Figure 1: A hierarchy of research data needs. First, research data need to be stored and preserved, so that the data is saved for future use. Second, it needs to be accessible, discoverable and citable, so that other researchers can find and retrieve the data. Last, it needs to be comprehensible, reviewed, reproducible and reusable, so that it can be trusted and built upon.

Data fraud detection tools will need to be an important focal point for NIH as well. In recent scientific fraud causes, fraud was detected as data that was statistically, "too good to be true." Similarly, image manipulation for scientific articles has been observed and is being addressed by a number of publishers at high cost due to the manual labor involved. To avoid future problems and resulting distrust in our data-drive scientific approaches, NLM and publishers will need to work together to find efficient and effective ways to detect data fraud before data sharing and publication.

Elsevier's research data policy (http://www.elsevier.com/about/research-data) commits us to encouraging and supporting researchers to making their research data freely available with minimal reuse restrictions wherever possible. Alongside our policy, we have developed a range of best-practice tools and services to support researchers to store, share, access, and preserve research data. These include our Open Data and Data Profile pilots, our DataLink search tool and database linking program, and our data journals, such as *Genomics Data and Data in Brief*.

Collectively, the NIH should work with other stakeholders in thinking about the big picture goal of enabling researchers to properly collect and annotate their research data initially in ways that lead to archiving, auditing, reproducibility, and interoperability. This might include making vocabularies and other data models available in the researchers' workflow (e.g., controlled vocabularies and drop-downs in Electronic Lab Notebooks; preferred use of DOI's for data sets). This is especially for vocabularies, databases, and other data models that identify entities that define research data (anatomy, diseases, organisms, etc.). Making this available in formats that foster interoperability is a big part of this. This way, unique identifiers and codes are captured early on and can stay with the research data through its entire lifecycle (whether or not research ends up getting published).

Part I: Elsevier's Response to Request for Information NOT-AI-15-045: Input on NIAID Data Sharing Repository, Immunology Database and Analysis Portal (ImmPort), and Services

Innovation is central area in promoting use of research data and maintaining an open ecosystem while allowing for the creation of services that provide added value. Innovations can range from search services to aggregators and analytical tools. For example, the Open PHACTS project in Europe provides a developer friendly API that enables applications to build across public domain pharmacology data. Their service is supported by pharmaceutical companies through a foundation. Importantly, this service allows proprietary commercial data to sit alongside public data. Three lessons for the NIH arise from this example:

1) Innovation developments should ensure that it is possible to develop a range of services with different business models that store, access, and query various forms of research data. In providing an open model, both in funding and with respect to technological solutions, the NIH can create a flexible framework that allows academic and industry parties to develop components that optimally mesh together and enable systems that can change over time and are tailored to the needs of specific medical and scientific communities;
2) The NIH should seek to develop reporting mechanisms such that downstream aggregators and users can ensure that upstream, publicly funded data providers can receive credit; and,
3) While standardization is helpful for downstream data users, it is important to note that a flexible and open ecosystem can help manage complexity. Therefore, it is preferable to recommend vs. mandate data standards, and any mandates must have the flexibility to allow for change in capabilities and community practice over time.

Elsevier is very interested in supporting a system that evaluates the performance of various components of the biomedical Research Data Management cycle. We are currently actively engaged in a number of conversations with academic and industry partners to enable components to such a shared set of metrics, and systems to support them. We are interested in working in partnership with the NIH and other stakeholders on a workbench that enables quantitative evaluation of the usefulness and usability of different tools pertaining to research data storage, sharing, and search. Questions that one can ask of such a system could include:

- Which data standards, metadata systems, and curation efforts optimally improve outcome of a particular use case, such as data search, or data reuse?
- What metrics can be used for successful data storage or curation: reuse, amount of queries/downloads, or other—possibly social—metrics?
- What systems can act across the spectrum of biomedical repositories, publications, and other research outcomes to track and combine these metrics?

Finally, the NIH should seek opportunities to collaborate effectively with publishers to avoid duplication of effort and costs associated with research data sharing and to minimize administrative costs to research institutions and burden to researchers. By way of example, in conjunction with the Professional and Scholarly Publishing Division (PSP) of the Association of American Publishers (AAP), Elsevier has been involved with the CHORUS service; which leverages existing infrastructure, tools, and services across publishers that have committed to collaboration with federal funding agencies around the public access of research articles.

Part I: Elsevier's Response to Request for Information NOT-AI-15-045: Input on NIAID Data Sharing Repository, Immunology Database and Analysis Portal (ImmPort), and Services

7. OTHER EMERGING TECHNOLOGIES OR RESEARCH INITIATIVES THAT MAY IMPACT THE FUTURE DEVELOPMENT OF IMMPORT

Understanding that a recognition economy is the dominant environment in which academic and government researchers operate, it is essential to consider the drivers of research data sharing at the individual researcher level to maximize rapid and efficacious sharing. The NIH needs to address data sharing incentives and rewards for researchers in development of its policies and procedures. Relying only on the "stick" of mandated policy compliance, the full potential to stimulate and motivate broad sharing of research data will go unmet and will face challenges similar to those related to posting to PubMed Central and ClinicalTrials.gov. Elsevier encourages the NIH to review and operationalize the literature that provides an evidence base for understanding what drives researchers to be participatory data donors and we encourage the NIH to develop *new* research funding programs to extend empirical knowledge about this area of science policy. One approach might be for the NIH to partner with the NSF's Science of Science Innovation and Policy (SciSIP) program to develop a research data stream and funding resources to support new research grants in this area.

The free, public Mendeley Research Data Sharing group contains a rich library of such research data sharing resources. Contained therein, references describe the need to develop a reward and recognition system that affords researchers ongoing attribution, recognition, and professional reward for their sharing efforts. The literature also calls on policy makers, funders, and research organizations to consider the resources necessary for researchers and their institutions to comply with policy mandates, such as necessary skills, time & effort, and ongoing finances. Furthermore, the literature demonstrates the need for stakeholders to take into account the impact of sharing and potential for misuse on individual competitiveness, an essential consideration given the current hypercompetitive funding landscape.

8. DATA MODEL AND DATA REPOSITORY INFRASTRUCTURE THAT SUPPORT EFFICIENT DATA COLLECTION, CURATION, ANNOTATION, INTEGRATION, AND PUBLIC SHARING

Much of what was presented in Section 1 above is relevant here as well. For example, Elsevier's data articles and data linking program are proven parts of an effective larger data infrastructure.

In its new Plan for Increasing Access to Scientific Publications and Digital Scientific Data from NIH Funded Scientific Research (the Plan), it is very good to see that the, "NIH recognizes the benefit of collaborating with other federal agencies and public and private stakeholders to adopt consistent practices for citation of data sets across scientific communities and other data set attribution systems and will work toward this goal." And a broader context for this can also be found in the HSS Guiding Principles document, which talks about developing healthdata.gov as the basis for a "data commons approach across agencies," specifically the development of an internal HHS Enterprise Data Inventory that will serve as the internal catalog for all HHS data assets and be linked to healthdata.gov, the external-facing platform through which the public will be able locate and access federally funded research data. Next to Elsevier being co-creator of the Force11 Data Citation Principles, it has best-practice linking services that could add to this initiative by expanding the reach of healthdata.gov datasets.

The NIH's recent Plan also explains that "As part of the data discovery index, a system for unique identifiers for datasets generated by NIH-funded research will be developed, analogous to the PubMed Central identification number (PMCID) that is assigned to all submitted publications resulting from NIH-funded research. The identifier would also provide a means of linking the data with the biomedical literature via associated PubMed records." We would like to take this opportunity to share our thoughts around the NIH participating in development of an open, international standard identifier system built on DOIs.

Data DOI's are becoming a globally recognized standard for biomedical and other types of research data identification. Worthy of noting, a number of big data repositories, including the NIH Protein Data Bank (PDB), have assigned DOIs for all its accession numbers. DataCite, for example, has a valuable set of services connected with it offered at no cost and that make it easier to connect with other systems and DataCite has plans to expand its services to accommodate use cases that it currently cannot support (e.g., unpublished data that is early on in the lifecycle, and which is still subject to change). DataCite could be positioned to become a resolver for all other data accession numbers, which simplifies the entire research data infrastructure. The mapping of the Data DOI to an accession number is in the DataCite metadata, and so the DataCite API can be used to map accession numbers and then benefit from metadata for that record in DataCite. Other organizations are also focused on collaborative digital data standards development, including: APARSEN; Opportunities for Data Exchange (ODE); CoData; and, NISO/NFAIS Supplemental Journal Article Materials Project.

Elsevier recommends that NIH focus on the use of Data DOIs as the primary open, international identifier option for data that is published in any formal sense, rather than developing a identifier schema. And if the NIH is to develop a new accession number schema, then it must include assigned DOIs as well.

Elsevier further encourages the NIH to leverage the significant amount of work that has gone into developing common ways to *expose and cite* data. For example, the community effort of the FORCE11 Joint Data Citation Implementation Group has led to the creation of a standard for citing data within article publishing (the NISO JATS 1.1d2 XML schema). The Joint Data Citation Principles has been endorsed by over 90 institutions. The paper, "Achieving human and machine accessibility of cited data in scholarly publications," describes how to

operationalize those principles. As described in the Partnership section above, this effort further exemplifies the benefits of collaboration between major stakeholders in the scholarly communication ecosystem, focused on biomedical research and other types of research and data more broadly. By leveraging these community-driven efforts, a common basis for new models of sustainability will emerge.

Elsevier is an active partner with the Research Data Alliance (RDA) and ICSU World Data System (ICSU WDS). With such a wide range of stakeholders across for-profit and nonprofit sectors around the world, and an understanding that biomedical research data is a subset of research data more broadly, it is crucial for the NIH to be partner with these collaborative efforts so as not to duplicate work nor move in a direction specific only to research funded by the NIH.

The basis for Elsevier's involvement in partnerships is that we recognize that creating a research data infrastructure (including the technical infrastructure but also policies, best practices, standards, etc.) has to be a collaborative, cross-stakeholder and international effort where all the different players work together. Elsevier is proud to contribute our deep expertise and perspective from our position as a world leader in research information and appreciate having a voice in development of a synergistic and interoperable emerging research data infrastructure.

The RDA is a great forum for such an approach, as it brings together thought leaders in research data from various stakeholder groups (data centers, research institutes, libraries, publishers, funders, interest group, etc.) and individuals working in the research data field with different expertise and focus, all the way from deep technical expertise to policy-making. The primary value of the RDA is that it has become the forum where stakeholder groups come together to interact and work on issues and focus on making realistic progress on a swift timescale (e.g., 18 mos is the typical lifespan of an RDA working group).

Specifically, Elsevier is involved in a number of working groups under the "Data Publication" umbrella Interest Group (IG) of the Research Data Alliance (RDA) and encourages NIH to join in the partnership. All of these working groups began as ICSU WSD working groups and now have dual ICSU WDS/RDA mandate:

- Data Publication Bibliometrics
- Publishing Data Cost Recovery for Data Centres (for more details, see previous paragraph)
- Data Publication Services

The joint RDA/ ICSU World Data System Publishing Data Cost Recovery for Data Centres scope aligns with this RFI. Co-chair Anita de Waard of Elsevier and her colleagues recently interviewed 22 data centers about their ideas around cost recovery methods, now and in the future. In summary, Elsevier supports the collaborative efforts of the joint RDA/ICSU WDS Interest Group (IG) to elucidate the full cost of data management throughout its lifecycle–from inception through publication to storage and curation—by engaging funders, researchers, repositories, and other stakeholders in the research data management lifecycle. Specifically, the IG finds that data repositories are looking for new funding mechanisms – including charging deposit fees, access fees, and working through public-private partnerships—but are having trouble finding the time and resources to actively explore these new models. Elsevier is very interested in supporting further work regarding these questions, whether within the scope of the RDA or in direct collaboration with the repositories and/or the NIH.

Part I: Elsevier's Response to Request for Information NOT-AI-15-045: Input on NIAID Data Sharing Repository, Immunology Database and Analysis Portal (ImmPort), and Services

The NIH should strive to work in partnership with other stakeholder groups to develop consistent preservation criteria. To do so, it will be important to address some key questions, such as: Should all versions of data be preserved? Should research data be overwritten with newer data? For how long should data be preserved? Is indefinite preservation sustainable?

9. DATA STANDARDS AND TRANSFORMATION METHODS FOR INTEGRATING DISPARATE DATASETS

Research data adds huge value to the users of published research articles. An important focus is twofold: 1) Attach and make available publicly the methods and data underlying published research; and, 2) Develop standard markups (XML) to allow machine interpretation of the data (this is an area that Elsevier's Mendeley team is currently working on). It will be important for NIH to work in close partnership with a broad stakeholder group to consider the most effective approach to enforcing data transparency and developing a set of markup standards.

There is a need for data standards, but it should also be recognized that such standards do develop continuously. So any standardization proposal should include a proposal for continuous maintenance and further development of the standard. It should also be noted that data standards have to be discipline, perhaps even subdiscipline, specific, and will always have some element of least common denominator as science, by definition, goes beyond what has been standardized.

Tools for automatic mapping of data would indeed be extremely useful as they can provide the input for data search engines. Furthermore, such tools can help scientists to better comply with funder requirements to share data in a meaningful way, especially when such tools are combined with proper (provenance) annotation capabilities.

Elsevier would be very interested in working with the NIH, other publishers, and data archive managers on mechanisms to connect articles and related datasets. It would be valuable for publishers to link plug-ins into their systems, such that authors could submit the data to the archive of their choice and simultaneously link this to an article.

We also feel that it is important that the NIH work with stakeholders on developing capabilities (at a variety of levels) to validate data and mark it as "OK" following a certain hierarchy of quality, from data has been well-described to data that has been fully reproduced in a different environment by a different team. Elsevier's data articles and microarticles do provide one of the steps in this continuum of quality/integrity validation, but there are additional levels beyond peer-review that need to be considered and built into developing systems.

With regards to the quality criteria and quality stamps for data archives, there has been considerable discussion in this space, especially in the EU, but it is essential that there be commonly shared view on what a data repositories should adhere to, e.g., the National Digital Stewardship Alliance (NDSA) levels of preservation do make a step in one dimension of data repositories (archives), but there are many more dimensions to consider.

UMLS provides a wide range of medical vocabularies. These by themselves are valuable for determining names of medical concepts and alternative names for the same concepts. More importantly, UMLS maps equivalent notions from different vocabularies. Those notions are classified into a reasonable number of semantic groups, which is helpful for us at Elsevier processes our content and looks for relations between things such as classes of drugs and types of diseases. The UMLS browser is helpful for quick lookups of vocabulary and relation data. NLM also provides tagging tools like MetaMap, useful in work on recognizing medial entity mentions. Elsevier's EMMeT Taxonomy uses UMLS as the primary source for the taxonomy. ClinicalKey licenses the PubMed taxonomy and proposes its content in the ClinicalKey suite of products. GoldStandard sends its drug data to RxNorm to get it coded. These three resources are very important contributors to our product offerings.

Part I: Elsevier's Response to Request for Information NOT-AI-15-045: Input on NIAID Data Sharing Repository, Immunology Database and Analysis Portal (ImmPort), and Services

In terms of vocabularies representation and alignment, MeSH and MedDRA are critical resources for our projects. What would be useful in the future would be a "graph of biomedical data" linking biomedical data across MeSH and MedDRA (and ideally all of UMLS) using Linked Data formats. The current work on representing MeSH in RDF is a very exciting step, but a SKOS/SKOS-XL representation would also have a lot of value and would make the integration with our own datasets easier. Elsevier is also interested in the multi-lingual aspect of some UMLS vocabularies, for building cross-language bridges; here again, MeSH and MedDRA are key.

Part II: Elsevier's Response to Request for Information NOT-AI-15-045: Input on NIAID Data Sharing Repository, Immunology Database and Analysis Portal (ImmPort), and Services

For NOT-AI-15-045, areas of possible comment include but are not limited to:

1. Best practices in maintaining public data sharing repositories.
2. Innovative bioinformatics or data analysis tools or methods for research data visualization that are currently missing from or need to be improved upon in ImmPort.
3. **Metadata analysis tools and methodology for extracting new information and knowledge from studies in public data repositories that are currently missing from or need to be improved upon in ImmPort.**
4. Existing barriers that prevent maximum utilization of ImmPort including specific obstacles related to accessibility, readability, or usability of data from ImmPort or to the data submission process.
5. Outcomes from utilizing the ImmPort dataset and tools including, but not limited to: new collaborations, manuscripts, grant proposals, research proposals, research funding, and consultations.
6. Ability to use ImmPort in conjunction with other databases and analytical tools.
7. Other emerging technologies or research initiatives that may impact the future development of ImmPort.
8. Data model and data repository infrastructure that support efficient data collection, curation, annotation, integration, and public sharing.
9. Data standards and transformation methods for integrating disparate datasets.
10. Suggestions for improving ImmPort.

Responses below are provided for the BOLDED areas above

Elsevier is appreciative for the opportunity to provide a response to NOT-AI-15-045, a Request for Information NOT-AI-15-045: Input on NIAID Data Sharing Repository, Immunology Database and Analysis Portal (ImmPort), and Services. Our response is split into two parts (this is Part II) and were submitted by Holly Falk-Krzesinski, PhD, Vice President, Strategic Alliances, Global Academic Relations, on behalf of Elsevier, July 30, 2015

3. METADATA ANALYSIS TOOLS AND METHODOLOGY FOR EXTRACTING NEW INFORMATION AND KNOWLEDGE FROM STUDIES IN PUBLIC DATA REPOSITORIES

Elsevier has a long track record of data and metadata standards, dating back to the 1990s when we led the TULIP project. The Elsevier XML specifications for journal articles and book chapters are widely known and in use for 3000+ propriety and society journals and the metadata for 20,000+ journals. Content, including 12M journal articles, resides in a content repository that is accessible through restful APIs. Its metadata model is described using RDF serialized as JSON-LD. The API payloads and responses in JSON-LD are treated in the same way as our main content standards.

Our content is stored in multiple content-type-specific "warehouses." Through a metadata repository, this is made in to a virtual whole, called our Virtual Total Warehouse. Our content model and metadata standards are especially focused on content versioning. "Generations" of content assets keep various files together that together constitute a version. This Virtual Total Warehouse (VTW) plays a role in acquisition, editing and curating content (in our case, journal articles, book chapters, drug monographs, patents, patient education, and much more) and a Content Enrichment Framework takes this content and can, in principle, run any semantic process on the content, depositing the results back in VTW.

Elsevier also has a linked data repository adhering to the standards of linked data and linked open data.

Part II: Elsevier's Response to Request for Information NOT-AI-15-045: Input on NIAID Data Sharing Repository, Immunology Database and Analysis Portal (ImmPort), and Services

Elsevier's approach to unstructured information: The vast majority of information exists as an unstructured text which makes it unsuitable for efficient analysis by humans. The area of computational assistance to analysis of large volumes of textual information is traditionally split into two (somewhat overlapping) approaches - information retrieval and information extraction.

Information Retrieval (IR) systems concentrate on finding documents containing information deemed relevant to a particular topic of interest. Usually this is done by analyzing the word content of the documents using statistical methods based on keywords or word co-occurrence. IR methods are by their nature generic and to a large degree language-independent; the output of IR systems is *intended for human readers*.

Unlike IR, Information Extraction (IE) focuses on extracting information contained within the documents in a form *suitable for automatic processing*. IE systems use an *ontology* (or knowledge representation schema) as a model of a particular domain, and thus are domain-specific. The simplest form of an ontology is a list (or, even better, a hierarchical tree) of concepts relevant to the domain. More advanced forms of ontology also specify possible semantic types of relationships between the concepts. Extracting information with high precision involves deep understanding of the actual meaning of the text; as a result, IE systems are language-specific.

In developing solutions for vertical markets, Elsevier takes the IE road. Instead of building one generic, language- and domain-independent system that deals with large number of topics but provides little depth when it comes to the subject matter, we focus on extracting structured information specific for a particular domain from English text.

Elsevier's Information Extraction (IE) technology: Elsevier Text Mining

Within its Elsevier Text Mining portfolio, Elsevier has developed a proprietary natural language processing (NLP)-based technology called MedScan for extraction of structured information from unstructured text. It is a good fit for automatic indexing of NIH's content as the MedScan Thesaurus/Taxonomy was built mostly based on NIH thesauri and has all the NIH identifiers integrated (MeSH Headings, NCI Metathesaurus IDs, Entrez Gene IDs, Organism Tax IDs, etc.). The technology works by first recognizing domain-specific named entities (concepts) in the input text, and then uses natural language processing techniques to extract *attributed, directional semantic relationships* between them. The relationships can be of any complexity from simplest binary (X affects Y) to n-ary (X protects Y from Z) and complex multi-level nested ones (effect of X on Y depends on Z).

Elsevier IE technology has modular architecture. Each module performs its specific function and has well-defined and documented input/output format. Modules with compatible interfaces can be combined into different text processing pipelines, as required by the application. All modules are written from scratch to achieve our flexibility/precision/performance goals. The modules are portable C/C++ applications interacting via files and pipes.

MedScan Technical Description: MedScan is a proprietary natural language processing (NLP)-based technology for extraction of structured information from unstructured text. Structured information is captured and formally represented using a conceptual model (ontology) of the domain. The ontology consists of a set of conceptual named entities (e.g. Proteins, Small molecules, Cellular processes, Diseases, etc) and a set of categorized relationships (Binding, Protein Modification, Expression regulation, Molecular Transport, etc) between them.

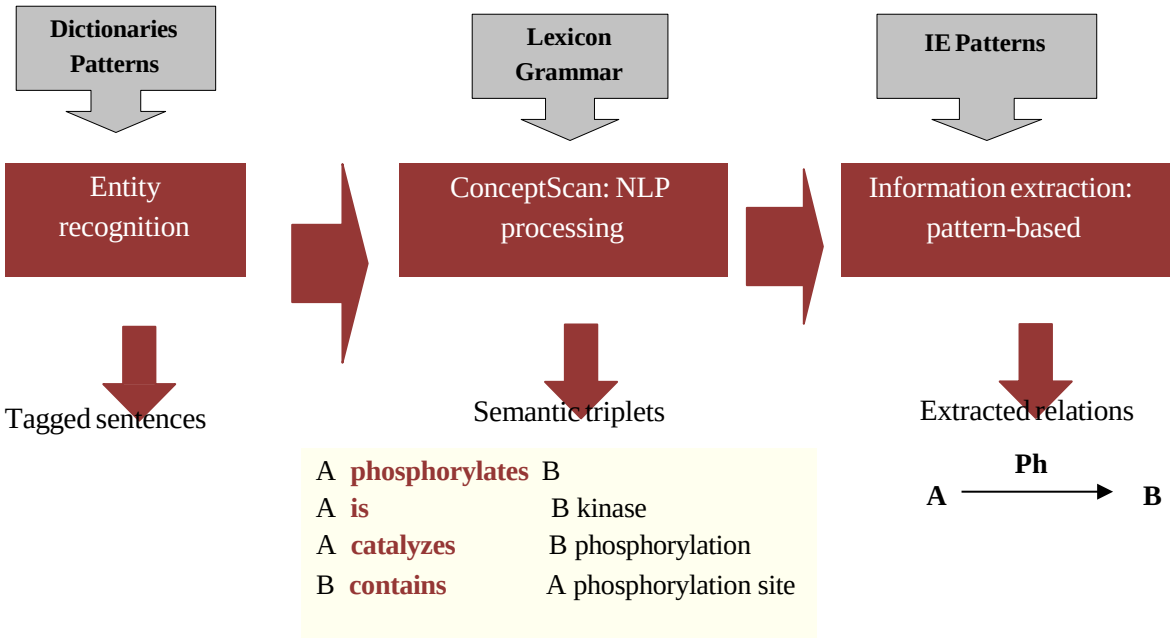


Figure 2. An overview of MedScan Architecture

MedScan first recognizes different domain-specific named entities (gene/protein names, cellular processes, cellular components, diseases, tissues, organs, etc.) in the input text, and then extracts functional relations (binding, regulation, association, molecular transport, etc.) between them. Figure 2 shows an overview of MedScan architecture.

The Entity Recognizer module utilizes hand-crafted dictionaries of domain-specific entities in combination with an advanced matching algorithm to detect them in input text.

To extract entity relationships from the text, MedScan utilizes two modules. The natural language processing module, ConceptScan, analyzes the sentence structure and decomposes each sentence into a deterministic set of Subject-Verb-Object triplets, each representing a single semantic relationship between two singular noun phrases. Next, Pattern Matcher matches carefully designed linguistic patterns over the triplets to extract and encode the entity relationships.

MedScan has been field-tested and is proven to be fast, efficient, and accurate information extraction technology. It is currently used to process the content of the entire Medline database along with more than 40 freely available full-text journals in order to extract more than 3.5 million individual facts (relations) about functions of proteins

with an overall accuracy of 90% and recall of 70%. The entire processing cycle can be completed in less than 24 hours on a regular PC.

Dictionaries and Named Entity Recognition:

Entity type	Number	Main sources
Proteins	136,000	Entrez Gene
Prot. Classes	7,500	GO, Enzymes, PubMed
Cell components	740	GO, PubMed
Cell processes	5,200	GO, PubMed
Diseases	6,300	MESH, PubMed
Small Molecules	270,000	MESH, PubChem, PubMed
Tissues	100	MESH, UMLS, NCI, EVoc
Cell types	360	MESH, UMLS, NCI, EVoc
Organs	2,875	MESH, UMLS, NCI, EVoc
Clinical parameters	1,786	Pubmed, ClinicalTrials.gov
Cell lines	2,500	PubMed

Table 1. MedScan Dictionaries

The Entity recognition module of MedScan utilizes hand-curated dictionaries of biomedical entities to detect them in the input text. Dictionaries are manually compiled and curated from the number of various public-domain resources (EntrezGene and SwissProt for protein names, PubChem and MESH for small molecules, GO for cell processes and components, MESH for diseases, NCI thesaurus for organs, tissues and cells, etc). Whenever possible the entities are hyperlinked to those outside resources for reference. Many additional aliases and terms are also added directly from the literature resources, e.g. PubMed. Table 1 shows the content of MedScan dictionaries. MedScan uses number of different algorithms to achieve accurate detection of entities in text. It can also use rule- and regular expression- based approaches to detect specific types of entities (abbreviations, numbers, dates, etc). The dictionaries are in a simple tab-delimited format so they can be easily extended or modified.

The input text can be in various formats (plain text, Microsoft Office, HTML, reasonable forms of PDF, zip/tar/gzip archives of the above, etc.) The output of the entity recognition step consists of individual sentences labeled to preserve their origin with identified named entities marked up with entity IDs, using **ID{***number***=….}** format (shown in red):

15986412:5 Enzyme assay, Western blot and ID{4000000,4106278=reverse-transcription} polymerase chain reaction (RT-PCR) results demonstrated that protein and mRNA expressions of human simple **ID{445329=phenol sulfotransferase}** (**ID{6799=P-PST}**), human ID{6818=monoamine sulfotransferase} (ID{6818=M-PST}), human ID{6822=dehydroepiandrosterone sulfotransferase} (ID{6822=DHEA-ST}) and human ID{6783=estrogen sulfotransferase} (ID{6783=EST}) were induced in ID{10000000,11012376=Hep G2

cells}; ID{6818=M-PST} and ID{6822=DHEA-ST} were induced in ID{10000000,11010382=Caco-2 cells}. The type of entity is encoded in its numerical range.

Natural Language Processing: The central idea of Elsevier's NLP algorithm (called ConceptScan) is decomposing natural language sentences into semantic relationships (which we will also call semantic triplets). Each triplet is designed to represent a single semantic relationship between two singular noun phrases (NPs). An example below illustrates this paradigm using a complex artificially constructed sentence.

11940574:7 Because **Axin2** has been shown to associate with and inhibit **beta-catenin** abundance and function, we hypothesized that **Axin2**, which is affecting proliferation of MEF cells can work in a negative feedback pathway, regulating **Wnt** signaling and thus controlling apoptotic process.

Triplets:

Axin2	associate	beta-catenin abundance
Axin2	inhibit	beta-catenin function
Axin2	associate	beta-catenin abundance
Axin2	inhibit	beta-catenin function
Axin2	affect	MEF cell line proliferation
Axin2	work	negative feedback pathway
Axin2	regulate	Wnt signaling
Axin2	control	apoptotic process

The extracted triplets capture the main facts expressed in a sentence. The ConceptScan is used in conjunction with named entity detection algorithm to index relationships between biomedical entities and to extract entity relationships.

ConceptScan parses sentences in several sequential algorithmic steps (See figure below)

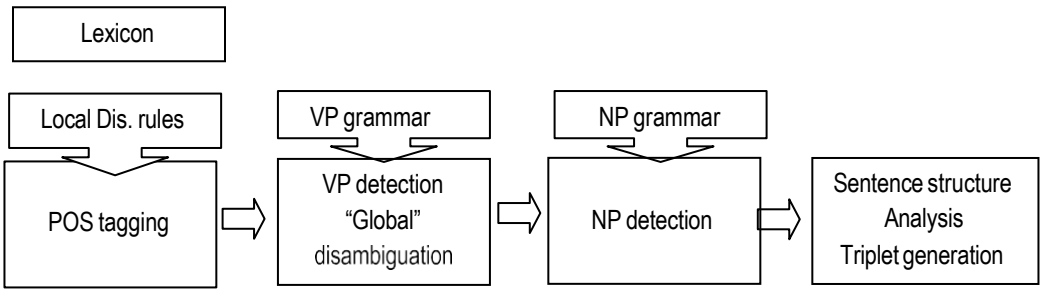


Figure 3. ConceptScan algorithm

The first step of NLP is part-of-speech tagging and local disambiguation. During this step, the words in a sentence are reduced to all possible uninflected forms, looked up in the lexicon and annotated with the respective syntactic categories. After initial POS tagging, the local disambiguation algorithm, encoded by a set of contextual regular expression-like rules, is applied. Notably, not all ambiguities can be resolved locally. The unresolved ambiguities are preserved for subsequent processing steps. The next step is identification of verbal phrases. Verbal phrase (VP) grammar is encoded in a single but complex deterministic finite-state automaton (DFA), with more than 25,000 states. It is matched over the sequence of syntactic categories assigned to sentence words at the POS-

tagging step. NP grammar is matched after detection of verbal phrases is complete. Similarly to VP grammar, it is encoded by a DFA. The structure of NP grammar covers prepositional attachment, conjunctions, relational constructs, appositions and exemplifications. Once VPs and NPs have been identified, ConceptScan analyzes the structure of the entire sentence.

Information extraction: The specific relationships between entities are extracted using separate module - Pattern Matcher. It utilizes a formalism closely resembling regular expressions to detect specific linguistic constructs expressing entity relations and to capture the expressed relations. It is specifically tailored to deal with linguistic input; it operates on the level of individual words rather than symbols and supports advanced linguistic features like matching all word forms and multi-word lexemes. Pattern matching also supports all regular expression features: wildcards, sets, negation, etc. The figure below shows a sample information extraction pattern.

```
CONTROL
{
  ControlType = "ProtModification"
  in  = %Protein1(Protein)
  out = %Protein2(Protein)
}
:
%Protein1 $MODAL? $ADV* phosphorylate~ %Protein2         |
%Protein2 $MODAL? $BE $ADV* phosphorylated by %Protein1 |
Phosphorylation of %Protein2 by %Protein1                |
;
```

Figure 4. An example of the information extraction pattern. The head template encodes the name of the output frame and templates for the values of its slots, which can be literals or other frames. Named entity variables (%Protein1 and %Protein2) are distinguished by the leading '%'. The head template can restrict the named entity variables to take values of specific semantic type(s) by providing the list of types in parentheses. Named word sets are distinguished by the leading '$'. They can be defined anywhere in the pattern file and can be used in multiple patterns. In the above example $MODAL is the set of modal verbs (can, may, might, etc). The '~' postfix indicates that the preceding word can be matched in any grammatical form. Multiple patterns extracting identical information are separated by the '|' separator.

MedScan output: The output of MedScan is in an XML-based format describing entities and relation between them (see an example below):

```
<resnet mref="16377759:4" msrc="The catalytic domain of ID{820019=S6K1} could be
phosphorylated by Arabidopsis ID{841259=3-phosphoinositide-dependent protein kinase-1}
(ID{830330=PDK1}), indicating the involvement of ID{830330=PDK1} in the regulation of
ID{820019=S6K}.">
  <nodes>
   <node local_id="N1" urn="urn:agi-llid:841259">
    <attr name="NodeType" value="Protein" />
    <attr name="Name" value="at1g48390" />
   </node>
   <node local_id="N2" urn="urn:agi-llid:820019">
    <attr name="NodeType" value="Protein" />
    <attr name="Name" value="AT3G08720" />
   </node>
  </nodes>
  <controls>
   <control local_id="L1">
    <link type="in" ref="N1" />
    <link type="out" ref="N2" />
    <attr name="ControlType" value="ProtModification" />
    <attr name="ModificationType" value="phosphorylation" />
   </control>
  </controls>
 </resnet>
```


Figure 5. An example of a MedScan output

MedScan Ontology of Relationships: Elsevier has developed ontology of different types of relations between biological entities. Each type of relation has a very specific semantic definition and is typically attributed with additional information, e.g. sign of relations (e.g. positive, negative or unknown) or mechanism (e.g. phosphorylation, methylation, etc). There are three set of patters currently used by MedScan to extract biological relations – patterns focused on extraction of different aspects of protein functions, small molecule functions and disease biomarkers. The Table 2 below shows the scope of biological relationships currently extracted by MedScan.

The current scope of the information extracted by MedScan can be extended by developing new dictionaries covering other aspects of biomedical domain (e.g. focused more on medical or clinical entities) and/or by developing novel information extraction patterns to capture other types of entity relationships.

The Pattern Matcher is extremely fast: it runs through more than 16,000,000 entity-tagged sentences from the entirety of Medline in less than 20 minutes.

- Protein -> Protein
  - Binding
  - Protein modification
  - Expression (positive/negative/unknown)
  - Promoter regulation/Binding
  - Regulation (positive/negative/unknown)
- Protein -> Small Molecules
  - Synthesis/Degradation
  - Mol. Transport
- Protein -> Cell processes
- Protein -> Disease
  - Positive/negative regulation
- Disease -> Protein/Small molecules
  - Changed concentration/expression (positive/negative/unknown)
  - Mutations
  - Activity (positive/negative/unknown)
- Small molecules -> Protein
  - Binding
  - Direct regulation
  - Expression
  - Indirect regulation (positive/negative)
- Small molecules -> Disease/Cell processes (positive/negative/unknown)

Table 2. Relationships currently extracted by MedScan

MedScan Customizations: MedScan is flexible platform open for two types of end-user modifications. First, MedScan taxonomy and dictionaries can be extended to include new concepts and even new concept classes. Dictionaries are provided in a simple text-based tabular format and new concepts and concept aliases can be added to the files. Second, the scope of extracted information can be extended to include new relationships by modifying information extraction rules. The rules are recorded in a well-documented textual format and new rules can be created and added to MedScan.

MedScan Features and competitive advantages: Elsevier's IE engine has been designed and implemented from scratch to address flexibility, precision/recall and performance problems of the off-the-shelf NLP tools. Our design efforts focused on issues specific for texts in vertical application domains characterized by complex sentence and relationship structure, highly specialized entity notation, proliferation of abbreviations and synonyms. As a result of this focus, we have surpassed the 90% precision / 60% coverage mark on technical texts in our current application domains (biology and medicine). Our engine has an unmatched performance – it can process up to 1000 sentences per second on a regular PC, which is 2-3 orders of magnitude faster than prevailing NLP technologies. High performance allowed us to achieve clean separation between modules where traditional approaches intertwine distinct functions like parsing and ontology-based information extraction to cut down on the amount of information exchanged between modules. Also, much attention has been paid to keep domain-specific information in dictionaries and rule files, to simplify maintenance and extending the coverage to other domains.

The engine achieved production quality in 2003 and since then has been installed on many sites, including both individual and corporate-wide licenses.

Elsevier's Information Extraction (IE) technology: Fingerprint Engine

A back-end software system, the Elsevier Fingerprint Engine mines the text of scientific documents – publication abstracts, funding announcements and awards, project summaries, patents, proposals/applications, and other sources – to create an index of weighted terms which defines the text, known as a Fingerprint™ visualization.

By aggregating and comparing Fingerprints, the Elsevier Fingerprint Engine enables institutions to look even beyond metadata and expose valuable connections among people, publications, funding opportunities and ideas.

The Elsevier Fingerprint Engine powers many solutions including Pure, comprehensive information management system, and Reviewer Finder, Elsevier's tool for finding reviewers.

The Elsevier Fingerprint Engine uses a variety of thesauri to support applications pertaining to different subject areas. By applying a wide range of thesauri, Elsevier can develop solutions in but not limited to: the life sciences, engineering, earth and environmental sciences, arts and humanities, social sciences, mathematics and agriculture. Thesauri provided by an institution or specific research domain can also be incorporated.

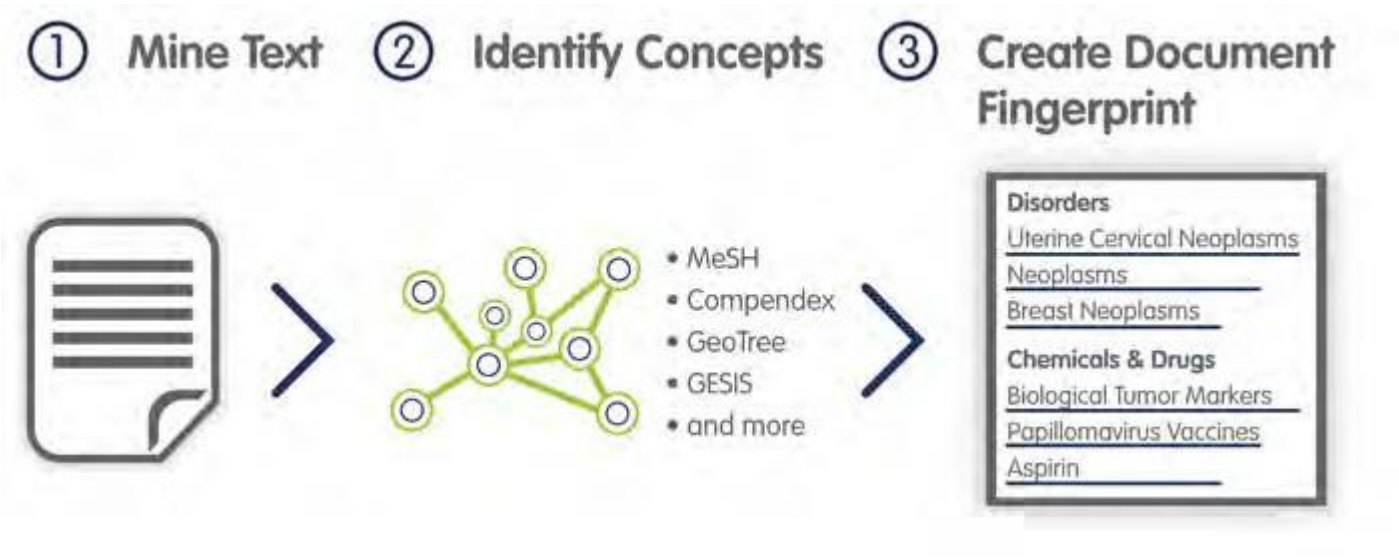


Figure 6: The Elsevier Fingerprint Engine creates Fingerprints via a three-step process

1. The Elsevier Fingerprint Engine applies a variety of Natural Language Processing (NLP) techniques to mine the text of scientific documents including publication abstracts, funding announcements and awards, project summaries, patents, proposals, applications and other sources

2. Key concepts that define the text are identified in thesauri spanning all the major disciplines

3. The Elsevier Fingerprint Engine creates an index of weighted terms that defines the text, known as a Fingerprint.

Applying Fingerprints to inform decision making: By aggregating and comparing Fingerprints of people, publications, funding opportunities and ideas, the Elsevier Fingerprint Engine can reveal insightful connections with practical applications. Here are some examples of how Fingerprints are currently used to bring scholarly business intelligence to institutional data.

- Pure aggregates the Fingerprints of individual documents to create unique Fingerprints that reveal your researchers' distinctive expertise. Pure also matches the Fingerprints of funding opportunities in SciVal® Funding to researchers' Fingerprints, recommending appropriate funding opportunities and suggested collaborators.

- Reviewer Finder compares document Fingerprints with researcher Fingerprints, making it easier to identify reviewers and raise awareness about potential conflicts of interest.

- Elsevier Journal Finder helps researchers find journals that could be best suited for publishing their articles. Journal Finder matches abstracts to Elsevier journals, scanning Elsevier's 2,200+ titles in the Health Sciences, Life Sciences, Physical Sciences and Social Sciences.

NIH Request for Information (RFI): Input on Sustaining Biomedical Data Repositories

For NOT-ES-15-011, the NIH is seeking information that addresses, but is not limited to, the following areas:

- Financial Models – New business models for sustaining digital repositories, including but not limited to examples cited in http://datacommunity.icpsr.umich.edu/sites/default/files/WhitePaper_ICPSR_SDRDD_121113.pdf and http://www.sr.ithaka.org/research-publications/guide-best-revenue-models-and-funding-sources-your-digital-resources.
- Innovation – Sustaining data repositories while enabling new innovations in finding, accessing, integrating and reusing their contents by a wide variety of stakeholders.
- Evaluation - Criteria to determine which data repositories require sustained funding models or no longer need to be sustained, including, but not limited to metrics for measuring the value of given repositories and data within those repositories.
- Best Practices - Current, new, and emerging means or practices to sustain data repositories for the long-term.
- Partnerships - The type, form, and governance of partnerships to ensure long-term access to essential data repositories including, but not limited to, private-sector organizations, non-profit foundations, universities, national and international government agencies, and combinations thereof.
- Technical – Technological developments needed to sustain data repositories in a more cost-effective way while furthering accessibility and usability to a broad set of stakeholders.
- Human Capital – Models to enhance efficiency in the application of human capital associated with data repositories.
- Life Cycle – Consideration of the evolution of value, cost, and scale as data repositories emerge, reach maturity, and either gain or lose relevance in the long term.

Response submitted by Holly Falk-Krzesinski, PhD on behalf of Elsevier, March 18, 2015

Elsevier values the NIH focus on research data and research data repositories and is appreciative for the opportunity to provide a response to NOT-ES-15-011, a Request for Information (RFI) on **Input on Sustaining Biomedical Data Repositories**.

Financial Models

Elsevier is involved in a number of working groups under the "Data Publication" umbrella Interest Group (IG) of the Research Data Alliance, notably the joint RDA/ ICSU World Data System Publishing Data Cost Recovery for Data Centres. The scope of this IG is greatly overlapping with this RFI. Co-chair Anita de Waard of Elsevier and her colleagues recently interviewed 22 data centers about their ideas around cost recovery methods, now and in the future. In summary, Elsevier supports the collaborative efforts of the joint RDA/ICSU WDS Interest Group (IG) to elucidate the full cost of data management throughout its lifecycle–from inception through publication to storage and curation—by engaging funders, researchers, repositories, and other stakeholders in the research data management lifecycle. Specifically, the IG finds that data repositories are looking for new funding mechanisms – including charging deposit fees, access fees, and working through public-private partnerships—but are having trouble finding the time and resources to actively explore these new models. Elsevier is very interested in supporting further work regarding these questions, whether within the scope of the RDA or in direct collaboration with the repositories and/or the NIH. The RDA/ICSU WDS IG is submitting a separate, detailed response to this RFI.

Innovation

Innovation is central area in promoting use of research data and maintaining an open ecosystem while allowing for the creation of services that provide added value. Innovations can range from search services to aggregators and analytical tools. For example, the Open PHACTS project in Europe provides a developer friendly API that enables applications to build across public domain pharmacology data. Their service is supported by pharmaceutical companies through a foundation. Importantly, this service allows proprietary commercial data to sit alongside public data. Three lessons for the NIH arise from this example:

1) Innovation developments should ensure that it is possible to develop a range of services with different business models that store, access, and query various forms of research data. In providing an open model, both in funding and with respect to technological solutions, the NIH can create a flexible framework that allows academic and industry parties to develop components that optimally mesh together and enable systems that can change over time and are tailored to the needs of specific medical and scientific communities;
2) The NIH should seek to develop reporting mechanisms such that downstream aggregators and users can ensure that upstream, publicly funded data providers can receive credit; and,
3) While standardization is helpful for downstream data users, it is important to note that a flexible and open ecosystem can help manage complexity. Therefore, it is preferable to recommend vs. mandate data standards, and any mandates must have the flexibility to allow for change in capabilities and community practice over time.

Evaluation

One element that Elsevier is interested in working with the NIH on is defining the difference between data posting and data publishing. When researchers *post* a description of their research on the web, it is not validated by peers. When the text describing the data is *published*, then others know that the associated research data is peer-reviewed and validated, and thus can be trusted. It is important to make a similar distinction between *data posting* and *data publishing*: validating and quality stamping the data is becoming an ever more important element of a data-driven research community. We need to develop a hierarchy of trust levels of data where at some moment reproducibility levels and algorithms to detect data become a part of that as well. Data validation and data publishing are areas in which Elsevier has deep expertise that we can lend to this.

Elsevier is very interested in supporting a system that evaluates the performance of various components of the biomedical Research Data Management cycle. We are currently actively engaged in a number of conversations with academic and industry partners to enable components to such a shared set of metrics, and systems to support them. We are interested in working in partnership with the NIH and other stakeholders on a workbench that enables quantitative evaluation of the usefulness and usability of different tools pertaining to research data storage, sharing, and search. Questions that one can ask of such a system could include:

- Which data standards, metadata systems, and curation efforts optimally improve outcome of a particular use case, such as data search, or data reuse?
- What metrics can be used for successful data storage or curation: reuse, amount of queries/downloads, or other—possibly social—metrics?
- What systems can act across the spectrum of biomedical repositories, publications, and other research outcomes to track and combine these metrics?

Best Practices/Policy

In its new Plan for Increasing Access to Scientific Publications and Digital Scientific Data from NIH Funded Scientific Research (the Plan), it is very good to see that the, "NIH recognizes the benefit of collaborating with other federal agencies and public and private stakeholders to adopt consistent practices for citation of data sets across scientific communities and other data set attribution systems and will work toward this goal." And a broader context for this can also be found in the HSS Guiding Principles document, which talks about developing healthdata.gov as the basis for a "data commons approach across agencies," specifically the development of an internal HHS Enterprise Data Inventory that will serve as the internal catalog for all HHS data assets and be linked to healthdata.gov, the external-facing platform through which the public will be able locate and access federally funded research data. Elsevier has linking services that could add to this initiative by expanding the reach of healthdata.gov datasets.

The Plan also indicates that, "the NIH will expect funded researchers to deposit data in 'appropriate, existing, publicly accessible repositories before considering other means of making data available,' but where needed, NIH will take steps to support the development of 'selected community-based data repositories and standards.' To help researchers find an appropriate repository to deposit their data, NIH will expand its database of existing repositories and plans to develop guidance and criteria to aid researchers in identifying 'acceptable repositories' not funded by NIH." While we are assuredly in favor of establishing authentication methods for data repositories we contend that researchers need the flexibility to choose where to deposit their research data into repositories and are the most knowledgeable about determining the repository best suited to their data and research. This principle should be at the center of any criteria NIH seeks to develop, and its criteria should not inadvertently limit data publication routes, such as linking data, data journals, interactive data plots, etc.

Rigid funder-specific mandates lead to directing depositing of research data to a limited number of more generic repositories, running the risk of losing discipline- and domain-specific repositories that add significant value for data reuse and reproducibility. Similarly, mandates that require depositing to a single funder's repository will lead to fragmentation on the basis of country, which is counterproductive to the ever-expanding global nature of (biomedical) science and creation and use of (biomedical) research data by international teams of researchers working across sectors. Research data should be created in formats that allow deposition in a multitude of repositories, and published or deposited in any repository that best suits the research and the discipline. It is also important for the NIH not to put a policy in place that requires undue burden on researchers. It should take special care to ensure that NIH-supported investigators working in international collaborations don't find that they are required to meet multiple—and especially not disparate—funder data posting mandates.

That said, the NIH should be a strong partner in defining data repository quality requirements and ensuring that repositories are validated. This would offer the NIH the opportunity for a more flexible policy that allows research data to be stored at repositories that meet specific the quality levels; more flexibility will facilitate compliance on the part of researchers and their institutions. Moreover, quality of repositories must also relate to unfettered access and linking abilities by multiple stakeholders. Recognizing that quality of data repositories is critical, Elsevier encourages the development of data repository certification standards building on initiatives like the Data Seal of Approval, an effort by several data repositories (working in partnership with other research data community stakeholder groups) to ensure sustainable and trusted data repositories.

Partnerships

As stated above, Elsevier is an active partner with the Research Data Alliance (RDA) and ICSU World Data System (ICSU WDS). With such a wide range of stakeholders across for-profit and nonprofit sectors around the world, and an understanding that biomedical research data is a subset of research data more broadly, it is crucial for the NIH to be partner with these collaborative efforts so as not to duplicate work nor move in a direction specific only to research funded by the NIH.

The basis for Elsevier's involvement in partnerships is that we recognize that creating a research data infrastructure (including the technical infrastructure but also policies, best practices, standards, etc.) has to be a collaborative, cross-stakeholder and international effort where all the different players work together. Elsevier is proud to contribute our deep expertise and perspective from our position as a world leader in research information and appreciate having a voice in development of a synergistic and interoperable emerging research data infrastructure.

The RDA is a great forum for such an approach, as it brings together thought leaders in research data from various stakeholder groups (data centers, research institutes, libraries, publishers, funders, interest group, etc.) and individuals working in the research data field with different expertise and focus, all the way from deep technical expertise to policy-making. The primary value of the RDA is that it has become the forum where stakeholder groups come together to interact and work on issues and focus on making realistic progress on a swift timescale (e.g., 18 mos is the typical lifespan of an RDA working group).

Specifically, Elsevier is involved in a number of working groups under the "Data Publication" umbrella Interest Group (IG) and encourages NIH to join in the partnership. All of these working groups began as ICSU WSD working groups and now have dual ICSU WDS/RDA mandate:

- Data Publication Bibliometrics
- Publishing Data Cost Recovery for Data Centres (for more details, see previous paragraph)
- Data Publication Services

Technical

The NIH's recent Plan for Increasing Access to Scientific Publications and Digital Scientific Data from NIH Funded Scientific Research explains that "As part of the data discovery index, a system for unique identifiers for datasets generated by NIH-funded research will be developed, analogous to the PubMed Central identification number (PMCID) that is assigned to all submitted publications resulting from NIH-funded research. The identifier would also provide a means of linking the data with the biomedical literature via associated PubMed records." We would like to take this opportunity to share our thoughts around the NIH participating in development of an open, international standard identifier system built on DOIs.

Data DOI's are becoming a globally recognized standard for biomedical and other types of research data identification. Worthy of noting, a number of big data repositories, including the NIH Protein Data Bank (PDB), have assigned DOIs for all its accession numbers. DataCite, for example, has a valuable set of services connected with it offered at no cost and that make it easier to connect with other systems and DataCite has plans to expand its services to accommodate use cases that it currently cannot support (e.g., unpublished data that is early on in the lifecycle, and which is still subject to change). DataCite

could be positioned to become a resolver for all other data accession numbers, which simplifies the entire research data infrastructure. The mapping of the Data DOI to an accession number is in the DataCite metadata, and so the DataCite API can be used to map accession numbers and then benefit from metadata for that record in DataCite. Other organizations are also focused on collaborative digital data standards development, including: APARSEN; Opportunities for Data Exchange (ODE); CoData; and, NISO/NFAIS Supplemental Journal Article Materials Project.

Elsevier recommends that NIH focus on the use of Data DOIs as the primary open, international identifier option for data that is published in any formal sense, rather than developing a identifier schema. And if the NIH is to develop a new accession number schema, then it must include assigned DOIs as well.

Elsevier further encourages the NIH to leverage the significant amount of work that has gone into developing common ways to *expose and cite* data. For example, the community effort of the FORCE11 Joint Data Citation Implementation Group has led to the creation of a standard for citing data within article publishing (the NISO JATS 1.1d2 XML schema). The Joint Data Citation Principles has been endorsed by over 90 institutions. The paper, "Achieving human and machine accessibility of cited data in scholarly publications," describes how to operationalize those principles. As described in the Partnership section above, this effort further exemplifies the benefits of collaboration between major stakeholders in the scholarly communication ecosystem, focused on biomedical research and other types of research and data more broadly. By leveraging these community-driven efforts, a common basis for new models of sustainability will emerge.

Finally, Elsevier is very interested for the NIH to develop open architectures to which other parties (including commercial) can contribute.

Human Capital

Understanding that a recognition economy is the dominant environment in which academic and government researchers operate, it is essential to consider the drivers of research data sharing at the individual researcher level to maximize rapid and efficacious sharing. The NIH needs to address data sharing incentives and rewards for researchers in development of its policies and procedures. Relying only on the "stick" of mandated policy compliance, the full potential to stimulate and motivate broad sharing of research data will go unmet and will face challenges similar to those related to posting to PubMed Central and ClinicalTrials.gov. Elsevier encourages the NIH to review and operationalize the literature that provides an evidence base for understanding what drives researchers to be participatory data donors and we encourage the NIH to develop *new* research funding programs to extend empirical knowledge about this area of science policy. One approach might be for the NIH to partner with the NSF's Science of Science Innovation and Policy (SciSIP) program to develop a research data stream and funding resources to support new research grants in this area.

The free, public Mendeley Research Data Sharing group contains a rich library of such research data sharing resources. Contained therein, references describe the need to develop a reward and recognition system that affords researchers ongoing attribution, recognition, and professional reward for their sharing efforts. The literature also calls on policy makers, funders, and research organizations to consider the resources necessary for researchers and their institutions to comply with policy mandates, such as necessary skills, time & effort, and ongoing finances. Furthermore, the literature demonstrates

the need for stakeholders to take into account the impact of sharing and potential for misuse on individual competitiveness, an essential consideration given the current hypercompetitive funding landscape.

Finally, the NIH should seek opportunities to collaborate effectively with publishers to avoid duplication of effort and costs associated with research data sharing and to minimize administrative costs to research institutions and burden to researchers. By way of example, in conjunction with the Professional and Scholarly Publishing Division (PSP) of the Association of American Publishers (AAP), Elsevier has been involved with the CHORUS service; which leverages existing infrastructure, tools, and services across publishers that have committed to collaboration with federal funding agencies around the public access of research articles.

Life Cycle

With regards to life cycle, the NIH should strive to work in partnership with other stakeholder groups to develop consistent preservation criteria. To do so, it will be important to address some key questions, such as: Should all versions of data be preserved? Should research data be overwritten with newer data? For how long should data be preserved? Is indefinite preservation sustainable?

Previous RFI Responses

Elsevier recently submitted a response that included information about research data and data repositories to NOT-OD-15-067, a Request for Information (RFI) on Soliciting Input into the Deliberations of the Advisory Committee to the NIH Director (ACD) Working Group on the NLM (NLM Elements RFI). The following is excerpted verbatim from that NLM Elements RFI response. In addition, we wish to call your attention to the NLM Elements RFI response that was submitted by the Professional & Scholarly Publishing Division (PSP) of the Association of American Publishers (AAP; refer to 'Research data' in Comment 5). In addition, the PSP/AAP will be submitting a response to this RFI as well.

Submitted by Holly Falk-Krzesinski, PhD on behalf of Elsevier on March 13, 2015:
Research Data: Elsevier would like to see the NLM allow mining of all database content inside the suite of databases managed and curated by the NLM and provide actionable copyright metadata elements on all NLM content so we understand what we can mine/use for commercial and non-commercial purposes.

Elsevier's research data policy (http://www.elsevier.com/about/research-data) commits us to encouraging and supporting researchers to making their research data freely available with minimal reuse restrictions wherever possible. Alongside our policy, we have developed a range of tools and services to support researchers to store, share, access, and preserve research data. These include our open data pilot, our database linking program, and our data journals, such as *Genomics Data and Data in Brief*. Collectively, Elsevier as partners with NLM, we should to be thinking about the big picture goal of enabling researchers to properly collect and annotate their research data in ways that lead to archiving, auditing, reproducibility, and interoperability. This might include making vocabularies and other data models available in the researchers' workflow (e.g., controlled vocabularies and drop-downs in Electronic Lab Notebooks). This is especially for vocabularies, databases, and other data models that identify entities that define research data (anatomy, diseases, organisms, etc.). Making this available in formats that foster interoperability is a big part of this. This way, unique identifiers and codes are

captured early on and can stay with the research data through its entire lifecycle (whether or not research ends up getting published).

Research data adds huge value to the users of published research articles. An important focus is twofold: 1) Attach and make available publicly the methods and data underlying published research; and, 2) Develop standard markups (XML) to allow machine interpretation of the data (this is an area that Elsevier's Mendeley team is currently working on). It will be important for NLM to work in close partnership with a broad stakeholder group to consider the most effective approach to enforcing data transparency and developing a set of markup standards.

Data fraud detection tools will need to be an important focal point for NLM. In recent scientific fraud causes, fraud was detected as data that was statistically, "too good to be true." Similarly, image manipulation for scientific articles has been observed and is being addressed by a number of publishers at high cost due to the manual labor involved. To avoid future problems and resulting distrust in our data-drive scientific approaches, NLM and publishers will need to work together to find efficient and effective ways to detect data fraud before data sharing and publication.

Regarding research data repositories, we think it is most useful to think in terms of data management plans and data archives. Elsevier is supportive of mandates for data management plans where researchers have the flexibility to choose where to deposit their data and that data publication routes are not limited (e.g., linking data, data journals, interactive data plots, etc.). Importantly, as efforts on research data repositories advance, it will be essential for the NLM to seek out collaboration opportunities with a broad and diverse range of stakeholders across sectors to ensure that collective expertise and experience as leveraged, a duplication of effort and resources are minimized, and cost savings and administrative efficiency are maximized.

There is a need for data standards, but it should also be recognized that such standards do develop continuously. So any standardization proposal should include a proposal for continuous maintenance and further development of the standard. It should also be noted that data standards have to be discipline, perhaps even subdiscipline, specific, and will always have some element of least common denominator as science, by definition, goes beyond what has been standardized.

Tools for automatic mapping of data would indeed be extremely useful as they can provide the input for data search engines. Furthermore, such tools can help scientists to better comply with funder requirements to share data in a meaningful way, especially when such tools are combined with proper (provenance) annotation capabilities.

Elsevier would be very interested in working with the NLM, other publishers, and data archive managers on mechanisms to connect articles and related datasets. It would be valuable for publishers to link plug-ins into their systems, such that authors could submit the data to the archive of their choice and simultaneously link this to an article.

We also feel that it is important that the NLM work with stakeholders on developing capabilities (at a variety of levels) to validate data and mark it as "OK" following a certain hierarchy of quality, from data has been well-described to data that has been fully reproduced in a different environment by a different team. Elsevier's data articles and microarticles do provide one of the steps in this continuum of quality/integrity validation, but there are additional levels beyond peer-review that need to be considered and built into developing systems.

With regards to the quality criteria and quality stamps for data archives, there has been considerable discussion in this space, especially in the EU, but it is essential that there be commonly shared view on what a data

repositories should adhere to, e.g., the National Digital Stewardship Alliance (NDSA) levels of preservation do make a step in one dimension of data repositories (archives), but there are many more dimensions to consider.

Elsevier values its multi-faceted and synergistic relationship with the NIH and appreciates the opportunity to provide a response to NOT-OD-16-133, Request for Information (RFI): Metrics to Assess Value of Biomedical Digital Repositories. Submitted on behalf of Elsevier by:

Holly J Falk-Krzesinski, PhD
Vice President, Strategic Alliances
Global Academic Relations
Elsevier
h.falk-krzesinski@elsevier.com
New York, NY, USA

Response Contents

Part 1: Research Data Definition and Research Data Metrics.....1-8
Part 2: Research Data Repositories.....8-10
Part 3: Data Discoverability.....10-13
Part 4: Recognition and Reward.....13-16

Part 1: Research Data Definition and Research Data Metrics

Definition and Disciplinarity

Research Data Definition

Elsevier's working definition is, "research data refers to the results of observations or experimentation that validate research findings." Research data can also be defined as, "the recorded factual material commonly accepted in the scientific community as necessary to validate research findings."[1] Research data covers a broad range of information types[2], and digital data can be structured and stored in a variety of file formats.

The main goal of research data sharing is to enable other researchers to reuse data. Thus, reusability should always be taken into account when designing systems that create and store research data. We believe that data reuse could be optimized by aligning the 10 aspects of data listed below, Figure 1. This pyramid[3] – loosely modeled on Maslow's hierarchy of human

[1] OMB Circular 110, https://www.whitehouse.gov/omb/fedreg_a110-finalnotice

[2] From 'Defining Research Data' by the University of Oregon Libraries: Documents (text, Word), spreadsheets; Laboratory notebooks, field notebooks, diaries; Questionnaires, transcripts, codebooks; Audiotapes, videotapes Photographs, films; Protein or genetic sequences; Spectra; Test responses; Slides, artifacts, specimens, samples; Collection of digital objects acquired and generated during the process of research; Database contents (video, audio, text, images); Models, algorithms, scripts; Contents of an application (input, output, logfiles for analysis software, simulation software, schemas); Methodologies and workflows; and, Standard operating procedures and protocols.

[3] See figure in '10 aspects of highly effective research data' at https://www.elsevier.com/connect/10-aspects-of-highly-effective-research-data.

needs – can be seen as an extension of the FAIR Data Principles[4] (data should be Findable, Accessible, Interoperable and Reusable) and can function as a roadmap for the development of better data management processes and systems throughout the data lifecycle.

Figure 1: This pyramid can function as a roadmap for the development of better data management processes and systems.

Disciplinarity of Data

While this RFI specifically indicates *biomedical* repositories, it is important to recognize the increasingly interdisciplinary nature of biomedical, life sciences, and health sciences research and the overlaps of research data types from other disciplines.

In a parallel effort, the NSF has been focused on open data and research data through the Open Data Workshop Series[5], the first of which was held in November, 2015. While the workshop's initial focus was on generating discipline-specific responses from the Mathematical and Physical Sciences research communities to the federal policy requiring open data and the recently-released NSF policy statement on open data, there is considerable alignment with the NIH biomedical domain as it relates to research data: decide how and what to preserve in terms of research data for public consumption; the manner by which research data will be stored and accessed; and, the level of burden implied by conservation that is placed on the individual investigator.

[4] Force11 The FAIR Data Principles, https://www.force11.org/group/fairgroup/fairprinciples
[5] NSF MPS Open Data Workshop Series, https://mpsopendata.crc.nd.edu/index.php

International standards organizations, such as the National Data Service (NDS)[6], Research Data Alliance (RDA),[7] and ICSU-World Data System (WDS)[8], have been leading the charge to develop consensus and standards related to research data across disciplines. Elsevier, along with other publishers and research information providers, and additional research ecosystem stakeholders have been working in close partnership with these organizations, and have been engaged with the NSF initiative, as well as working with NIST[9]. These joint efforts have already begun to make significant strides in defining how to publish, find, and reuse research data. We thus recommend that the NIH also participate in this collaborative approach to:

1. Adopt flexible, broad standards and principles related to research data so that all disciplines have the maximum opportunity to interpret research data metrics and demonstrate research impact according to their field and across domains;

2. Consider how to combine quantitative with qualitative inputs; this to ensure that all disciplines, and all agencies and institutions regardless of their disciplinary focus, can share and interpret outcomes and research impact in a similar way;

3. Highlight the full range of types of research data deposit and reuse relevant to many research disciplines, so researchers have the widest opportunity to demonstrate maximum research impact of their work.

Research Metrics

This response focuses on research data, which constitutes an important part of the comprehensive ecosystem of research recognition. We would like to note the following types of research impact that should be considered across the research workflow (Figure 2, below):

1. Research activity – production of outputs leading to enhanced knowledge and understanding, such as original research in journal publications and books, research data, reports, designs, software, etc.; securing income to support ongoing research activities.

2. Research impact – recognition of the influence of research activity on subsequent research through viewing activity, and the receipt of citations from that subsequent research.

[6] NDS, http://www.nationaldataservice.org/
[7] Research Data Alliance, https://rd-alliance.org/
[8] ICSU-WDS, https://www.icsu-wds.org/
[9] Public Access to NIST Research, https://www.nist.gov/open

3. Scholarly impact – the wider recognition of research, beyond citing previous work, within the scholarly community, such as the receipt of prizes, requests to edit a journal and to peer review funding applications, and so on.

4. Economic impact – the production of commercializable outputs such as registered and granted patents and spin-out companies, and income generated from these outputs.

5. Social impact – the achievement of societally relevant outcomes, the enhancement of well-being to society as a result of research outputs and/or activities.

A well-rounded, inclusive recognition system can be assessed on all of the facets mentioned above, including research data, by the responsible use of research metrics as good approximations (proxies) of the actual level of performance. The research metrics that are selected should be complemented by the occasional use of narrative inputs such as case studies, firstly as a sanity check that the research metrics are indeed reflective of performance, and secondly in cases where research metrics cannot capture the full value of the research output or outcome.

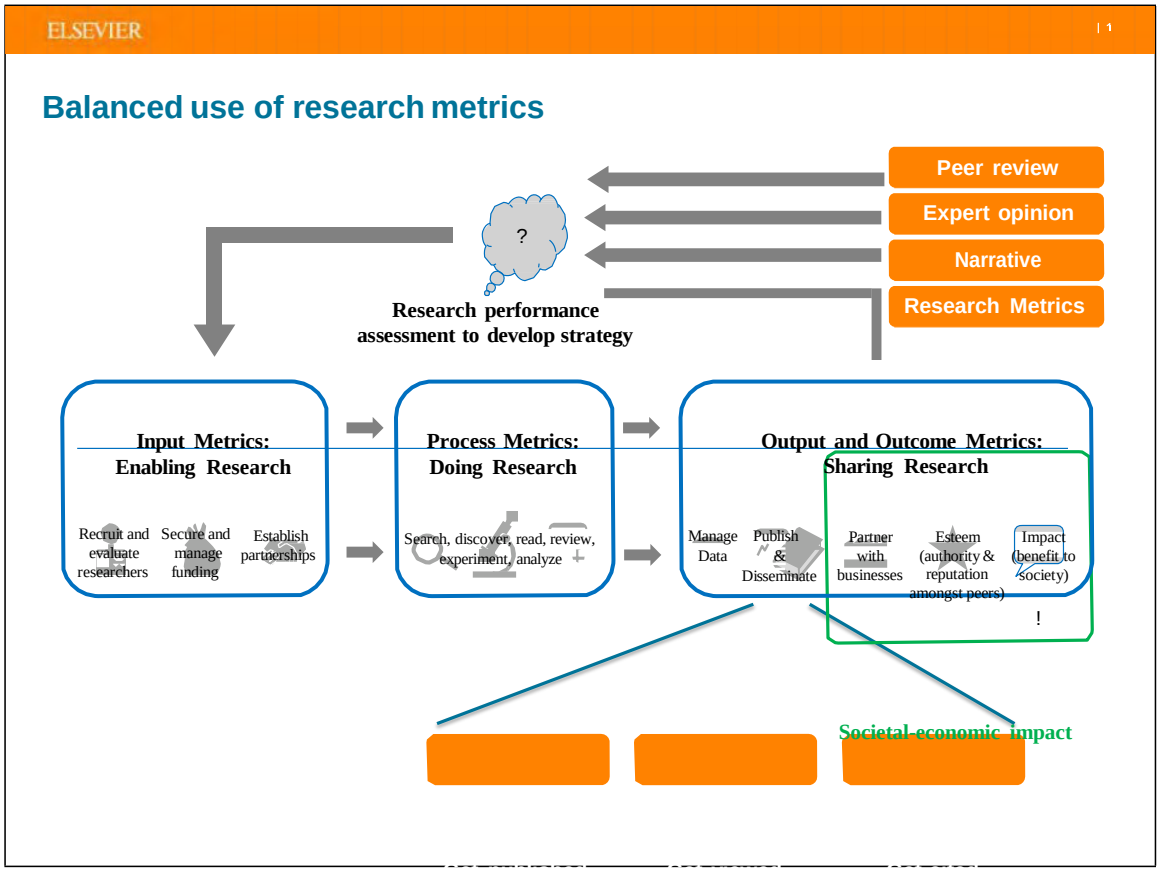


Figure 2: Balanced use of research metrics across the full research workflow.

4

Golden Rules

Elsevier's work with the research community has led us to recommend two "Golden Rules" for working with research metrics:

1. All decisions and participants benefit from a combination of both quantitative indicators and qualitative (e.g., case studies) input;

2. Quantitative input should always be based on at least two metrics (refer to Table 1 and Table 2 below for examples).

These Golden Rules are a practical reflection of the fact that the highest confidence in decision making is achieved when based on the most complete picture possible, which in turn depends on diverse inputs. Indicators reflect a version of the complete truth that is represented in research data repositories, and as such are an effective proxy for performance. The combination of these indicators can create a good impression of a comprehensive picture, as when a jigsaw has enough pieces in place to gain a good impression of the image, but the indicator jigsaw retains gaps, even when the underlying data sources are comprehensive and a broad set of indicators are used. Consequently, we recommend always complementing quantitative input from indicators with qualitative input from narratives to bring the view into sharper focus, and equally, we recommend that qualitative inputs are always used in combination with indicators.

Basket of Metrics

In close partnership with the research community, we have developed a 'basket of metrics' approach to using research metrics representing all types of research activity across the research workflow (Figure 2); research data metrics are no exception. In the next section, we list research data metrics that would be useful to help measure research impact, but would like to make some general comments about the advantages of an approach that builds on a multiplicity of research metrics here. The advantages of a 'basket of metrics' are:

1. Research excellence, even in one area such as research data, covers a broad range of concepts, and this diversity is best captured by considering a broad range of research metrics.

2. Funders and institutions need flexibility to determine the most appropriate research metrics to demonstrate research impact.

3. The set of research metrics offered can be read out in different ways, which accommodates the expectation by the research community for both simple research

metrics and more sophisticated, but complex, ones. Our research[10] shows that both types are needed and appreciated by users, and both types are important in offering the most complete picture of performance.

a. Simple research metrics such as total counts of activity, and counts normalised by university or faculty size (expressing the indicator as a proportion (%) of total, or by dividing the total count by number of researchers or outputs), are useful for offering transparency and clarity on the underlying data, and for showing the magnitude of activity in absolute terms.

b. More complex research metrics, such as field-normalised algorithms, take into account different behavior between fields and so enable the fair comparison of relative performance in physics with that in biology, for instance.

4. Our work with the community has led us to recommend Two Golden Rules of using research metrics. We discussed the first, always using quantitative measurements together with qualitative inputs, in question 4. The second Golden Rule is to always use at least two quantitative indicators as input into any decision. We recommend that any instance of research impact is demonstrated by using at least two research metrics, because:

   a. Every single indicator has its weaknesses as well as its strengths, and these weaknesses can be complemented, or balanced, by the strengths of other indicators.

   b. It reduces the likelihood of game playing. There is not, and will never be, one single research metrics that encompasses all aspects of excellent performance. If we try to reduce excellent performance to any single research metric, we will almost certainly drive unbalanced, undesirable behaviour; the researchers could work out how to optimise their performance according to that one research metric. It is much more difficult to see how researchers could adjust their behaviour when the outcomes of that behaviour are measured by using two, or three, or five different research metrics, except by doing genuinely better research across a range of outcomes – which is a result that the NIH is aiming to encourage.

[10] Extensive user research is represented in L. Colledge and C. James, 2015, A "basket of metrics"—the best support for understanding journal merit, European Science Editing 41(3), p61-65; http://europeanscienceediting.eu/articles/a-basket-of-metrics-the-best-support-for-understanding-journal-merit/

Metrics for Research Data

According to the Digital Curation Centre (DCC)[11], "a key measure of the worth of research is the impact it has or, to put it another way, the difference it is making both within the academic community and beyond." It is therefore in the interests of researchers, institutions, and funders to track the impact of research, starting with the impact of research outputs. Historically, research output used to evaluate impact was primarily peer-reviewed research articles. In recent years, other forms of research output are being recognized. The NIH now identifies research data as a legitimate type of 'research product' that can be listed in the "Contributions to Science" section of biosketches submitted as part of a grant application, carrying equal weight with publications.

Elsevier, through Scopus, is leading the way in displaying and collecting journal, article, and author level metrics around scientific literature[12], and intends to do the same for research data (see more below in the "Citation in Practice – The Scopus Model" section). Elsevier's Metrics team, with input from members of the NIH Big Data to Knowledge (BD2K) team, has developed an initial set of quantitative research data metrics (Table 1 and Table 2). All of the research data metrics presented in both tables can be calculated at multiple levels of aggregation (e.g., institution or discipline).

Table 1: Types of Research Data Metrics

Category	Research Data Metric	Description
Collaboration	Collaboration	Proportion of research data outputs with international, or national, or institutional, or no co-authors
Posting	Research Data Outputs	Total count of research data outputs
Get Viewed	Search Count	Total count of times research data outputs have been returned in a search
Get Viewed	Views Count	Total count of views
Get Viewed	Views Percentile measurement	For an individual piece of research data, this would be its percentile according to views received, compared to similar research data outputs For an aggregate entity like an institution, this will be proportion of research data outputs that fall into the top 1%, 5%, 10% or 25% of the world of research data outputs
Get Cited	Citation Count	Total count of citations

[11] Why measure the impact of research data?, http://www.dcc.ac.uk/resources/how-guides/track-data-impact-metrics#why-measure-the-impact-of-research-data
[12] Scopus metrics, https://www.elsevier.com/solutions/scopus/features/metrics

Get Cited	Cited Research Data Outputs	Proportion of Research Data Outputs that have been cited at least once
Get Cited	Citations Percentile measurement	As Views Percentile Measurement
Economic Impact	Academic-Corporate Collaboration	Proportion of research data outputs with both academic and corporate co-authors
Scholarly Impact	Scholarly Activity	This is the total of Mendeley deposits, CiteULike deposits, and similar kind of activity. You can then slice and dice by each individually
Scholarly Impact	Scholarly Commentary	Total mentions in e.g. F1000. You can then slice and dice by each individually
Social Impact	Social Activity	This is the total of Tweets, Facebook likes, and similar kind of activity. You can then slice and dice by each individually
Social Impact	Mass Media	Total mentions in mass media. There are a few variants of this metric we have worked on for publications and which could be applied

Table 2: Research Data Repository Metrics

Category	Research Data Metric	Description
Data Reuse	Data Linkage	Proportion of papers with research data associated with them
Data Reuse	Data Depositing	Proportion of researchers that deposit research data within a certain time frame

Part 2: Research Data Repositories

Defining Trustworthiness

Elsevier has been actively working in robust and deep partnership with numerous national and international research data organizations developing standards for research data repositories. These organizations have made significant strides in defining the criteria that should be used to develop and certify trusted research data repositories.

The most advanced existing data repository certification schemes are:

- Data Seal of Approval (DSA)[13]
- World Data Scheme (WDS) Certification[14]
- Trusted Repositories Audit & Certification (TRAC)[15]

[13] DSA, http://www.datasealofapproval.org/en/
[14] WDS Certification, https://www.icsu-wds.org/services/certification

- Digital Curation Centre (DCC)'s Nestor Catalogue of Criteria for Trusted Digital Repositories[16]

DSA and WDS, whose schemas both rely on self-assessment, are combining their efforts through the Research Data Alliance (RDA)'s Repository Audit and Certification DSA–WDS Partnership Working Group[17] for "realizing efficiencies, simplifying assessment options, stimulating more certifications, and increasing impact on the community. The output from this WG is envisioned as a possible first step towards developing a common framework for certification and a service of trusted data repositories."

DSA includes 16 guidelines[18] covering data producers, data repositories, and data consumers. DSA already has a process in place for the full range of research data repositories to obtain certification, and it maintains a directory of repositories that have successfully acquired certification. The developing DSA-WDS Common Requirements[19] creates a harmonized set of criteria for certification of repositories at the core level addressing research data repository sustainability issues in the areas of organizational infrastructure, digital object management, technology, financial, and legal, etc. Furthermore, the DSA-WDS joint initiative plans to collaborate on a global framework for repository certification that moves from the core to the extended (NESTOR-Seal[20]), to the formal (ISO 16363[21]) level.

Rather than constructing schemas anew specific to biomedical repositories, the current DSA and WDS guidelines and developing Common Requirements must be applied to biomedical repositories to ensure the greatest potential for discoverability and reuse of research data that results from NIH-funded studies and other biomedical research.

Obtaining Certification

From its inception, Elsevier has incorporated the guidance developed by the aforementioned organizations into the development of our multidisciplinary data repository, **Mendeley Data**[22].

---

[15] TRAC, http://www.dcc.ac.uk/resources/repository-audit-and-assessment/trustworthy-repositories

[16] DCC Nestor Catalogue of Criteria for Trusted Digital Repositories, http://www.dcc.ac.uk/resources/repository-audit-and-assessment/nestor

[17] Repository Audit and Certification DSA–WDS Partnership WG, https://rd-alliance.org/groups/repository-audit-and-certification-dsa%E2%80%93wds-partnership-wg.html

[18] DSA Guidelines, http://www.datasealofapproval.org/en/information/guidelines/

[19] DSA-WDS Common Requirements, https://rd-alliance.org/system/files/DSA%E2%80%93WDS%20Catalogue%20of%20Common%20Requirements%20V2.2.pdf

[20] NESTOR Seal for Trustworthy Digital Archives, http://www.langzeitarchivierung.de/Subsites/nestor/EN/nestor-Siegel/siegel_node.html

[21] ISO 16363 Trusted Digital Repositories Management Systems, http://anab.org/programs/isoiec-17021/ms-accreditation-programs/digital-repositories-iso-16363/

[22] Mendeley Data, https://data.mendeley.com/

A critical and absolute criterion of a trusted repository, but one often overlooked by many data repositories, is a mechanism for *long-term* preservation of digital assets. Elsevier has long been a leader in the area of permanent e-journal preservation and an advocate of publisher and research information provider responsibility for digital archiving. Just as Elsevier has done for content published in our journals, we teamed up with DANS (Data Archiving and Networking Services)[23] to ensure that all research datasets within Mendeley Data will be sent offsite to DANS, where they will ensure that the research data is safely archived.

Elsevier is also in the process of obtaining the Data Seal of Approval for Mendeley Data.

Part 3: Data Discoverability

Data Indexing

Elsevier's DataSearch[24] is a prototype research data search engine developed by Elsevier's Research Data Management team that allows users to search for research data across domains and types, from domain-specific, cross-domain, and institutional data repositories. The tool is an exploration of what a search engine for research data needs to look like (versus a web search engine or a document search engine). DataSearch currently indexes images, tables and supplementary data from content sources[25], considered 'research data components.' DataSearch also indexes a series of domain-specific repositories, as well as non-domain specific ones[26]. We are exploring how we might integrate DataSearch with our other offerings, such as Mendeley Data, Scopus, and Pure, to provide robust research data management solutions across the research workflow. And for both, we are working with BD2K on the inclusion of Mendeley Data and DataSearch into the NIH Data Commons.

DataSearch harvests data through APIs (application program interfaces) from various repositories or, in some cases, through database dump files provided to the project. We then normalize the data to our data model, index the data to make it searchable, and generate previews of data where possible. Users can go directly to the source repository from the preview page.

[23] DANS, https://dans.knaw.nl/en
[24] Elsevier DataSearch, https://datasearch.elsevier.com/
[25] Other than from Elsevier's ScienceDirect, DataSearch only indexes open data from open access repositories
[26] As of June 2016, DataSearch is indexing the following content sources: Tables, figures and supplementary data associated with papers in ScienceDirect, arXiv and PubMed Central; Mendeley Data; NeuroElectro; Dryad; PetDB; ICPSR; Harvard Dataverse; and ThemoML at NIST Thermodynamic Research Center (TRC). We are currently investigating DataSearch being able to index all of the NIH-supported data repositories (see https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html for list). We will continue to add more content sources in the future.

Elsevier uses a pilot set of criteria to select repositories to index in DataSearch, including the number of users, the ease of our ability to index the repository data, and relationships we have with data repository managers. We are committed to indexing all 63 NIH-supported repositories[27] in DataSearch; we cannot do them all at once, however, so we will seek input from the NIH on ranking/prioritization.

We are also engaging with data repositories to investigate how we can most effectively combine efforts regarding data discovery options, including having DataSearch power search on the repositories themselves. The DataSearch team is working with the NIH-funded bioCADDIE (biomedical and healthCAre Data Discovery Index Ecosystem)[28] team, which has been developing a data discovery index prototype[29] that indexes data that are stored elsewhere, and Elsevier is exploring how we can better collaborate through shared interfaces and API's.

Data Citation

For data to be discovered and acknowledged it must be widely accessible and cited in a consistent and clear manner in the scientific literature. Elsevier endorses the Joint Declaration of Data Citation Principles[30], which will render research data an integral part of the scholarly record, properly preserved and easily accessible, ensuring that researchers get proper credit for their work. The citation principles focus on Importance, Credit and Attribution, Evidence, Unique Identification, Access, Persistence, Specificity and Verifiability, and Interoperability and Flexibility. A data citation is included in the standard References list of an article, and treated on equal footing with article citations.

In Elsevier's ScienceDirect platform, this means readers will enjoy the same benefits with data as they do with article citations, including one-click deep links to the referenced material and the ability to quickly jump to the point in the article where the work was first cited (see Figure 3 below).

[27] NIH Data Sharing Repositories, https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html
[28] bioCADDIE, https://biocaddie.org/about
[29] DataMed, https://datamed.org/
[30] Joint Declaration of Data Citation Principles, https://www.force11.org/group/joint-declaration-data-citation-principles-final

References

Barnett et al., 2013 C.L. Barnett, N.A. Beresford, L.A. Walker, M. Baxter, C. Wells, D. Copplestone
Element and radionuclide concentrations in representative species of the ICRP's reference animals and plants and associated soils from a forest in North-west England
NERC — Environmental Information Data Centre (2013) http://dx.doi.org/10.5285/e40b53d4-6699-4557-bd55-10d196ece9ea

Beresford, 2010 N.A. Beresford
The transfer of radionuclides to wildlife (Editorial)
Radiat Environ Biophys, 49 (2010), pp. 505–508

View Record in Scopus | Full Text via CrossRef | Citing articles (10)

Beresford et al., 2008a N.A. Beresford, M. Balonov, K. Beaugelin-Seiller, J. Brown, D. Copplestone, J.L. Hingston, *et al.*
An international comparison of models and approaches for the estimation of the radiological exposure of non-human biota
Appl Radiat Isot, 66 (2008), pp. 1745–1749

Article | PDF (272 K) | View Record in Scopus | Citing articles (27)

Figure 3: The image shows a reference list from the article "A new approach to predicting environmental transfer of radionuclides to wildlife: A demonstration for freshwater fish and caesium," published in *Science of the Total Environment* 2013.

Citation in Practice – The Scopus Model

Elsevier's Scopus is the largest abstract and citation database of peer-reviewed literature: scientific journals, books/book chapters, and conference proceedings. Delivering a comprehensive overview of the world's research output in the fields of science, technology, medicine, social sciences, and arts and humanities, Scopus features smart tools to track, analyze, and visualize research and its impact. Scopus' vision of research data aligns with the Force11 Joint Declaration of Data Citation Principles[31] which state that research data is as integral to recognizing and assessing the research output of modern researchers as are articles, reviews, books and all other "traditional" forms of research output (refer to Figure 2). Thus, research data must be:

- Discoverable
- Trustworthy
- Included in the author profile
- Creditable

DataSearch and Scopus are taking a complementary approach. Whereas DataSearch indexes a number of data sources and allows researchers to discover, access, and preview relevant data sets in multiple formats, the goal for Scopus is to integrate and curate DataSearch results to ensure that the research data discoverable via Scopus.com is trustworthy, in a manner consistent with the approach we take toward traditional content inclusion by way our independent Content Selection & Advisory Board (CSAB)[32].

[31] https://www.force11.org/group/joint-declaration-data-citation-principles-final

[32] ScopusCSAB, https://www.elsevier.com/solutions/scopus/content/scopus-content-selection-and-advisory-board

Presently the Scopus CSAB vets all journals indexed in Scopus to ensure high quality standards. We believe that a similar methodology should be applied to data repositories, to ensure transparent, consistent, high quality content

Integrating a research data search engine such as DataSearch in Scopus as a prototype will require a combination of human and algorithmic curation techniques to ensure that Scopus users can trust and rely on the results. In order to achieve this, we intend to apply rigorous selection criteria to both data repositories and data types (refer to the sections on "Research Data Definition" and "Defining Trustworthiness" above for criteria that we will consider).

After ensuring research data is discoverable, the next step will be for Scopus to integrate research data citations in Scopus Author Profiles, to appropriately link and assign credit to the author. Metrics can be applied to research data citations in Scopus just as they are now for articles (refer to the section above, "Metrics for Research Data").

Scopus is leading the way in displaying and collecting journal, article, and author level metrics around scientific literature[33], and intends to do the same for research data. Several parameters will be developed to attribute metrics to data. Scopus will collect and display these metrics in a way that is clear and imparts meaning and value to each metric. Through these efforts, Elsevier can enhance recognition across the research workflow (Figure 2) through enhancement of data search and credit for research data output.

Part 4: Recognition and Reward

While this RFI doesn't specifically identify the topic recognition and reward of research data to support widespread research data sharing, we think that the issue is inextricably linked to the sustainability of research data repositories.

At the SciDataCon 2016 conference in September, 2016, there was a session entitled, *Getting the incentives right: Removing social, institutional and economic barriers to data sharing*[34]. The session description indicates that while "much work has been done relating to the technical aspects of scientific data sharing…[progress toward research data sharing]…has been particularly hampered by a lack of awareness that the barriers and risks to be addressed are socio-technical concerns, with the non-technical concerns –the social, institutional and economic aspects of data sharing, often overlooked."

[33] Scopus metrics, https://www.elsevier.com/solutions/scopus/features/metrics
[34] http://www.scidatacon.org/2016/sessions/37/

Elsevier has been working with the research data community to compile a body of literature addressing the socio-technical aspects of research data sharing rewards and incentives, as well as relevant references on knowledge sharing incentive systems (Table 3)[35]. We recommend that this literature be comprehensively evaluated with the goal of developing recommendations for effective policies and practices that the NIH (and other funders), research institutions, and faculty promotion & tenure committees can employ to promote research data sharing.

Table 3: References on Rewards and Incentives for Research Data Sharing

1. Anderson MS, Ronning E a., De Vries R, Martinson BC. The perverse effects of competition on scientists' work and relationships. Sci Eng Ethics. 2007;13:437–61.
2. Arzi S, Rabanifard N, Nassajtarshizi S, Omran N. Relationship among Reward System, Knowledge Sharing and Innovation Performance. Interdiscip J Contemp Res Bus. 2013;5(6):115–41.
3. Bartol K. Encouraging Knowledge Sharing: The Role of Organizational Reward Systems. J Leadersh & Organ Stud. 2002;9(1):64–76.
4. Birney E, Hudson TJ, Green ED, Gunter C, Eddy S, Rogers J, et al. Prepublication data sharing. Nature [Internet]. 2009;461(7261):168–70. Available from: http://dx.doi.org/10.1038/461168a
5. Borgman CL. The Conundrum of Sharing Research Data. SSRN Electron J [Internet]. 2011;63(6):1–40. Available from: http://www.ssrn.com/abstract=1869155
6. Boudreau KJ, Lakhani KR. "Open" disclosure of innovations, incentives and follow-on reuse: Theory on processes of cumulative innovation and a field experiment in computational biology. Res Policy. 2015;44(1):4–19.
7. Bourne PE, Lorsch JR, Green ED. OUTLOOK BIG DATA IN BIOMEDICINE Sustaining the big-data ecosystem. 2015;
8. Carrara W, Fischer S, Steenbergen E van. Open Data Maturity in Europe 2015: Insights into the European state of play. European Data Portal Open. 2015.
9. Chia-Shen C, Shih-Feng C, Chih-Hsing L. Understanding Knowledge-Sharing Motivation, Incentive Mechanisms, and Satisfaction in Virtual Communities. Soc Behav Personal An Int J [Internet]. 2012;40(4):639–47. Available from: http://proxy.indianatech.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=s3h&AN=75245509&site=ehost-live&scope=site
10. Costello MJ. Motivating Online Publication of Data. Bioscience [Internet]. 2009;59(5):418–27. Available from: http://www.jstor.org/stable/10.1525/bio.2009.59.5.9
11. Cress U, Barquero B, Schwan S, Hesse FW. Improving quality and quantity of contributions: Two models for promoting knowledge exchange with shared databases. Comput Educ [Internet]. 2007 [cited 2016 Sep 15];49(2):423–40. Available from: http://www.sciencedirect.com/science/article/pii/S0360131505001375
12. Denis J, Goëta S. Exploration, Extraction and "Rawification". The Shaping of Transparency in the Back Rooms of Open Data. Soc Sci Res Netw [Internet]. 2014; Available from: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2403069&download=yes
13. Edwards PN, Mayernik MS, Batcheller AL, Bowker GC, Borgman CL. Science friction: Data, metadata, and collaboration. Soc Stud Sci [Internet]. 2011 Aug 15 [cited 2012 Mar 22];41(5):667–90. Available from: http://sss.sagepub.com/cgi/content/abstract/0306312711413314v1
14. Ember C, Hanisch R, Alter G, Berman H, Hedstrom M, Vardiagn M. Sustaining Domain Repositories for Digital Data: A White Paper. 2013;(February):1–16.
15. Enke N, Thessen A, Bach K, Bendix J, Seeger B, Gemeinholzer B. The user's view on biodiversity data sharing - Investigating facts of acceptance and requirements to realize a sustainable use of research data -. Ecol Inform. 2012;11:25–33.
16. Fecher B, Friesike S, Hebing M. What Drives Academic Data Sharing? SSRN Electron J [Internet]. Berlin, Germany; 2014;10(2). Available from: http://papers.ssrn.com/abstract=2439645
17. Fecher B, Friesike S, Hebing M, Linek S, Sauermann A. A Reputation Economy: Results from an Empirical Survey on Academic Data Sharing [Internet]. Berlin, Germany; 2015. Report No.: 1454. Available from: http://d.repec.org/n?u=RePEc:diw:diwwpp:dp1454&r=sog
18. Fitch P, Craglia M, Pollock R, Cox S, Fowler D. Getting the incentives right: removing social, institutional and economic barriers to data sharing [Internet]. International Data Week Conference Session. Denver, CO, USA; 2016. Available

[35] Holly Falk-Krzesinski, PhD, at Elsevier can be contacted directly to be added to the growing reference group, h.falk-krzesinski@elsevier.com

from: http://www.scidatacon.org/2016/sessions/37/

19. Friesike S, Fecher B, Hebing M, Linek S. Reputation Instead of Obligation : Why We Need to Forge New Policies to Motivate Academic Data Sharing [Internet]. Blog Post. Alexander von Humboldt Institute for Internet and Society; 2015. p. 5–8. Available from: http://www.hiig.de/en/23202/
20. Gardner D, Toga AW, Ascoli G a, Beatty JT, Brinkley JF, Dale AM, et al. Towards effective and rewarding data sharing. Neuroinformatics.2003;1(3):289–95.
21. Gaulé P, Maystre N. Getting cited: Does open access help? Res Policy. 2011;40(10):1332–8.
22. Goëta S. Instaurer des données, instaurer des publics- Une enquête sociologique dans les coulisses de l'open data: Instantiate data, instantiate publics- a sociological inquiry in the backrooms of open data. [Paris, France, France]: Télécom Paris Tech; 2016.
23. Gorgolewski KJ, Margulies DS, Milham MP. Making data sharing count: A publication-based solution. Front Neurosci. 2013;(7 FEB).
24. Hung SY, Durcikova A, Lai HM, Lin WM. The influence of intrinsic and extrinsic motivation on individuals knowledge sharing behavior. Int J Hum Comput Stud. 2011;69(6):415–27.
25. Ingram C. How and why you should manage your research data: a guide for researchers [Internet]. JISC; 2016. Available from: https://www.jisc.ac.uk/guides/how-and-why-you-should-manage-your-research-data?mkt_tok=3RkMMJWWfF9wsRonuqjMZKXonjHpfsX56%2B4pW6S%2BlMI%2F0ER3fOvrPUfGjI4ATMRmI%2BSLDwEYGJlv6SgFTrLHMa1izLgNUhA%3D
26. Jahani. Is Reward System and Leadership Important in Knowledge Sharing Among Academics? American Journal of Economics and Business Administration. 2011. p. 87–94.
27. Kim Y, Adler M. Social scientists' data sharing behaviors: Investigating the roles of individual motivations, institutional pressures, and data repositories. Int J Inf Manage. 2015;35(4):408–18.
28. Koers H. How do we make it easy and rewarding for researchers to share their data? – a publisher's perspective. J Clin Epidemiol [Internet]. 2015 Jul [cited 2015 Jul 14]; Available from: http://www.sciencedirect.com/science/article/pii/S089543561500325X
29. Li Y-M, Jhang-Li J-H. Knowledge sharing in communities of practice: A game theoretic analysis. Eur J Oper Res [Internet]. 2010;207(2):1052–64. Available from: http://www.sciencedirect.com/science/article/pii/S0377221710003899
30. Lin S-W, Lo LY-S. Mechanisms to motivate knowledge sharing: integrating the reward systems and social network perspectives. J Knowl Manag [Internet]. 2015;19(2):212–35. Available from: http://www.emeraldinsight.com.ezp.lib.unimelb.edu.au/doi/full/10.1108/JKM-05-2014-0209
31. Longo DL, Drazen JM. Data Sharing. N Engl J Med [Internet]. 2016;374(3):276–7. Available from: http://dx.doi.org/10.1056/NEJMe1516564\nhttp://www.nejm.org/doi/full/10.1056/NEJMe1516564
32. Mayernik MS, Callaghan S, Leigh R, Tedds J, Worley S. Peer Review of Datasets: When, Why, How. Bull Am Meteorol Soc. 2014;1–32.
33. Mueller-Langer F, Andreoli-Versbach P. Open Access to Research Data: Strategic Delay and the Ambiguous Welfare Effects of Mandatory Data Disclosure. Berlin, Germany; 2014. Report No.: 239.
34. Muller R, Spiliopoulou M, Lenz H. The Influence of Incentives and Culture on Knowledge Sharing. System Sciences, 2005 HICSS '05 Proceedings of the 38th Annual Hawaii International Conference on [Internet]. 2005. p. 247b--247b. Available from: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1385745
35. Nelson B. Data sharing: Empty archives. Nature [Internet]. 2009;461:106–63. Available from: http://www.nature.com/news/2009/090909/full/461160a.html
36. Niu JJ. Reward and Punishment Mechanism for Research Data Sharing. IASSIST Q. 2006 May;(Winter).
37. Pham-Kanter G, Zinner DE, Campbell EG. Codifying collegiality: Recent developments in data sharing policy in the life sciences. PLoS One [Internet]. 2014;9(9). Available from: http://journals.plos.org/plosone/article/asset?id=10.1371/journal.pone.0108451.PDF
38. Pisani E, AbouZahr C. Sharing health data: good intentions are not enough. Bull World Health Organ [Internet]. 2010;88(6):462–6. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2878150&tool=pmcentrez&rendertype=abstract
39. Piwowar HA, Chapman WW. A review of journal policies for sharing research data. Open Scholarship: Authority, Community, and Sustainability in the Age of Web 20 - Proceedings of the 12th International Conference on Electronic Publishing, ELPUB 2008 [Internet]. 2008. p. 1–14. Available from: http://elpub.scix.net/cgi-bin/works/Show?001_elpub2008
40. Poline J-B, Breeze JL, Ghosh S, Gorgolewski K, Halchenko YO, Hanke M, et al. Data sharing in neuroimaging research. Front Neuroinform [Internet]. 2012;6:9. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3319918&tool=pmcentrez&rendertype=abstract
41. Ranganathan K, Ripeanu M, Sarin a., Foster I. Incentive mechanisms for large collaborative resource sharing. IEEE Int Symp Clust Comput Grid, 2004 CCGrid 2004 [Internet]. 2004;1–8. Available from:

http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1336542

42. Roche DG, Kruuk LEB, Lanfear R, Binning SA. Public Data Archiving in Ecology and Evolution: How Well Are We Doing? PLOS Biol [Internet]. 2015;13(11):e1002295. Available from: http://dx.plos.org/10.1371/journal.pbio.1002295
43. Rolland B. Data Sharing and Reuse: Expanding Our Concept of Collaboration [Internet]. Team Science Toolkit Blog. 2016 [cited 2016 Jan 1]. Available from: https://www.teamsciencetoolkit.cancer.gov/Public/ExpertBlog.aspx?tid=4
44. Rood RB, Edwards PN. Climate Informatics: Human Experts and the End-to-end System. Earthzine [Internet]. 2014;1–14. Available from: http://earthzine.org/2014/05/22/climate-informatics-human-experts-and-the-end-to-end-system/
45. Šajeva S. Encouraging Knowledge Sharing among Employees: How Reward Matters. Procedia - Soc Behav Sci [Internet]. 2014;156(April):130–4. Available from: http://www.sciencedirect.com/science/article/pii/S1877042814059540
46. Sarathy R, Muralidhar K. Secure and useful data sharing. Decis Support Syst. 2006;42(1):204–20.
47. Sayogo DS, Pardo TA. Exploring the determinants of scientific data sharing: Understanding the motivation to publish research data. Gov Inf Q. 2013;30(SUPPL. 1).
48. Schofield PN, Bubela T, Weaver T, Portilla L, Brown SD, Hancock JM, et al. Post-publication sharing of data and tools. Nature [Internet]. 2009;461(7261):171–3. Available from: http://dx.doi.org/10.1038/461171a\nhttp://www.nature.com/doifinder/10.1038/461171a
49. Seonghee K, Boryung J. An analysis of faculty perceptions: Attitudes toward knowledge sharing and collaboration in an academic institution. Libr Inf Sci Res. 2008;30(4):282–90.
50. Shuman LJ. Data Sharing in Engineering Education. Adv Eng Educ [Internet]. 2016;5(2). Available from: http://advances.asee.org/summer-2016-volume-5-issue-2/
51. Stanley B, Stanley M. Data sharing: The primary researcher's perspective. Law Hum Behav. 1988;12(2):173–80.
52. Sterling TD, Weinkam JJ. Sharing scientific data. Commun ACM [Internet]. 1990;33(8):112–9. Available from: http://portal.acm.org/citation.cfm?id=79182\nhttp://delivery.acm.org.ezp2.bath.ac.uk/10.1145/80000/79182/p112-sterling.pdf?key1=79182&key2=8630174821&coll=GUIDE&dl=GUIDE&CFID=104901268&CFTOKEN=61271004
53. Tenopir C, Allard S, Douglass K, Aydinoglu AU, Wu L, Read E, et al. Data Sharing by Scientists: Practices and Perceptions. Neylon C, editor. PLoS One [Internet]. 2011 Jun 29 [cited 2011 Jun 30];6(6):e21101. Available from: http://dx.plos.org/10.1371/journal.pone.0021101
54. Thorisson G a. Accreditation and attribution in data sharing. Nat Biotechnol [Internet]. 2009;27(11):984–5. Available from: http://dx.doi.org/10.1038/nbt1109-984b
55. Toronto International Data Release Workshop Authors. Prepublication data sharing. Nature. 2009;461(September):168–70.
56. Van Noorden R. Data-sharing: Everything on display. Nature. 2013;500:243–5.
57. Wallis JC, Rolando E, Borgman CL. If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology. PLoS One [Internet]. 2013;8(7):e67332. Available from: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0067332
58. Weber NM, Baker KS, Thomer AK, Chao TC, Palmer CL. Value and context in data use: Domain analysis revisited. Proc Assoc Inf Sci Technol [Internet]. 2012;49(1):1–10. Available from: http://onlinelibrary.wiley.com/doi/10.1002/meet.14504901168/abstract
59. Yang HL, Wu TCT. Knowledge sharing in an organization. Technol Forecast Soc Change. 2008;75(8):1128–56.
60. Zimmerman A. Data Sharing and Secondary Use of Scientific Data: Experiences of Ecologists [Internet] [phdthesis]. 2003. Available from: http://deepblue.lib.umich.edu/dspace/handle/2027.42/39373
61. Zinner D, Pham-Kanter G, Campbell E. The Changing Nature of Scientific Sharing and Withholding in Academic Life Sciences Research: Trends From National Surveys in 2000 and 2013. Acad Med [Internet]. 2016;91(3):433–40. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26675188

Submission Date

01/17/2017

Submitter Name

Vincent Mor, PhD

Name of Organization

Brown University

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

Health Services Research

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

• Databases generated through intense research efforts such as the consolidation of information found through publicly available web searches, for example a collection of topic-specific policies or geocodes for health care or residential facilities. These types of data efforts consolidate objective information that are useful across a broad spectrum of research questions and are often labor intensive to the team that undertakes the initial effort. • Survey data of health care entities such as nursing homes, hospitals, etc., on a particular topic, stripped of identifying information. • Our Center primarily uses administrative data from covered entities acquired through HIPAA waivers, the use of which is covered under the strict terms of data use agreements (DUAs). Any direct data sharing by the researcher is strictly prohibited. These include not only data from the Centers for Medicare and Medicaid Services (CMS), but electronic health records directly from nursing homes, for example. Given the complexity of using such data, it is forseeable that the ability to share analytic files that have gone through a measure of data cleaning and merging would be useful to the larger research community, in order to not only reproduce published findings and allow metadata analysts to more completely understand the characteristics of each individual analyses, but also to allow researchers to extend findings or use the cohort and set of data for a new purpose altogether. How this might be done is described below.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

There is no limit to the length of time these data should be made available for secondary research purposes. The first two types of data described above can be stored under existing digital repositories such as those maintained by many Universities that have already assumed the costs of maintaining such resources. The third, at least those data supplied initially by CMS, would require policy changes whereby CMS would either agree to store the final analytic files themselves for use by future researchers entering into their own DUAs, or else allow the researchers that created the files to give them to NIH to store and further distribute for ongoing research efforts, under the appropriate provisions, which may include a Systems of Records Notice (SORN). This would require new users to enter into a DUA with NIH. Along with the datafile itself, the original researcher could provide a list of specified steps with raw data counts that delineate the inclusion and exclusion criteria set in place to get from the starting raw data that the original researcher received to the analytic file(s) ultimately provided to NIH.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

There would be no real stewardship of the data stored on the digital repositories apart from the repositories maintaining the accessibility to the files themselves. Aside from the documentation originally provided by the original researcher to be stored with the data, no further follow-up with the original researchers would be necessary, nor feasible. Stewardship of the person level analytic datafiles derived from administrative data would necessarily lie with NIH (or in the case of CMS data, with CMS if they preferred to keep control of the distribution of the files). NIH (or CMS)

would be responsible for all costs associated with maintaining these files in a secure environment and developing a process for the redistribution of the data for future use (to include HIPAA waivers, DUAs, and security oversight) which may include the creation of an enclave that would allow researchers (with proper approvals in place) to work in a secure, NIH- or CMS-controlled environment.

4. Any other relevant issues respondents recognize as important for NIH to consider

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

b. Inclusion of a link to the data/software resource with the citation in the report

c. Identification of the authors of the Data/Software products

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

4. Any other relevant issues respondents recognize as important for NIH to consider

Additional Comments

Submission Date

01/17/2017

Submitter Name

David Carr

Name of Organization

Wellcome Trust

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

Biomedical sciences (broadly)

1. The highest-priority types of data to be shared and value in sharing such data

As a global research foundation, Wellcome considers that the priorities for data sharing are those types of data that have the greatest potential value for secondary use by others not connected to the original study team. Data that can be deposited with sufficient metadata to enable methodologically sound re-use or linkage to other datasets are the 'low-hanging fruit' that should be shared by default unless there are sound reasons not to. Genomic and clinical trial data are key priorities for sharing and we also consider that there is a particular imperative that research data collected in public health emergencies should be shared quickly (with appropriate safeguards for participant confidentiality) to enable rapid evidence-based clinical and public health responses. Data underpinning published research articles should also be considered a priority and, as a minimum, this data should be shared at the time of publication. There can be enormous value in promoting research data sharing. Some of the key advantages are: increasing the efficient use of expensive primary data; avoiding unnecessary duplication; increasing the statistical power of datasets; enabling novel and innovative research questions to be asked; increasing authors' scrutiny over their analyses; aiding reproducibility and speeding scientific advances.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

We are signatories to the UK Concordat on Open Research Data (http://www.rcuk.ac.uk/documents/documents/concordatonopenresearchdata-pdf/), which states that the data underlying publications should be made available for ten years after publication (Principle 8). Ensuring this principle can be adhered to requires strategic investment in data infrastructures and their maintenance. Careful thought needs to be given to what data should be retained long-term and what is unlikely to be of value to other users, especially where research generates vast file sizes such as in genomics. Research funders should jointly consider how best to sustain these data resources in the long-term, rather than taking a piece-meal approach to supporting ad hoc data management and maintenance for individual grants. If funders wish their research communities to take data sharing seriously, resources need to be costed both for the initial formatting, cleaning and providing adequate metadata to transform the data into a useable form, and for the longer term management and curation of the data. We do not consider there is an appropriate 'one size fits all' model for the long-term support for data infrastructures.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

There are substantial technical, logistical, legal and cultural barriers to data stewardship and sharing on both the supply and demand sides. Many of these are documented in the 2014 report of the UK Expert Advisory Group on Data Access report 'Establishing Incentives and Changing Cultures to support Data Access' (https://wellcome.ac.uk/sites/default/files/establishing-incentives-and-changing-cultures-to-support-data-access-eagda-may14.pdf) and although significant progress has been made in some fields, such as genomics and social science, many of the same challenges arise in different scientific domains. Key issues include: significant time and resource costs

required to format data; lack of adequate recognition for data generators; fears that making data available will allow authors to be 'scooped' or for their analyses to be refuted by others; lack of adequate storage and curation facilities; institutional legal complexities over establishing data sharing agreements; lead scientists retaining 'territorialism' over their data; and lack of a leadership culture that promotes and values data sharing. There are, however, many innovative ideas being generated to overcome these barriers, and over time these in turn could help shift scientific culture towards one of openness and recognising the value of data as a research output in its own right. It is important that funders take a leadership role in incentivising and rewarding data sharing, and work with the broader research community to drive this change in culture.

4. Any other relevant issues respondents recognize as important for NIH to consider

The recent upsurge of interest in the use of preprints in the life sciences – spearheaded by the ASAPbio initiative and a growing number of research funders who recognize the value of sharing research outputs as preprints – provides an opportunity to determine good data and software sharing practices for these types of outputs. Specifically, funders could determine that preprints can be cited in grant applications and end of grant reports but only if the preprint includes a data/software availability statement. Though such a change may be small, it may be sufficient to get researchers to start thinking about making their outputs accessible at an early stage.

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

We have no empirical evidence on this issue to date but support the intention behind increasing reporting and citation of data and software sharing in grant applications, as part of a broader drive to promote consistent approaches for data citation in the wider community. In the UK we are doing this through seeking to improve recognition of data sharing in the Research Excellence Framework, which allocates funding to public universities on the basis of an assessment of research quality. In our own guidance to applicants, we encourage the use of persistent identifiers for datasets and software outputs to enable credit to be given to data generators. We agree this is an area in which stronger and consistent guidance from funders might help drive change: encouraging grant holders to track and report sharing and re-use of data and software could help build the evidence base on the benefits of sharing these outputs, and provide a basis upon which these activities could receive due recognition and reward.

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

We would support the adoption of data and software citations in reports to NIH, in line with recognized good practices of which inclusion of a persistent identifier is key.

b. Inclusion of a link to the data/software resource with the citation in the report

We would support the inclusion of a link to the data/software resource within reports and published articles. By way of example the recently launched Wellcome Open Research platform (wellcomeopenresearch.org) specifically includes sections which require the authors to specify where data and software (referenced in the article) can be accessed from.

c. Identification of the authors of the Data/Software products

It is important that authors get credit for making data and software available to others – so we would support this. One way to facilitate this is through the adoption of the ORCID identifier. At Wellcome we require grant applicants to cite their ORCID in their application and are encouraging publishers to consider mandating this for peer reviewed publications. Again, for our own publishing platform (Wellcome Open Research) we have mandated that the corresponding author must cite their ORCID. By requiring ORCIDs – and encouraging publishers to include a data/software availability statement – it should be easier in the longer term for authors to get credit for their data and software outputs.

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

Alongside Wellcome and the Howard Hughes Medical Institute, the NIH supports the Open Science Prize (www.openscienceprize.org) and we believe this demonstrates a strong public commitment to promoting the value of data sharing and re-use. The innovative and exciting projects shortlisted for the inaugural prize also help showcase what can be achieved with better data sharing. Celebrating these projects provides the best possible advertisement for strengthening and incentivising others to begin shifting mindsets towards more and better sharing of data and software. Wellcome is keen to explore further opportunities for partnership with NIH and other funders to support innovative approaches to data sharing and open research.

4. Any other relevant issues respondents recognize as important for NIH to consider

The recent upsurge of interest in the use of preprints in the life sciences – spearheaded by the ASAPbio initiative and a growing number of research funders who recognize the value of sharing research outputs as preprints – provides an opportunity to determine good data and software sharing practices for these types of outputs. Specifically, funders could determine that preprints can be cited in grant applications and end of grant reports but only if the preprint includes a data/software availability statement. Though such a change may be small, it may be sufficient to get researchers to start thinking about making their outputs accessible at an early stage.

Additional Comments

Submission Date

01/17/2017

Submitter Name

Stephanie Marvin

Name of Organization

Brigham and Women's Hospital

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

big data sharing

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

Polysomnography cohorts, so researchers can further expand upon their knowledge of sleep medicine.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

Indefinately. For example Sleep Heart Health was collected a long time ago but is referenced and used in today's research.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

Keeping a website up to date. Having staff to help with issues with posted data.

4. Any other relevant issues respondents recognize as important for NIH to consider

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

b. Inclusion of a link to the data/software resource with the citation in the report

c. Identification of the authors of the Data/Software products

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

4. Any other relevant issues respondents recognize as important for NIH to consider

Additional Comments

Submission Date

01/18/2017

Submitter Name

Jeffery R Smith

Name of Organization

American Medical Informatics Association (AMIA)

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

Health Informatics, broadly defined

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

Please see attached document for AMIA's response to all questions.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

4. Any other relevant issues respondents recognize as important for NIH to consider

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

Please see attached document for AMIA's response to all questions.

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

b. Inclusion of a link to the data/software resource with the citation in the report

c. Identification of the authors of the Data/Software products

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

4. Any other relevant issues respondents recognize as important for NIH to consider

Additional Comments

AMIA Response to NIH Data Management, Sharing and Citation RFI_FINAL.pdf (296 KB)

January 18, 2017

Carrie D. Wolinetz, Ph.D.
Associate Director for Science Policy
Office of Science Policy
National Institutes of Health

Submitted electronically at: http://osp.od.nih.gov/content/nih-request-information-strategies-nih-data-management-sharing-and-citation

Re: Request for Information: Strategies for NIH Data Management, Sharing and Citation

Dear Dr. Wolinetz:

The American Medical Informatics Association (AMIA) appreciates the opportunity to submit comments regarding the National Institutes of Health's (NIH) request for information (RFI) on data management, sharing and citation. AMIA is the professional home for more than 5,400 informatics professionals, representing researchers, front-line clinicians, public health experts, and educators who bring meaning to data, manage information and generate new knowledge across the health and research enterprise.

AMIA enthusiastically supports development of policies for data management, sharing and citation. Recently, AMIA published the first in a series of Policy Principles & Positions.[1] Among them was an articulation of our belief that data sharing among researchers is foundational to advance scientific discovery, foster a culture of transparency, and improve reproducibility.[2]

In considering this RFI, AMIA members identified three key institutional incentives as necessary to improve data management and sharing: (1) Making data sharing plans scorable elements of applicable grants; (2) Financially supporting data curation and sharing; and (3) identifying ways to support academic advancement for scholars who create and/or contribute to useful public datasets and software.

First, AMIA recommends NIH make Data Sharing Plans a "scorable" element of grant applications subject to the existing policy.[3] Data sharing has become such an important proximal output of research that we believe the relative value of a proposed project should include consideration of how its data will be shared. Making data sharing plans scorable would enable those

[1] "AMIA Public Policy Principles and Policy Positions, 2016 – 2017," available at http://bit.ly/2gPB52N
[2] Ibid., Data Sharing in Research Policy Principle (pg. 10)
[3] National Institutes of Health, "NIH Data Sharing Policy and Implementation Guidance," March 2003 https://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm

January 18, 2017

projects that prioritize systematic and strategic data sharing, through use of standards and accepted best-practice, to garner higher scores. By using the peer-review process, we will make incremental improvements to interoperability, while identifying approaches to better data sharing practices over time. Expert and peer review of data sharing plans will lead to improved data sharing across the NIH portfolio, which will greatly improve interoperability, and research rigor, transparency, traceability, and reproducibility. An important component of this recommendation is funding.

Second, AMIA recommends NIH earmark support for data sharing as part of applicable grants' direct costs. In order for researchers to dedicate additional time and energy to produce (or collaborate) on development and execution of a quality data sharing plan, specified funding is needed. Mandating robust sharing plans, and elevating them to be scorable without corresponding funding would be counterproductive, and likely diminish the impact of such a policy.

Finally, AMIA recommends NIH identify ways to provide institutional awards for scholars who create and/or contribute to useful public datasets and software. We note that our academic reward system does not adequately recognize or provide incentives for those who create as well as those who analyze data.[4] NIH should look to scale platforms such as dbGaP, which enable investigators to receive points for others using their data, to other types of research.

Below we outline our recommendations in more detail, and we address NIH's specific questions related to this RFI. A group of AMIA members, listed in Appendix A, has provide detailed responses to this RFI in Table 1 of the enclosed document. These responses have been reviewed and duly approved by AMIA's Public Policy Committee and the AMIA Board of Directors. Should you have any questions or require additional information, please contact AMIA Vice President for Public Policy Jeffery Smith at jsmith@amia.org or (301) 657-1291 ext. 113. We, again, thank NIH for the opportunity to comment and look forward to continued dialogue.

Sincerely,

Douglas B. Fridsma, MD, PhD, FACP, FACMI
President and CEO
AMIA

Thomas H. Payne, MD, FACP, FACMI
AMIA Board Chair
Medical Director, IT Services, UW Medicine
University of Washington

Enclosed: Detailed AMIA Recommendations and Comments to NIH Questions

[4] Piwowar, H., Vision, T., "Data reuse and the open data citation advantage," *Peer J.* 2013. 1:e175

4720 Montgomery Lane, Suite 500 | Bethesda, Maryland 20814

January 18, 2017

Detailed AMIA Recommendations and Comments to NIH Questions

SECTION 1. Data Sharing Strategy Development

High-priority types of data to be shared

Any data necessary to the process of reproducing research are of high-value. While we understand this casts a wide net, we believe reproducibility and provenance are especially important in the basic sciences where investigators are using pre-clinical or other data to achieve the same results, in large epidemiological data sets used for population-based research, and data used in the review of pharmaceutical and medical devices.

We also note sharing is most important for data that would be expensive to re-create. However, if the data are inexpensive to generate, clear methodological instructions might be sufficient for re-creation, making sharing less important.

The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

In discussing length of time, AMIA members note a common tension concerns how quickly data should be shared, in addition to how long they should be public. We applaud NIH for finalizing policy, concerning how quickly data must be deposited in ClinicalTrails.gov,[5] and encourage it to harmonize similar requirements beyond the registry and results database.

AMIA recommends NIH endeavor to make publicly-funded research data available for secondary use for at least ten years after it is first published, and we recommend NIH develop policies for both unrestricted and controlled access. Data governance and permission policies should be an additional area of NIH focus. We note the NIMH Limited Access Datasets (LAD) project as an exemplar of defining access requirements. In whatever way NIH proposes to identify data retention and access policies, AMIA recommends NIH harmonize requirements across its Centers and Institutes.

Where possible, NIH should leverage existing and proven environments to maintain and sustain publicly-funded data through platforms such as Dryad,[6] Dataverse,[7] Cancer Imaging Data,[8] Figshare,[9] Zendo[10] and BioCADDIE.[11] Consistent with previous AMIA recommendations on

[5] 42 CFR Part 11. Clinical Trials Registration and Results Information Submission; Final Rule.
[6] http://datadryad.org
[7] https://dataverse.harvard.edu
[8] http://www.cancerimagingarchive.net
[9] https://figshare.com
[10] https://www.zenodo.org
[11] https://biocaddie.org

American Medical Informatics Association | AMIA
4720 Montgomery Lane, Suite 500 | Bethesda, Maryland 20814

January 18, 2017

digital data repositories, we support development of metrics to evaluate the quality and fit-for-purpose of various repositories.[12] Key among these metrics should be consideration of the repositories sustainability and/or business model. Should NIH designate an existing, independently operated repository, researchers depositing data need to be assured of their continued existence and availability. Additionally, NIH should prefer repositories that store the data in a non-proprietary (i.e. open) data format. Should a repository shutter or fail to meet its contractual obligations, it is important to protect the data from being "locked-in."

Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

AMIA supports dedicated funding from research sponsors for data curation and donation efforts so there are sufficient incentives to share, collaborate, and advance data sharing capabilities.[13] We recommend NIH earmark a percentage of grant funds for such activities as a way to overcome cost barriers. In combination with scoring data sharing plans (DSPs), explicitly setting aside funds to carry-out the DSP will improve data stewardship and sharing. Further, ensuring adherence to FAIR principles – Findability, Accessibility, Interoperability and Reusability – will help demonstrate value to overcome cost concerns.

Any other topics respondents recognize as important for NIH to consider

Additional aspects NIH may want to consider surround the role of participants in research data sharing. Specifically, NIH should ensure that data sharing policies are clearly articulated to both researchers and patients; that there are mechanisms for consent management; provisions of notification when data is used, and ways to share / return results in appropriate circumstances.

Additionally, NIH may wish to articulate expectations around pre-publication data management, including annotation, metadata, and provenance. For example, NIH should consider what documentation should be shared along with data that are necessary to support reuse, such as processes for transformation, imputation, coding, mapping standardization, data cleaning, and data quality assessments.

Finally, NIH should develop guidelines and best practices around data discoverability, including the use of model annotations, metadata schemas focused on a given domain (e.g. imaging) and minimal metadata expectations.

[12] AMIA Response to NIH RFI on Metrics to Assess Value of Biomedical Digital Repositories, October 5, 2016 available at http://bit.ly/2i2XF5a
[13] Borne, P., Lorsch, J., Green, E., "Perspective: Sustaining the big-data ecosystem," *Nature*. November 2015. 527, S16–S17

January 18, 2017

Section II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports (RPPR) and Grant Applications

The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

We see the utility in this kind of requirement; however, we reiterate our recommendation that there exist dedicated funding from research sponsors for data curation and donation efforts. Additionally, we recommend that reporting requirements be shared across venues (i.e. RPPR, publication in journals, etc.) with common guidance and metadata wherever possible. We further note that multiple approaches to point towards data, such as through data repository URLs, software source code hosting services, and DOIs, should be supported, and granularity issues should be supported in metadata models whenever possible.

Important features of technical guidance for data and software citation in reports to NIH, which may include:

- *Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI) (https://www.iso.org/obp/ui/#iso:std:iso:26324:ed-1:v1:en)*
- *Inclusion of a link to the data/software resource with the citation in the report*
- *Identification of the authors of the data/software products*
- *Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately*
- *Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed;*

AMIA strongly supports development of policy to cite data and software developed by grantees. As discussed previously, institutional incentives are needed to encourage development of reusable data / software. Data journals[14] are a step in the right direction, and development of persistent UIDs, such as DOIs, for data / software citation would be an important contribution to this effort. However, we note that such DOIs for re-use are new and nascent. We encourage NIH to fund specific projects to improve the use of DOIs for data / software, and we encourage NIH to explore how open source code and software containers, which represent snapshots of entire operating system configurations of computers used to develop software, can be leveraged to improve research rigor.

Inclusion of data sharing activities in scientists' career assessments is a potentially powerful means of incentivizing data sharing. Recent "Alt-metrics" efforts have begun to build the framework for

[14] http://www.nature.com/sdata/

tracking data reuse and citation as meaningful measurements of researcher contributions. As a means to help develop these policies and to further encourage data sharing, AMIA recommends NIH host a roundtable of academic health and science leaders, and produce a handbook for integrating this type of "credit" into promotion and tenure decisions.

Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications;

Again, making data sharing plans scorable aspects of pertinent grant applications would enable reviewers to assess the mechanisms through which data / software will be shared, and it would encourage more systematic, robust sharing strategies. One example for how NIH could operationalize the scoring of data sharing plans, could be to score according to priority data types (e.g. sharing data for autism or rare disease) and data quality / usability, similar to the 5-star deployment scheme for Open Data.[15]

Further, development of policy around Data Management Plans (DMPs), and examples of DMPs, would strengthen data and software sharing, especially towards to goals of traceability and reproducibility of research. According to research in progress, shared for purposes of AMIA's response to this RFI, a review of 67 data management and sharing plan requirements documents "uncovered inconsistent requirements for written DMPs as well as high variability in required or suggested DMP topics among funder requirements."[16] Further, "DMP requirements were found to emphasize post-publication data sharing rather than upstream activities that impact data quality, provide traceability or support reproducibility. With the emphasis equalized, the forty-three identified topics can aid Data Managers in systematically generating comprehensive DMPs that support project planning and application evaluation as well as data management conduct and post-publication data sharing."

Any other topics respondents recognize as important for NIH to consider.

We point NIH reviewers to real-time open science projects that have developed citable reference, which add to the diversity of solutions available. Through a project called Thinklab, researchers can get feedback on their grant proposals, participate in open peer review, and even lead entirely open research projects.[17] An example of an open science project developed using Thinklab is "Rephetio: Repurposing drugs on a hetnet," where each of the discussions that were part of that project is developed as a citable reference.[18,19] For example, https://thinklab.com/discussion/incorporating-

[15] http://5stardata.info/en/
[16] Zousus, M., Williams, M. "Data Management Plans, The Missing Perspective," available in a forthcoming issue of *Journal of Biomedical Informatics*
[17] https://thinklab.com/about
[18] https://thinklab.com/p/rephetio
[19] https://thinklab.com/p/rephetio/discussion

January 18, 2017

drugcentral-data-in-our-network/186 can be cited as "Daniel Himmelstein, Oleg Ursu, Mike Gilson, Pouya Khankhanian, Tudor Oprea (2016) Incorporating DrugCentral data in our network. *Thinklab*. doi:10.15363/thinklab.d186." While we acknowledge that such citations do not have the same rigor as peer-reviewed publication it may provide another avenue for recognition for researchers.

We also point NIH reviewers to the concept of "nanopublications" as means to disseminate individual data as independent publications with or without an accompanying research article.[20] Because nanopublications can be attributed and cited, they provide incentives for researchers to make their data available in standard formats that drive data accessibility and interoperability.

Increasingly, the research community is becoming aware that reproducibility requires that all software used to collect, transform and analyze the data must be publically available for inspection, modification, and reuse, along with the data.[21] We encourage NIH reviewers to become more sensitive to open source concepts relevant to ensuring scientific software practices support reproducibility such as the value of (1) using well-defined, standard open source licenses approved by the Open Source Initiative (OSI)[22] and (2) open source best practices of a public code repository (e.g., on GitHub[23] or Bitbucket[24]) and a public issue tracker.

Finally, any guidelines or requirements issued by NIH should consider *when* data must be shared. Timing requirements for data sharing may require tradeoffs between the goals of data originators hoping to retain exclusive access to data as needed to publish papers and those interested in prompt access to data for secondary use and replication. The NIH should develop models and policies designed to balance these potentially competing needs. Models such as pre-publication of protocols, as required by ClinicalTrials.gov, might be a partial solution. Alternatively, timeliness of data sharing efforts might be considered in the review of data management plans.

[20] http://nanopub.org/wordpress/
[21] Ince, D., et al., "The case for open computer programs." *Nature. February 2012*. **482**, S485-488. doi:10.1038/nature10836
[22] https://opensource.org/licenses
[23] https://github.com/
[24] https://bitbucket.org/

January 18, 2017

Appendix A: Response Team Members

These comments are official AMIA policy and endorsed by the full AMIA membership and Board of Directors, as evidenced by the President & CEO and Board Chair signatures. Below are the names of AMIA members who participated in the drafting and development of these policies recommendations.

Response Team: NIH RFI on Data Management, Sharing & Citation			
First	**Last**	**Organization**	**Email**
Tudor	Oprea	UNM School of Medicine	toprea@salud.unm.edu
Edwin	Young	Mount Sinai Health System	edwin.young@mountsinai.org
Michael	Cantor	NYULMC	michael.cantor@nyumc.org
Jorge	Caballero	Distal, Inc.	jorge@distal.co
Satyajeet	Raje	National Library of Medicine	satyajeet.raje@nih.gov
Harry	Hochheiser	U. Pittsbugh	harryh@pitt.edu
Ashish	Sharma	Biomedical Informatics, Emory Univ.	ashish.sharma@emory.edu
Rashmi	Mishra	NIDCR	rashmi.mishra2@nih.gov
Eileen	Healy	NIH	healye@slu.edu
Brian	Fung	Mayo Clinic	fung.brian@mayo.edu
Tod	Yates	Blue Cross Blue Shield AZ Advantage	tod.yates@azbluemedicare.com
Meredith	Zozus	Univ. of Arkansas for Medical Sciences	mzozus@uams.edu
Ken	Goodman	University of Miami	kgoodman@med.miami.edu
Leon	Rozenblit	Prometheus Research, LLC	leon@prometheusresearch.com
Jorge	Ferrer	VA	Jorge.Ferrer@va.gov
Carolyn	Petersen	Mayo Clinic	Petersen.Carolyn@mayo.edu

Submission Date

01/18/2017

Submitter Name

Rebecca Boyles

Name of Organization

RTI International

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

Public Health

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

The scientific potential of unshared data sets remains largely unknown; however, studies that have captured variables and outcomes can be used to understand complex contributors to disease, such as environmental exposures. From a practical view, high-priority data are those which are reputable, high quality, and well-documented. From these two perspectives, small studies can be as valuable as larger, more powered studies.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

To echo concerns stated above, data should be maintained in relation to its value for secondary research. The value of data is extremely difficult to access, except that poorly documented data sets, or those of poor quality, are less valuable. Resources to maintain and sustain these should be minimized. While no one seeks to create useless data, there are many examples in noteworthy repositories of data that is not reusable. Resources would be more efficiently allocated to ensure that future data is of high value (i.e., FAIR). In our experience, this will require an investment in a data infrastructure that can easily be adapted into researchers current workflows. Furthermore, education and outreach to encourage adoption of these tools cannot be overlooked.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

Central to the discussion of data sharing is data management. This needs to be a funded component of any study. Investment in the creation and adoption of reusable tools to enable good data management throughout the lifecycle are wise investments that can create efficiencies. While there are many components to consider in establishing a data management infrastructure that can support innovative data reuse and analysis, Common Data Elements (CDEs) can provide a critical building block to a broader data infrastructure. Notably, CDEs address many of the barriers to data sharing and require little change in current research practices. For both the data producer and consumer, CDEs clarify the data collection method, reducing the risk of misinterpretation. By providing standard documentation they reduce the time required for data documentation and preparation. Adoption of CDEs increases the value of repositories by enhancing the value of their data assets and decreases the time often required for data cleaning and documentation.

4. Any other relevant issues respondents recognize as important for NIH to consider

Increasingly the public is engaging in research as participants and expects to be able to access resources created through their participation. Much of the present conversation has centered around data sharing among researchers, but in the near future a broader audience will seek data access. It is essential for the validity of the biomedical research endeavor that steps be taken to ensure data quality and reuse. A lack of investment to support a data reuse infrastructure that includes sustainable repositories, CDE efforts, data standards, and adoption/communication efforts could undermine the future of research support and funding.

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

Increased reporting requirements are a welcome step. Sharing is not sufficient unless steps are taken to ensure the data being shared is reusable (i.e., FAIR). Documentation is critical.

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

PUIs are welcome. However, sharing is not sufficient unless steps are taken to ensure the data being shared is reusable (i.e., FAIR). Documentation is critical.

b. Inclusion of a link to the data/software resource with the citation in the report

c. Identification of the authors of the Data/Software products

Data and software are scholarly contributions that are only increasing in importance. Authors who contribute to high-quality, reusable data and software resources need to be acknowledged to create the appropriate incentive structure that will enable meaningful data sharing in the future.

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

4. Any other relevant issues respondents recognize as important for NIH to consider

Increasingly the public is engaging in research as participants and expects to be able to access resources created through their participation. Much of the present conversation has centered around data sharing among researchers, but in the near future a broader audience will seek data access. It is essential for the validity of the biomedical research endeavor that steps be taken to ensure data quality and reuse. A lack of investment to support a data reuse infrastructure that includes sustainable repositories, CDE efforts, data standards, and adoption/communication efforts could undermine the future of research support and funding.

Additional Comments

Submission Date

01/18/2017

Submitter Name

Shirely Y. Hill, Ph.D.

Name of Organization

University of Pittsburgh School of Medicine

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

Psychiatry -- Family Studies Imaging and Genetics

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

4. Any other relevant issues respondents recognize as important for NIH to consider

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

b. Inclusion of a link to the data/software resource with the citation in the report

c. Identification of the authors of the Data/Software products

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

4. Any other relevant issues respondents recognize as important for NIH to consider

Additional Comments

REQUEST FOR INFORMATION_1_18_17.pdf (26 KB)

Request for Information:

The highest-priority types of data to be shared and the value in sharing such data.

Response: Currently there is an underdeveloped set of guidelines pertaining to the potential risks and associated costs of sharing data of particular types. Answering the question of what types of data should be of highest priority requires a better understanding of the parameters surrounding data sharing. While there may be some benefit in sharing of specific types of de-identified data, the risk associated with re-identification has not been addressed in depth, nor policies developed and circulated to researchers who will provide the data. Moreover, because one of the largest group of stake holders are the principal investigators of NIH grants who will be needed to insure the data is usable and can remain de-identified, their place at the table where ongoing NIH policies are developed should be increased in number.

Cost/Benefit Analysis: Costs: The potential costs include: (1) lost productivity, (2) fiscal costs, and (3) loss of privacy for research participants. (4) The potential for loss of privacy and confidentiality is not only an issue of utmost ethical importance for researchers, but may impact the decisions of potential research participants to take part in research studies by sharing their private information. This is especially true where potentially stigmatizing conditions such as substance use disorders are revealed for the participant and his/her family members. *Decisions regarding NIH requirements for sharing of data should include a cost/benefit analysis.*

(1) Lost Productivity: The costs include lost time from the principal Investigators' plans for collecting, analyzing and disseminating new data that they have agreed to perform in their peer-reviewed funded grant proposals in order to prepare older data for sharing. Overseeing the preparation of data for sharing is an enormous burden on researchers who are already limited in available time to perform the research because of increased regulatory requirements, and reduced NIH funding over the past few years that has required resubmission of many more grant applications than in previous years.

(2) Fiscal Cost: Data that has already been collected for internal use may not meet the requirements needed for sharing with others outside one's own research group. This may require extensive recoding of the data along with detailed documentation. Many researchers have informally discussed concern that the costs for covering data base managers needed to make this transition have not been present in the NIH plans for PIs being asked to share. While it is recognized that NIH has described the need for researchers to outline the costs of preparing data for sharing in new applications, there is an enormous amount of data that has already been collected for which there is no mechanism for data preparation. To date, costs have been discussed regarding additional monies needed at the NIH level to manage the shared data bases.

(3) Loss of Privacy and Confidentiality (LPC): An extended discussion of possible risks to human participants who have volunteered their private information is urgently needed. This is especially true for sensitive data (HIV status, illegal drug use). The recognition that two pieces of demographic data (birth date and zipcode) can often uniquely identify an individual brings up the question of whether any data can truly be deidentified (See Sweeney, 2000) which is further compounded by the ease with which medical data can be accessed through arrangements made in 33 states to sell such data (Sweeney, 2015). Not only should researchers uphold the principles of the Common Rule and guarantee the promises they have made to volunteers who have entrusted the researchers with their most private information, but the NIH should ensure this as well. If public perception is that participating in research has the potential for loss of privacy, then the willingness for human volunteers to participate broadly in research will be harmed.

(3a) LPC and Individual Data versus Family Data: Consideration should be given to whether the data to be shared is for one individual from a family or for multiple individuals' data collected from the same family. Sharing of data from a single individual from a family will provide

less risk for re-identification than providing an entire family. However, genetic analyses require having the sex, age, and family relationship to be of use. Where pedigrees are extensive, individuals from a pedigree can recognize their own family particularly if data is being submitted from a single city rather from a multi-site collaboration. Sharing of pedigree data makes it possible for a family member to uncover the disease status of other family members. This may be less stressful in the case of common medical conditions such as diabetes but can be devastating to families with addiction. Even if all members consent to participate and are interviewed regarding their own addiction they should be confident that this will not be revealed to other participating members. An additional level of concern is present for individuals within a family who are not consented and do not participate yet have their private information put at risk. This occurs when one's family is identified as addiction prone because of recognition that one belongs to a particular family. This can also occur where participating family members provide family history data on the non-participating member. This concern was emphasized by the Office for Protection from Research Risks (OPRR) (The Scientist, 2000) in a review of research conducted at Virginia Commonwealth University in which a family member of consenting individual objected to his information being used without his consent. OPRR noted that "Researchers spending federal tax dollars should diligently consider ethics in their work". Although OPRR's opinion was troubling to some genetic researchers fearing they could not obtain needed information on family clustering of diseases, subsequent experience has proven that ethical principles can be upheld without loss of research value. As noted by Botkin (2001) the Belmont report has noted that a fundamental principle of contemporary research ethics is that participating in research should not place anyone at risk of harm without an explicit informed consent dialogue between the researcher and participant.

(3b) LPC, Sensitive Data, and its Implications: Historically, NIH has been concerned about the implications of identifying an individual as having an addiction problem or being prone to addiction. In its Executive Summary, the Committee on Vaccines Against Drugs of Abuse noted in regard to use of newer immunotherapies that "Enthusiasm for the new medications should not obscure the fact that fully informed and voluntary consent is necessary under any and all circumstances. These medications can produce long-lasting biological markers (raising issues of confidentiality and potential for discrimination)... " In my view, the potential for re-identification as a result of data sharing should be of equal concern.

(4) Loss of Public Trust and Willingness of Individuals to Participate in Research: Prospective Versus Retrospectively Collected Data Requiring Sharing: Currently NIH does not have a policy that distinguishes between the requirement for researchers to share data they agree to share in new NIH grant applications developed after the White House directive and data acquired prior to the implementation of the NIH sharing policy. It is suggested that procedures can be put in place for new data to be acquired after the initiation of the NIH policy that may limit the risks to research participants. However, it is of great concern that individuals who participated in data collection prior to the initiation of this policy had the full expectation that their data would not be shared without reconsenting. Obtaining new consents for persons participating more than a decade ago is impractical. Policies need to be developed to clearly define those instances where data should be excluded from the NIH sharing requirement to protect the privacy and confidentiality of former research subjects and to lessen the burden on researchers to attempt to locate and reconsent such participants.

Benefits of Sharing: There is no question that science builds on the sharing of discoveries across research groups. At some level, there is no question that the tax-paying public deserves to receive the benefits of the research that is paid for by NIH funds. Steps that have been taken to date requiring researchers to deposit statistically analyzed data in form of peer reviewed manuscripts accepted for publication in medical and scientific journals is a valuable development improving the accessibility of findings to the general public and to researchers in countries where

university libraries may not have the funds to subscribe to a vast array of medical and scientific journals.

Limitations of Accessible Data: Even when all of the costs are taken into consideration and deemed to be acceptable, there are major limitations to what the general public can derive from raw data that has not been statistically analyzed, a discussion presented offering strengths and limitations of the data set and the scientific/medical implications of the results. Offering raw data to untrained citizens may lead to erroneous conclusions that could be potentially harmful. For example, access to raw data from clinical trial information could lead an individual to conclude that a particular drug is or is not beneficial causing them to change their plan to comply with a given treatment prescribed by their physician. Researchers trained in experimental design, statistics and health-related information are in the best position to weigh the results offered by raw data. This information should be peer reviewed and published in medical/scientific journals and then offered to the public.

Researchers Secondary Analysis of Data: It may be argued that the "general public" includes researchers wishing to engage in secondary analysis of someone else's data should have free access to that data. The argument goes that perhaps these individuals will find something new that the original researcher who collected the data will not have thought of in the initial analysis. While this may be true, the possibility of secondary analysis has always been present when researchers informally contact each other to suggest a project for analysis. This is the ideal way to mine the data because the original researcher can participate in the analysis and dissemination of the findings. Every data set has nuances that are known only to the person designing, collecting and analyzing the data. No amount of written "data documentation" that is submitted with the accompanying data to a large scale NIH resource can provide this. Ongoing collaboration between the data originator and the secondary analyzer insures that questions that arise as the data is analyzed are answered adequately. Without participation of the original researcher erroneous conclusions may be drawn which ultimately sets back scientific/medical progress rather than enhancing it.

Incentives for Good Stewardship: Why should researchers put their creativity energy and an enormous amount of their professional lives into collecting data, especially investing many years of their labor to provide stewardship of longitudinal studies and their associated data, if it is to be mined by others before they have an opportunity to publish their ideas. Individual incentives to collect and maintain databases, particularly longitudinal databases are greatly reduced if the researcher has limited opportunity to utilize the fruits of their labors. This is not to say that data should not ultimately be shared, or even shared in ongoing manner with collaborators who make a request to the PI. Rather, the principal investigators who originally design a study, recruit participants and keep them actively engaged over many years, and who provide many hours of administrative management (e.g., progress reports, renewal applications, meeting regulatory requirements) while simultaneously managing and maintaining these complex data sets should have the opportunity to mine the data first. The US government can own the data they have paid for with NIH funds while at the same time providing a lifetime lease, or alternatively, a 20 year post collection window lease, as is allowed by patent law, to the Principal Investigator who originates the data. This would insure their commitment to maintain the databases they have created and insure that secondary users of the data are not inappropriately handling the data either statistically or ethically.

Conclusions: In order to comply with the White House memorandum for the Heads of Executive Departments and Agencies of February 22, 2013 entitled "Increasing Access to the Results of Federally Funded Scientific Research", the NIH has developed policies that are outlined in eight elements. Among these are "(b) a strategy for improving the public's ability to locate and access digital data resulting from federally funded research .... (d) a plan for notifying awardees and other federally funded scientific researchers of their obligations (e.g., through guidance,

conditions of awards, and /or regulatory changes)." These policies have required researchers to immediately deposit raw data with the NIH or risk loss of funding for their ongoing research. A public discussion with adequate representation from the research community and from bioethicists concerning the risks and benefits of this plan was not forthcoming before implementation of the plan. Now, we are faced with a situation where researchers who wish to uphold ethical practices of not comprising the privacy and confidentiality of their research participant's data may not be able to do so. This situation is particularly concerning where participants signed consent forms that did not include sharing of their data. Hearings should be conducted and/or panels convened to study the question of data sharing further and to develop a clear set of policies with prominent participation of NIH funded researchers, experts in data privacy (Sweeney, 2015), and by legal/ethical scholars (Rothstein, 2015; Harrell and Rothstein,2016).

As noted by Mark A. Rothstein,JD (2015) "Despite unique aspects, such as data sources, scale and open access provisions, the ethical issues surrounding Big Data are similar to those involving traditional biomedical research. Without question, the regulation of research can be improved in many ways. The development of new data analytic tools, however, such as Big Data, should not serve as a catalyst for abandoning foundational principles of research ethics." The issues are complex and deserve further study. As noted by Harrell and Rothstein (2016) "Research and privacy laws are applicable to both biobanks and researchers. The laws are often complicated and confusing, and therefore determining which laws are even applicable to biobanks can be particularly difficult."

1, L. Sweeney, Simple Demographics Often Identify People Uniquely. Carnegie Mellon University, Data Privacy Working Paper 3. Pittsburgh 2000.

2. Sweeney L. Only You, Your Doctor, and Many Others May Know. Technology Science. 2015092903. September 29, 2015. http://techscience.org/a/2015092903.

3. Botkin JR. Protecting the privacy of family members in survey and pedigree research. JAMA, 285:207-2011, 2001.

4. Amber D. Case at VCU Brings Ethics to Forefront. The Scientist, 14, 1, 2000.

5. New Treatments for Addiction: Committee on Immunotherapies and Sustained-Release Formulations for Treating Drug Addiction: Panel formed by NIDA/NIH to study the Behavioral, Ethical, Legal, and Social Questions. Henrick J. Harwood and Tracy G. Myers, *Editors*: National Research Council and the Institute of Medicine of the National Academies. The National Academies Press, Washington, D.C. **www.nap.edu**

6. Rothstein MA. Ethical Issues in Big Data Health Research. Journal of Law, Medicine and Ethics, 425-429, 2015.

7. Harrell HL and Rothstein MA. Biobanking Research and Privacy Laws in the United States. Journal of Law, Medicine and Ethics, 106-227, 2016.

Submission Date

01/19/2017

Submitter Name

Iain Hrynaszkiewicz

Name of Organization

Springer Nature

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

All research domains

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

As a publisher of research, primarily via scholarly journals and books, data that support peer-reviewed publications are important to share, to enable reuse of research by the research community and independent replication and verification of results. For more specific types of research data and disciplines, priority might be given to research data that are difficult to generate and hard to recreate, such as data from human research subjects and rare or vulnerable species. Data important to public health, such as in response to a global epidemic, are also important to share rapidly. Similarly, data relevant to public policy and/or with high social impact might be viewed as high priority. Publishers have responded to public health emergencies by making research articles freely available. However publishers - which provide services for the research community - are generally not well placed to set the priorities for the types of data that should be shared but should support the needs of the research community in their policies and services.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

The length of time research data remains useful to the research community is best determined by the research community and its funding agencies, and these expectations should be supported by publisher policies and services. Several UK Research Councils' data policies stipulate that "research data is securely preserved for a minimum of 10 years from the date that any researcher 'privileged access' period expires or, if others have accessed the data, from last date on which access to the data was requested by a third party" (e.g. https://www.epsrc.ac.uk/about/standards/researchdata/expectations/) While we can publish small datasets (up to 10-20Mb) as electronic supplementary materials in journals, a requirement of all Springer Nature journal data policies is the preference for archiving of research data in repositories. Public datasets supporting publications should be preserved indefinitely (whether in repositories or journals) to maintain the integrity of the published record. Community specific data repositories are preferred and general/institutional repositories can also be used. Springer Nature manages a list of more than 80 recommended repositories across all research domains (http://www.springernature.com/gp/group/data-policy/repositories) including health sciences (http://www.springernature.com/gp/group/data-policy/repositories#c10106444). Our criteria for approving repositories (http://www.nature.com/uploads/ckeditor/attachments/3243/SciData_respository_evaluation_Aug2016.docx) include ensuring long-term preservation of datasets. In practice this means sustainability plans and data preservation for a minimum of 10 years.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

Barriers to data sharing reported by researchers include: 1. Copyright and licensing 2. Data standards 3. Uncertainty about compliance with funder policy 4. Lack of time 5. Perceived lack of credit 6. Lack of a data repository 7. Uncertainty about covering costs 8. Concerns over inappropriate data reuse 9. Protecting human research participant privacy Mechanisms to overcome them (taking each of the above stated barriers in turn): 1. Data repositories and publishers

with clear terms of use for data. 2. Promote community standards (e.g. https://biosharing.org/standards/) and provide for researcher training 3. Funder policy compliance advisory services (e.g. Springer Nature's Research Data Support helpdesk http://www.springernature.com/gp/group/data-policy/helpdesk) 4. Appoint or promote involvement of research data management experts in research, and recommend services/products for researchers to provide data management/sharing services, including those which might already be available commercially 5. Encourage formal data citation and quality/time stamping e.g. badges for open practices 6. Promote data repositories including generalist repositories (http://blogs.nature.com/scientificdata/2016/11/14/expanding-our-generalist-data-repository-options/) 7. Provide a proportion of grant funding for research data management and sharing (e.g. 5%) 8. Promote scholarly norms of citation and provenance/evidence when research data are reused (https://www.biomedcentral.com/about/policies/open-data) 9. For sensitive data, establish data use agreements (DUAs) in partnership with specialist repositories (see http://researchintegrityjournal.biomedcentral.com/articles/10.1186/s41073-016-0015-6 for a list of these repositories) and provide resources for anonymisation of datasets for sharing; and modify participant consent procedures accordingly (https://trialsjournal.biomedcentral.com/articles/10.1186/1745-6215-11-9).

4. Any other relevant issues respondents recognize as important for NIH to consider

We are happy to be contacted for more information on this submitted response at researchdata@springernature.com or iain.hrynaszkiewicz@nature.com

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

Firstly, a likely impact will be creating cultural change amongst researchers and research assessment procedures to increase the recognition of data and software as legitimate, - shareable and citable - scholarly outputs. The NSF in 2014 began asking for research "products" rather than papers in principal investigators' bibliographic sketches (http://www.nature.com/nature/journal/v493/n7431/full/493159a.html). Secondly, reporting of data and software - particularly if done via persistent identification mechanisms - will help improve reproducibility and provenance tracking of research outputs. It will, also, help in the assigning of credit for researchers' outputs that are not papers/publications. Thirdly, reuse and assessment of data and software might be encouraged, improving scholarly discourse and return on investment in research, which might help also increase the reliability of scientific research funded by the NIH. Finally, funding agencies encouraging or requiring the citation of data and software in RPPRs (and grant applications) might help make this practice more widespread in other scholarly literature.

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

For many researchers citation of research data and, even more so, research software is a new concept. It is important, therefore, that data citation with DOIs is communicated as a simple and easily achievable practice that is supported by publishers. Citing datasets and software can be done in much the same way as citing papers, provided the necessary information (metadata) about the data and software are available. Supporting these basic principles of data citation are mandatory parts of all Springer Nature's standard research data policies: "Datasets that are assigned digital object identifiers (DOIs) by a data repository may be cited in the reference list. Data citations should include the minimum information recommended by DataCite: authors, title, publisher (repository name), identifier." (http://www.springernature.com/gp/group/data-policy/policy-types#c10305772) There are other considerations in implementing effective data citation, from the more technical perspective of publishers and repositories and persistent identifier issuing bodies. A working group to define a roadmap for implementing data citation (co-chaired by Springer Nature https://www.force11.org/group/dcip/eg3publisherearlyadopters) is working to enable consistent implementation of data citation by publishers. For researchers who are more advanced with the practice of data citation there are standards emerging for the citation of dynamic objects and citation of data with additional metadata that infers more meaning to citations (e.g. data generated by a research study or data referenced/analysed by a study).

b. Inclusion of a link to the data/software resource with the citation in the report

Availability of research data and software are increasingly reported in research papers in a designated section, often called a "Data availability statement", "Code availability" statement or "Availability of data and materials" statement. This approach has several benefits: - Makes links to research data and software easier to find in publications - Supports the requirements of some funder policies, such as the UK Research Councils, which often require data availability statements in publications - They complement data citation in reference lists by providing a narrative to describe availability of data and software Springer Nature's research data policies support and provide detailed guidance - and examples - for writing data availability statements. For more information see http://www.nature.com/authors/policies/data/data-availability-statements-FAQs.pdf and http://www.springernature.com/gp/group/data-policy/data-availability-statements Many Springer Nature journals including the Nature and BioMed Central journals have policies on the sharing and reporting of software/code (e.g. http://www.nature.com/news/code-share-1.16232). Funding and research organisations and publishers should collaborate in the development of standards and policies for data/software citation, via specific fora within the Research Data Alliance (https://www.rd-alliance.org/groups/data-policy-standardisation-and-implementation) and Force 11 (https://www.force11.org/group/dcip/eg3publisherearlyadopters) communities, for example. Publishing technology can also support bidirectional linking between publications and repositories, such as embedding data viewers in publications, and initiatives such as Scholix (http://www.scholix.org/) to make data-article link exchange more widespread and interoperable.

c. Identification of the authors of the Data/Software products

The policy of Springer Nature's BioMed Central journals and of the journal Scientific Data do not just encourage (or, at some of these journals, require) the sharing of software but furthermore encourage the deposition of software in repositories that can assign persistent identifiers (DOIs) to software to enable citation and provenance tracking e.g. Scientific Data policy (http://www.nature.com/sdata/policies/editorial-and-publishing-policies#code-avail): "authors are encouraged to archive their code in a public repository that can assign it a DOI, such as figshare...we recommend using an open control version system (CVS), such as GitHub, in combination with a DOI providing repository...Code with an assigned DOI may be formally cited and listed in the References section of the manuscript." BioMed Central policy (http://www.biomedcentral.com/getpublished/editorial-policies#availability+of+data+and+materials): "include a link to the most recent version of your software or code (e.g. GitHub or Sourceforge) as well as a link to the archived version referenced in the manuscript. The software or code should be archived in an appropriate repository with a DOI or other unique identifier. For software in GitHub, we recommend using Zenodo." By encouraging sharing and citation of software via repositories that provide DOIs, as a consequence authors must be identified as part of deposition.

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

In our experience as an early adopter of data citation principles and practices, simplicity is very important in communicating data citation expectations. The journal Scientific Data recently shared guidance on data citation. A basic principle of data citation is to "Cite what you used" (http://blogs.nature.com/scientificdata/2016/07/14/data-citations-at-scientific-data/). If a researcher/author used associated datasets, especially data archived outside of a journal article and its supplementary material, then they should cite the data. Often it will be appropriate to cite both: the paper and any datasets used. The focus of Springer Nature's research data policies is on citing datasets that are assigned DOIs in reference lists (e.g. Nature's policy: www.nature.com/authors/policies/data/data-availability-statements-data-citations.pdf), but other types of dataset identifier can also be cited - either as in-text references or links or as formal citations. Very data focused journals, such as Scientific Data, that support the most stringent of Springer Nature's data policies, require any stably archived datasets mentioned in a publication to be formally cited with its persistent identifier, regardless of identifier type. In terms of what data should be reported and cited, this will vary by research community. The research data policies of Springer Nature in general concern the minimal dataset that supports the central findings of a published study (http://www.springernature.com/gp/group/data-policy/faq).

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

As described previously, following the standards of DataCite and including authors, title, publisher (repository name), identifier in citations of/references to data and software is important. It is also important that those citing data and software resources do not "invent" metadata. If for example "authors" or "title" for data or software are not clear from the information available from the source/repository, they should not be included. In our experience titles of data and software can be more difficult to report uniformly (http://blogs.nature.com/scientificdata/2016/07/14/data-citations-at-scientific-data/).

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

Possibilities include: 1. Encouraging or requiring the reporting and citation of data and software in scholarly publications (journal articles, monographs, preprints etc) as well as RPPRs 2. Require provision of consistent, standardised data and software availability/accessibility statements in RPPRs and publications 3. Monitor and encourage inclusion of information on data and software reuse in RPPRs: a)This could take the form of metrics such as "pull requests" for software in Github and citations and downloads of datasets, where this information is available from repositories. b) This could also be achieved anecdotally, by researchers providing case studies and examples of data reuse, such as the number of requests to share data they received. 4. Commit to working with publishers and other stakeholders to share information on data-article links and discuss policy standardisation for example via the Scholix framework (https://blogs.openaire.eu/?p=1589) and Research Data Alliance (https://www.rd-alliance.org/groups/data-policy-standardisation-and-implementation), respectively 5. Encouraging researchers to, where appropriate, publish data papers and software papers in journals to promote reuse of data and software and submit their data and software for consideration by peer reviewers of traditional papers 6. Encourage NIH funded repositories that provide accession IDs, to standardize data citation guidance for researchers in collaboration with publishers. http://identifiers.org/ may be a useful resource for this.

4. Any other relevant issues respondents recognize as important for NIH to consider

We are happy to be contacted for more information on this submitted response at researchdata@springernature.com or iain.hrynaszkiewicz@nature.com

Additional Comments

Submission Date

01/19/2017

Submitter Name

Ara Tahmassian

Name of Organization

Harvard University

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

All areas of biomedical research

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

4. Any other relevant issues respondents recognize as important for NIH to consider

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

b. Inclusion of a link to the data/software resource with the citation in the report

c. Identification of the authors of the Data/Software products

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

4. Any other relevant issues respondents recognize as important for NIH to consider

Additional Comments

HarvardUniversity_response_to_NIH_NOT-OD-17-015.pdf (103 KB)

We are responding to RFI NOT-OD-17-015 on behalf of Harvard University. We represent members of the Harvard University Data Group, which is charged with planning and executing the University's approach to storage and dissemination of data in the natural and social sciences. Several of us are co-PIs on NIH Grants in which large scale data production and sharing are requirements. Others of us play key roles on behalf of the University in maintaining data for NIH-funded programs: Ara Tahmassian (Harvard University Chief Compliance Officer); Mercè Crosas (Chief Data Science and Technology Officer at Harvard's IQSS, Force11 Data and Software Citation Principles working groups and FAIR DMP working group, BioCaddie,); Caroline Shamu (NIH LINCS Program U54 HL127365; U54HL127624-03S2); Rainer Fuchs (Chief Information Officer, Harvard Medical School); James Cuff (Assistant Dean for Research Computing, Harvard University); Rebecca Li, Kristen Bolt, and Barbara Beirer (Multi-Regional Clinical Trials Center of Harvard and Brigham and Women's Hospital). Jason Johnson (Chief Health Information Officer, Dana-Farber Cancer Center) has also joined as a signatory representing Dana-Farber.

Section I Data Sharing Strategy and Development

Comment on the length of time data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications:

There is no single timeframe relevant to all data types and disciplines that can be specified *a priori* as the appropriate length of time for which datasets should be made available. It must be recognized that, for any field, there will be datasets of different intrinsic value and longevity to the community. There could be some data, such as human genome sequences, that should probably be retained in perpetuity.

We advocate that producers and users of different types of data develop their own specific recommendations for their datasets, perhaps with review by a central NIH entity that includes external and internal individuals. We also note that datasets can still be accessible even if the modes of storage on which they are maintained have different retrieval times. In the extreme case, this could involve physical transfer of a disc (e.g. Amazon Snowball). Datasets determined by the community to be extremely useful—frequently re-used and cited—would in this model be retained for longer periods of time on high-end storage systems with fast retrieval times. Datasets that are used less often over time might move to less expensive storage tiers with slower retrieval times. However, in these cases or in cases in which the data must be deleted, if the dataset was used as part of a published finding, we recommend that a metadata record describing the dataset lives in perpetuity in a public repository (as the so called "dataset landing page"), so that it can continue being referenced even if the data files themselves are not quickly accessible.

Comment on barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers:

We are interested in seeing how the new Commons Cloud Credits program develops and is implemented by our NIH-funded investigators. We believe that credits should also be useful for university-based data storage and computation in the program as

well, although this would probably require compliance with accepted standards for reliability and longevity. Currently, the Conformant Cloud program appears to be geared exclusively to commercial providers of cloud data services. Harvard investigators have found that for many applications, use of local storage/compute resources at our university are cheaper and more effective than use of commercial cloud services. Thus, for now, it is important that they have access to both commercial and local storage/computing resources to carry out their work. The Cloud Credits program should recognize this need and it should also encourage (non-commercial) inter-university computation and data storage open cloud solutions (solutions built with open-source technologies such as OpenStack; an example is the Massachusetts Open Cloud).

A very substantial, and under-recognized barrier to data sharing is the effort required for curation of datasets—collection and documentation of metadata and protocols relevant to reagents, experiments, and data analysis, as well as dataset formatting and annotation. These activities are essential not only for re-use of data but for reproducibility too. From our experience in the NIH LINCS program, 15-20% of our total effort is put into curating and formatting datasets so that they are documented sufficiently for integration with other datasets and for use in many software tools. It is our experience that the incentives for high quality curation of data at the level of publication are nearly non-existent. Thus, as data sharing becomes required for all NIH projects, it is essential that adequate funds, incentives and software tools be developed to support appropriate curation of datasets. Resources to automate the data curation process are essential—for example databases of registered reagents such as the Resource Identification Portal (https://scicrunch.org/resources). Furthermore, we recommend that criteria similar to the set proposed by the FAIR principles (findable, accessible, interoperable, and reusable data) are followed to ensure high-quality curation for discovery and reusability. In addition, there could be a reward system associated with well-curated datasets. For example, a metadata rating or badge that would indicate whether a dataset is well-described.

NIH should also support mechanisms that make it easier for PIs to fulfill data publication requirements. At HMS, we routinely get requests from labs to help them set up data publication mechanisms such as websites or FTP servers. This patchwork approach is highly inefficient and places an undue burden especially on small labs that don't have much experience in that space or that have limited access to advanced computational resources at their institutions. There are major databases such as Genbank or PDB for certain data types. Some publishers provide mechanisms for data sharing. However, we believe there is ample opportunity to provide additional – and more effective – data sharing mechanisms for the dissemination of any type of research data. We encourage NIH to provide more direct support for repositories of research data such as Dataverse and help streamline options available for its grant recipients.

Section II Inclusion of Data and Software Citation in NIH RPPR and Grant Applications

Much work has been done by community-organized working groups on the topic of data and software citation, many of them led by the Force11 scholarly communication organization. We recommend that NIH consider the Joint Declaration of Data Citation Principles, and the Software Data Citation as guidance for data and software citation, including the recently published roadmap for data repositories on how to implement the data citation principles. The roadmap describes in the detail the use of persistent identifiers assigned to a dataset, a dataset landing page, and the minimal set of metadata fields needed for universal discovery (for example, using schema.org metadata to describe a dataset). As recommended by the Joint Declaration of Data Citation Principles, DOIs or other community accepted persistent identifiers (for example, handles), should be assigned to a dataset to provide a permanent link from a publication to the data; dataset metadata should include the publications(s) DOI(s) that are associated with that dataset.

Incentives for data sharing

We propose that NIH consider a system of recognition whereby data generators are identified through designation as "data authors." These "Data authors" would be recognized in any publications derived from the original dataset and cited in Medline, as well as searchable through the National Library of Medicine (and other search engines).

We envision that Data authors would be responsible for the integrity and curation of their data, but would not be responsible for the conclusions of the secondary analyses of the secondary authors. The outcome of such a system would be that high quality, usable datasets would be cited more commonly and reflected on an individual's curriculum vitae. Over time this system would incentivize data sharing through defining and building a system that recognizes and incentivizes quality (e.g. whether the data has been peer-reviewed, what data standards if any have been utilized, degree and completeness of metadata, number of times data has been reused, etc.).

There may also be ways to provide financial incentives that NIH may want to consider, such as a small add-on to base budgets in later years if grant recipients can provide clear evidence of effective data sharing. Note: this should NOT be implemented by holding back parts of the base budget as this would be a negative reinforcement method less likely to lead to a sustainable behavior shift.

Submission Date

01/19/2017

Submitter Name

Carmen Nitsche

Name of Organization

Pistoia Alliance

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

Across biopharmaceutical industry research and development

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

4. Any other relevant issues respondents recognize as important for NIH to consider

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

b. Inclusion of a link to the data/software resource with the citation in the report

c. Identification of the authors of the Data/Software products

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

4. Any other relevant issues respondents recognize as important for NIH to consider

Additional Comments

Submission Date

01/19/2017

Submitter Name

Mary Ellen Davis, Exec Dir

Name of Organization

Association of College and Research Libraries

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

Our nearly 11,000 members come from academic and research libraries of all types and partner with a

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

ACRL is the higher education association for librarians. Representing nearly 11,000 academic and research librarians and interested individuals, ACRL (a division of the American Library Association) develops programs, products and services to help academic and research librarians learn, innovate, and lead within the academic community. 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 OSTP 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. Too often, the data and articles resulting from research remains locked behind paywalls or siloed in proprietary computer systems. In order to unleash the power of this information and truly accelerate discovery, we need to ensure that research outputs are made immediately available to the global public, and that people are fully empowered to use it in new and innovative ways. At present, data underlying publications are of high priority. Sharing the data underlying publications in open and machine-readable formats through trust-worthy repositories (e.g., DSA/WDS certified, http://www.datasealofapproval.org/en/news-and-events/news/2016/11/25/wds-and-dsa-announce-uni-ed-requirements-core-cert/) should meet the minimal level of data sharing requirements. These data should include the associated documentation necessary for access, interoperability and reusability, such as data dictionaries, code and computational details. Ideally, the workflow associated with the project should be made available in a format that encourages reproducibility.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

To some extent, the question of how long the data should be accessible will depend on the area of research and the characteristics of the resulting data. As such, these time frames should be community-driven determinations. There are important implications for sustainability in this recommendation. It would be regrettable to see a situation develop where NIH-funded investigators must in future purchase access to data that was produced through NIH-funded research. Ultimately, a viable solution may require a creative consortial approach to maintaining long-term access to these data resources.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

Currently, cultural norms and perceived cost are significant barriers to data stewardship and sharing. While multiple approaches to this issue may increase odds of success, a significant factor is the expectation of the funder. In order to help shift researcher behavior, funder incentives, support, and requirements that encourage good data management and sharing practices is required. For example, grants using established standards for metadata, and best practices for documentation (both which facilitate the reuse of data) should receive continued funding as they offer evidence or responsible management of research data.

4. Any other relevant issues respondents recognize as important for NIH to consider

As noted in Fenner et. al (2016) scholarly data repositories, which are often within the remit of the University Library, deal with issues of data and software citation on a regular basis. We encourage NIH to reach out to members of this community who have developed practical expertise in these areas, and to consider librarians as active partners in their efforts to implement effective data and software citation. ACRL is happy to work with NIH as a bridge to the academic and research library community, helping to build effective collaborations and partnerships between communities. On behalf of the Association of College and Research Libraries, I urge you to seriously consider these recommendations so that the NIH can increase the re-use of data created through its funding. If you have any questions about these recommendations, please do not hesitate to reach out to me at mdavis@ala.org or 312-280-3248. Sincerely, Mary Ellen K. Davis ACRL Executive Director

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

No comment.

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

ACRL encourages NIH to sign on to the FORCE11 Data Citation Principles: https://www.force11.org/group/joint-declaration-data-citation-principles-final. Data citation is an important component of data sharing and incentivizing the practice. Additionally, NIH should explore the work available on issues of data citation implementation as presented in Starr, et al. (2015), https://doi.org/10.7717/peerj-cs.1.

b. Inclusion of a link to the data/software resource with the citation in the report

Barring instances when the data are protected, this should be a requirement. In keeping with the recommendations in 2a on implementing data citation, a persistent URL should be associated with a persistent identifier. This link should be part of the citation, and should resolve to a landing page specific to the resource (Starr et al., 2015; Fenner et al., 2016, http://dx.doi.org/10.1101/097196). This practice enables the findability that should be associated with all data/software resources and permits the gatekeeping that may be necessary in the case of restricted resources.

c. Identification of the authors of the Data/Software products

In line with recommendations from FORCE11, it is critical that data/software creators be identified and credited for their work. This is one crucial step towards larger cultural changes within disciplines towards recognizing data/software as scholarly products in their own right.

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

This is an evolving topic of discussion as the systems that host these data evolve and become more sophisticated. Documentation that clearly explains the level of granularity is one way to allow for diverse practices, and in the absence of an agreed upon standard, is necessary. Research fields or communities may also have their own requirements for granularity unique to their research needs.

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

We encourage NIH to recognize institutional and subject repositories as acceptable sharing platforms for data. Recognize that repositories that align with ISO 16363 provide a level of trustworthiness that is ideal.

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

While multiple approaches to this issue may increase odds of success, a significant factor is the expectation of the funder. In order to help shift researcher behaviour, funder incentives, support, and requirements that encourage good documentation and sharing practices that make the foundational findings of research findable, accessible, interoperable and reusable is required. A possible incentive could be funding projects that are focused on data and software reuse, or otherwise rewarding projects that produce data and software that meet the highest levels of effective sharing as demonstrated by their degree of reuse.

4. Any other relevant issues respondents recognize as important for NIH to consider

As noted in Fenner et. al (2016) scholarly data repositories, which are often within the remit of the University Library, deal with issues of data and software citation on a regular basis. We encourage NIH to reach out to members of this community who have developed practical expertise in these areas, and to consider librarians as active partners in their efforts to implement effective data and software citation. ACRL is happy to work with NIH as a bridge to the academic and research library community, helping to build effective collaborations and partnerships between communities. On behalf of the Association of College and Research Libraries, I urge you to seriously consider these recommendations so that the NIH can increase the re-use of data created through its funding. If you have any questions about these recommendations, please do not hesitate to reach out to me at mdavis@ala.org or 312-280-3248. Sincerely, Mary Ellen K. Davis ACRL Executive Director

Additional Comments

ACRL input to NIH on data sharing.pdf (418 KB)

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

TO: National Institutes of Health
DATE: Thursday, January 19, 2017
RE: Request for Information on Strategies for NIH Data Management, Sharing, and Citation

Submitted online at http://osp.od.nih.gov/content/nih-request-information-strategies-nih-data-management-sharing-and-citation

To Whom It May Concern,

On behalf of the Association of College and Research Libraries (ACRL), I am writing to offer comments on supporting data management, sharing, and citation.

Section I. Data Sharing Strategy Development

NIH recognizes that many factors must be considered when determining what, when, and how data should be managed and shared. These factors include, for example, the purpose for sharing, supporting data re-use and reproducibility, maturity of the science, the infrastructure uniqueness of the data, and ethical considerations.
The NIH seeks comment on any or all of the following topics to help formulate strategic approaches to prioritizing its data management and sharing activities:

1. The highest-priority types of data to be shared and value in sharing such data (Maximum: 250 words)
The Association of College & Research Libraries is the higher education association for librarians. Representing nearly 11,000 academic and research librarians and interested individuals, ACRL (a division of the American Library Association) develops programs, products and services to help academic and research librarians learn, innovate, and lead within the academic community. 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. Too often, the data and articles resulting from research remains locked behind paywalls or siloed in proprietary computer systems. In order to unleash the power of this information and truly accelerate discovery, we need to ensure that research outputs are made immediately available to the global public, and that people are fully empowered to use it in new and innovative ways.

At present, data underlying publications are of high priority. Sharing the data underlying publications in open and machine-readable formats through trust-worthy repositories (e.g., DSA/WDS certified, http://www.datasealofapproval.org/en/news-and-events/news/2016/11/25/wds-and-dsa-announce-uni-ed-requirements-core-cert/) should meet the minimal level of data sharing

requirements. These data should include the associated documentation necessary for access, interoperability and reusability, such as data dictionaries, code and computational details. Ideally, the workflow associated with the project should be made available in a format that encourages reproducibility.

2. **The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications** (Maximum: 250 words)
To some extent, the question of how long the data should be accessible will depend on the area of research and the characteristics of the resulting data. As such, these time frames should be community-driven determinations.

There are important implications for sustainability in this recommendation. It would be regrettable to see a situation develop where NIH-funded investigators must in future purchase access to data that was produced through NIH-funded research. Ultimately, a viable solution may require a creative consortial approach to maintaining long-term access to these data resources.

3. **Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers** (Maximum: 250 words)
Currently, cultural norms and perceived cost are significant barriers to data stewardship and sharing. While multiple approaches to this issue may increase odds of success, a significant factor is the expectation of the funder. In order to help shift researcher behavior, funder incentives, support, and requirements that encourage good data management and sharing practices is required. For example, grants using established standards for metadata, and best practices for documentation (both which facilitate the reuse of data) should receive continued funding as they offer evidence or responsible management of research data.

4. **Any other relevant issues respondents recognize as important for NIH to consider** (Maximum words: 250)
ACRL encourages NIH to continue to recognize and promote librarians as partners and sources of expertise with respect to the documentation, organization, preservation, stewardship, and curation of data.

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications
Currently, NIH grantees are required to report "other products of the research," including data, databases, and software, in section C5a of their annual RPPR submission (http://grants.nih.gov/grants/rppr/rppr_instruction_guide.pdf). However, limited guidance is available on how data, databases, and software should be reported or cited.
NIH recognizes that data and software citation indicates proof of productivity that translates to publications and patents. More thorough reporting of data and software products in the RPPR and in Competitive Grant Renewal applications may strengthen documentation of productivity and may also identify projects and investigators who most effectively share data and software.

The NIH seeks comment on any or all of the following topics:

1. **The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing** (Maximum words: 250)
No comment.

2. **Important features of technical guidance for data and software citation in reports to NIH, which may include**:

a. **Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)** * (Maximum words: 250)
ACRL encourages NIH to sign on to the FORCE11 Data Citation Principles: https://www.force11.org/group/joint-declaration-data-citation-principles-final. Data citation is an important component of data sharing and incentivizing the practice. Additionally, NIH should explore the work available on issues of data citation implementation as presented in Starr, et al. (2015), https://doi.org/10.7717/peerj-cs.1.

* (DOI: https://www.iso.org/obp/ui/#iso:std:iso:26324:ed-1:v1:en)

b. **Inclusion of a link to the data/software resource with the citation in the report** (Maximum: 250 words)
Barring instances when the data are protected, this should be a requirement. In keeping with the recommendations in 2a on implementing data citation, a persistent URL should be associated with a persistent identifier. This link should be part of the citation, and should resolve to a landing page specific to the resource (Starr et al., 2015; Fenner et al., 2016, http://dx.doi.org/10.1101/097196). This practice enables the findability that should be associated with all data/software resources and permits the gatekeeping that may be necessary in the case of restricted resources.

c. **Identification of the authors of the Data/Software products** (Maximum: 250 words)
In line with recommendations from FORCE11, it is critical that data/software creators be identified and credited for their work. This is one crucial step towards larger cultural changes within disciplines towards recognizing data/software as scholarly products in their own right.

d. **Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately** (Maximum words: 250)
This is an evolving topic of discussion as the systems that host these data evolve and become more sophisticated. Documentation that clearly explains the level of granularity is one way to allow for diverse practices, and in the absence of an agreed upon standard, is necessary. Research fields or communities may also have their own requirements for granularity unique to their research needs.

e. **Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed** (Maximum words: 250)
We encourage NIH to recognize institutional and subject repositories as acceptable sharing platforms for data. Recognize that repositories that align with ISO 16363 provide a level of trustworthiness that is ideal.

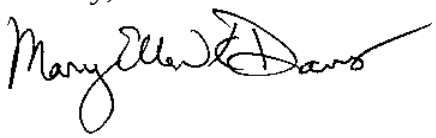
3. **Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications**
(Maximum: 250 words)
While multiple approaches to this issue may increase odds of success, a significant factor is the expectation of the funder. In order to help shift researcher behaviour, funder incentives, support, and requirements that encourage good documentation and sharing practices that make the foundational findings of research findable, accessible, interoperable and reusable is required.

A possible incentive could be funding projects that are focused on data and software reuse, or otherwise rewarding projects that produce data and software that meet the highest levels of effective sharing as demonstrated by their degree of reuse.

4. **Any other relevant issues respondents recognize as important for NIH to consider**
(Maximum: 250 words)
As noted in Fenner et. al (2016) scholarly data repositories, which are often within the remit of the University Library, deal with issues of data and software citation on a regular basis. We encourage NIH to reach out to members of this community who have developed practical expertise in these areas, and to consider librarians as active partners in their efforts to implement effective data and software citation. ACRL is happy to work with NIH as a bridge to the academic and research library community, helping to build effective collaborations and partnerships between communities.

On behalf of the Association of College and Research Libraries, I urge you to seriously consider these recommendations so that the NIH can increase the re-use of data created through its funding. If you have any questions about these recommendations, please do not hesitate to reach out to me at mdavis@ala.org or 312-280-3248.

Sincerely,



Mary Ellen K. Davis
ACRL Executive Director

Submission Date

01/19/2017

Submitter Name

Wendy Pradt Lougee

Name of Organization

University of Minnesota

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

all disciplines

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

When considering which data should be considered high priority for sharing and preserving, we recommend considering the re-use potential of the data as well as the potential of data to support verification of research findings, especially findings that impact public policies and health care. Additional value should be placed on data that are peer-reviewed through re-use or verification efforts. We encourage data that have considerable public impact to be shared widely and in manners that are accessible to the general public, such as in public use repositories indexed by search engines, and in formats that are consumable by general audiences. Data that are poorly documented are not worth the time and resources necessary for sharing and long-term access. Therefore, any high priority data must also be well curated. Curation (intentionally documenting, describing, and transforming data for long term access and reuse) is a necessary part of any sharing or preservation plan. A minimum level of documentation should be encouraged for others to successfully and ethically reuse shared or preserved data. We stress the importance of recognizing the legal and ethical requirements to protect sensitive or private individually identifiable data, especially as it is possible to re-identify participants through combinations of indirect identifiers in the data or from other sources that can be linked to the data (such as public repositories, on websites, or on social media). High priority data will need to be de-identified to remove as much risk to participants as possible before sharing broadly.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

We encourage NIH to establish a minimum guideline for how long data that support a publication or were collected as part of a grant should be preserved. Guidelines should consider disciplinary norms and practices, as well as potential reuse and public health implications of the data. Guidelines we use for local institutional data repository are to preserve data for at least 10 years and establish other criteria for longer preservation periods in accordance with University policies on research data management and record retention. We would advise against blanket statements for one size fits all preservation standards as they will depend on the purpose of the data, the data files themselves (file types, data format), and the institution/repository where they are stored. Policies surrounding data preservation and sharing for intramural researchers may have been previously established. Requirements for extramural researchers should mirror these policies whenever appropriate. Such groups without policies should be encouraged to develop policies appropriate to their specific data and usage. Because it is costly to preserve all data that are generated through a grant, NIH should encourage researchers to describe which pieces of data are going to be preserved and which will be appropriately disposed of (or not intentionally preserved) after the grant. Longer or indefinite time periods should be considered for preservation of data directly supporting claims in published research articles and data that would be difficult to re-collect, such as clinical data on rare conditions. NIH should provide guidance about appropriate data disposal and deaccessioning.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

Many university libraries have long-standing expertise and services that can help alleviate the burdens of data stewardship for individual researchers, including resources for data management training, documentation, metadata, and institutional repositories for sharing. These services take the form of broad trainings, individual or lab consultations, or partnerships on research projects. While these services are often provided at no cost for researchers, they come with significant infrastructure, training, and time costs for libraries, especially as services scale to meet growing needs for more specialized outreach across a growing number of disciplines. NIH should recognize the investment of time, energy, and resources to provide these services, as well as to curate and maintain access to datasets over time. Researchers should be encouraged to account for those costs in grant applications, whether from their institutional libraries or elsewhere. NIH should also encourage any newly-launched grant-funded repositories to have a business plan to ensure sustainability over time. To help meet the time costs of curation, our institution relies on a team of disciplinary experts to review and curate submitted datasets. We also serve as the lead institution on the Data Curation Network project (https://sites.google.com/site/datacurationnetwork), a Sloan-funded initiative to establish an inter-institutional network of disciplinary experts for the purpose of scaling data curation across universities nationwide. We would like to see budget and timeline guidance from NIH for typical cost, labor, and time required to prepare data for submission to a repository so applicants can include it in grants.

4. Any other relevant issues respondents recognize as important for NIH to consider

To further enhance discoverability of data that underlie articles and encourage validation of research findings, it is important to ensure data registries interoperate with PubMed publications and resources like clinicaltrials.gov. Requirements and standards for well structured metadata in both data and article repositories can make this possible. It is important to make researchers aware of the rights and requirements tied to specific datasets. NIH should recommend researchers share data with clear licenses (such as creative commons or other open-source licenses) and data use agreements, as applicable. The use of well-curated and certified data repositories, with staff that review and add substantive, machine-readable metadata to submissions, should also be encouraged. Emerging certifications (e.g., the Data Seal of Approval and TRAC) may assist. Data should be shared in formats that can ensure the ability to render and interact with the data over the long-term, such as non-proprietary and open source formats. Similarly, for verification and reproducibility purposes, NIH should encourage researchers to think of interoperability broadly, and to document computational resources in the generation of data and data formats. Wendy Pradt Lougee University Librarian McKnight Presidential Professor University of Minnesota Libraries

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

Compliance with data management and sharing requirements can be more easily facilitated through integration with RPPRs and grant applications. Despite implementation of the public access policy in 2008, NIH saw relatively low compliance rates continue until this policy was more directly connected to RPPRs. Following this, NIH saw aggregate submissions increase from an average of 5,158 articles per month in 2012 to 7,931 articles per month in 2013 according to Monthly Aggregate Submission Statistics. Positioning these items as research outputs would align with recent revisions to NIH biosketch requirements which allow researchers to include other types of scholarly outputs, including data sets, as evidence of their expertise.

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

Persistent Unique Identifiers are only useful if they remain persistent. DOIs and other persistent links should only be provided by reputable repositories (according to standards, such as the Data Seal of Approval) with policies for how they will ensure the sustainability of the links.

b. Inclusion of a link to the data/software resource with the citation in the report

Links to data/software resources should also include documentation about the computing environment and versions of the software used in creation/analysis.

c. Identification of the authors of the Data/Software products

We encourage use of ORCiD IDs (Open Researcher and Contributor ID) and other ways to uniquely identify authors.

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

With regard to granularity of data citations, while there are advantages in providing citations to each specific dataset in a project, it may be unduly burdensome to researchers to cite each individual piece of a larger study, especially as the academic norms are for citing a study, rather than the individual pieces (such as the methods or results section). We encourage NIH to provide guidance for researchers in justifying their rationale for their dissemination strategy. We also encourage researchers to add links to related articles, datasets, or grants as part of the metadata of a dataset in a repository.

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

We encourage inclusion of the repository as the publisher or distributor of the dataset in the citation. This is recommended by the International Association for Social Science Information Services & Technology (IASSIST): http://www.iassistdata.org/sites/default/files/quick_guide_to_data_citation_high-res_printer-ready.pdf

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

We encourage the NIH to value citations to well curated data submissions in all sections of the NIH biosketch and when considering the overall impact of a researcher. NIH should encourage researchers to cite both grants and datasets to enhance reporting and discoverability.

4. Any other relevant issues respondents recognize as important for NIH to consider

To further enhance discoverability of data that underlie articles and encourage validation of research findings, it is important to ensure data registries interoperate with PubMed publications and resources like clinicaltrials.gov. Requirements and standards for well structured metadata in both data and article repositories can make this possible. It is important to make researchers aware of the rights and requirements tied to specific datasets. NIH should recommend researchers share data with clear licenses (such as creative commons or other open-source licenses) and data use agreements, as applicable. The use of well-curated and certified data repositories, with staff that review and add substantive, machine-readable metadata to submissions, should also be encouraged. Emerging certifications (e.g., the Data Seal of Approval and TRAC) may assist. Data should be shared in formats that can ensure the ability to render and interact with the data over the long-term, such as non-proprietary and open source formats. Similarly, for verification and reproducibility purposes, NIH should encourage researchers to think of interoperability broadly, and to document computational resources in the generation of data and data formats. Wendy Pradt Lougee University Librarian McKnight Presidential Professor University of Minnesota Libraries

Additional Comments

Minnesota_NIHRFIResponse_January2017_0.pdf (104 KB)

Response to Request for Information: "Strategies for NIH Data Management, Sharing, and Citation," November 2016

January 17, 2017

Wendy Pradt Lougee
University Librarian
McKnight Presidential Professor
University of Minnesota Libraries

Thank you for the opportunity to comment on "Strategies for NIH Data Management, Sharing, and Citation." These comments are submitted on behalf of the University of Minnesota Libraries. The University of Minnesota is one of the leading public research institutions in the United States, and a key contributor to the entrepreneurial economy of the state of Minnesota, as well as to scholarship both nationally and internationally. As the value of a dataset depends on its quality, we strongly advocate for a policy that emphasizes curation as an instrumental part of the sharing and preservation process, and that recognizes the importance of discipline specific and institutional repositories in this space. We also encourage NIH to develop mechanisms to monitor compliance and recognize the time and resources associated with the responsible stewardship of data.

SECTION I. Data Sharing Strategy Development

**Replies for each section can not exceed 250 words*

(1) The highest-priority types of data to be shared and value in sharing such data

Comment 1:

- When considering which data should be considered high priority for sharing and preserving, we recommend considering the re-use potential of the data as well as the potential of data to support verification of research findings, especially findings that impact public policies and health care. Additional value should be placed on data that are peer-reviewed through re-use or verification efforts.
- We encourage data that have considerable public impact to be shared widely and in manners that are accessible to the general public, such as in public use repositories indexed by search engines, and in formats that are consumable by general audiences.
- Data that are poorly documented are not worth the time and resources necessary for sharing and long-term access. Therefore, any high priority data must also be well curated. Curation (intentionally documenting, describing, and transforming data for long term access and reuse) is a necessary part of any sharing or preservation plan. A minimum level of documentation should be encouraged for others to successfully and ethically reuse shared or preserved data.
- We stress the importance of recognizing the legal and ethical requirements to protect sensitive or private individually identifiable data, especially as it is possible to re-identify participants through combinations of indirect identifiers in the data or from other sources that can be linked to the data

(such as public repositories, on websites, or on social media). High priority data will need to be de-identified to remove as much risk to participants as possible before sharing broadly.

(2) The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

Comment 2:

- We encourage NIH to establish a minimum guideline for how long data that support a publication or were collected as part of a grant should be preserved. Guidelines should consider disciplinary norms and practices, as well as potential reuse and public health implications of the data. Guidelines we use for local institutional data repository are to preserve data for at least 10 years and establish other criteria for longer preservation periods in accordance with University policies on research data management and record retention.
- We would advise against blanket statements for one size fits all preservation standards as they will depend on the purpose of the data, the data files themselves (file types, data format), and the institution/repository where they are stored.
- Policies surrounding data preservation and sharing for intramural researchers may have been previously established. Requirements for extramural researchers should mirror these policies whenever appropriate. Such groups without policies should be encouraged to develop policies appropriate to their specific data and usage.
- Because it is costly to preserve all data that are generated through a grant, NIH should encourage researchers to describe which pieces of data are going to be preserved and which will be appropriately disposed of (or not intentionally preserved) after the grant. Longer or indefinite time periods should be considered for preservation of data directly supporting claims in published research articles and data that would be difficult to re-collect, such as clinical data on rare conditions. NIH should provide guidance about appropriate data disposal and deaccessioning.

(3) Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

Comment 3:

- Many university libraries have long-standing expertise and services that can help alleviate the burdens of data stewardship for individual researchers, including resources for data management training, documentation, metadata, and institutional repositories for sharing. These services take the form of broad trainings, individual or lab consultations, or partnerships on research projects. While these services are often provided at no cost for researchers, they come with significant infrastructure, training, and time costs for libraries, especially as services scale to meet growing needs for more specialized outreach across a growing number of disciplines.
- NIH should recognize the investment of time, energy, and resources to provide these services, as well as to curate and maintain access to datasets over time. Researchers should be encouraged to account for those costs in grant applications, whether from their institutional libraries or elsewhere.
- NIH should also encourage any newly-launched grant-funded repositories to have a business plan to ensure sustainability over time.
- To help meet the time costs of curation, our institution relies on a team of disciplinary experts to review and curate submitted datasets. We also serve as the lead institution on the Data Curation Network project (https://sites.google.com/site/datacurationnetwork), a Sloan-funded initiative to establish an inter-institutional network of disciplinary experts for the purpose of scaling data curation across universities nationwide.

- We would like to see budget and timeline guidance from NIH for typical cost, labor, and time required to prepare data for submission to a repository so applicants can include it in grants.

(4) Any other topics respondents recognize as important for NIH to consider.

Comment 4:

- Researchers should be provided with sufficient training and resources to appropriately de-identify datasets. Because the utility of some datasets may be compromised through de-identification efforts, NIH should allow for mediated sharing, such as restricted access or inclusion in a data catalog to facilitate discoverability while protecting participants.
- NIH should provide guidance to researchers for writing language in IRB protocols and Informed Consent documents that allow future sharing and preservation of the data while protecting the confidentiality of participant information. There are existing guidelines for writing non-restrictive informed consents: https://www.icpsr.umich.edu/icpsrweb/content/datamanagement/confidentiality/conf-language.html
- When recommending repositories, disciplinary specialization should be balanced with broad discoverability of data. Disciplinary repositories account for specialized needs of unique data types, and are essential in meeting the data preservation, security, and discovery needs of their communities. Likewise, institutional repositories (e.g., the Data Repository for the University of Minnesota, http://hdl.handle.net/11299/166578) are important resources for researchers in fields which lack a disciplinary data repository, and are often a source of non-published data, code, and software. NIH should emphasize the utility of these diverse repositories, and create common registries (building off existing efforts such as COAR and SHARE) to aid broad discoverability across repositories.
- NIH should provide guidance on the use of different types of repositories throughout the research lifecycle, including those meant for collaboration and version control, such as GitHub and the Open Science Framework (OSF), versus those intended for final archival storage.

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports (RPPR) and Grant Applications

(1) *The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing;*

Comment 1:

- Compliance with data management and sharing requirements can be more easily facilitated through integration with RPPRs and grant applications. Despite implementation of the public access policy in 2008, NIH saw relatively low compliance rates continue until this policy was more directly connected to RPPRs. Following this, NIH saw aggregate submissions increase from an average of 5,158 articles per month in 2012 to 7,931 articles per month in 2013 according to Monthly Aggregate Submission Statistics.
- Positioning these items as research outputs would align with recent revisions to NIH biosketch requirements which allow researchers to include other types of scholarly outputs, including data sets, as evidence of their expertise.

(2) Important features of technical guidance for data and software citation in reports to NIH, which may include:

(a) *Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI) (https://www.iso.org/obp/ui/#iso:std:iso:26324:ed-1:v1:en);*

Comment (a): Persistent Unique Identifiers are only useful if they remain persistent. DOIs and other persistent links should only be provided by reputable repositories (according to standards, such as the Data Seal of Approval) with policies for how they will ensure the sustainability of the links.

(b) *Inclusion of a link to the data/software resource with the citation in the report;*

Comment (b): Links to data/software resources should also include documentation about the computing environment and versions of the software used in creation/analysis.

(c) Identification of the authors of the data/software products;

Comment (c): We encourage use of ORCiD IDs (Open Researcher and Contributor ID) and other ways to uniquely identify authors.

(d) Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately;

Comment (d): With regard to granularity of data citations, while there are advantages in providing citations to each specific dataset in a project, it may be unduly burdensome to researchers to cite each individual piece of a larger study, especially as the academic norms are for citing a study, rather than the individual pieces (such as the methods or results section). We encourage NIH to provide guidance for researchers in justifying their rationale for their dissemination strategy. We also encourage researchers to add links to related articles, datasets, or grants as part of the metadata of a dataset in a repository.

(e) Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed;

Comment (e): We encourage inclusion of the repository as the publisher or distributor of the dataset in the citation. This is recommended by the International Association for Social Science Information Services & Technology (IASSIST): http://www.iassistdata.org/sites/default/files/quick_guide_to_data_citation_high-res_printer-ready.pdf

(3) Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications;

Comment 3:

- We encourage the NIH to value citations to well curated data submissions in all sections of the NIH biosketch and when considering the overall impact of a researcher.

- NIH should encourage researchers to cite both grants and datasets to enhance reporting and discoverability.

(4) Any other topics respondents recognize as important for NIH to consider.

Comment 4:

- To further enhance discoverability of data that underlie articles and encourage validation of research findings, it is important to ensure data registries interoperate with PubMed publications and resources like clinicaltrials.gov. Requirements and standards for well structured metadata in both data and article repositories can make this possible.
- It is important to make researchers aware of the rights and requirements tied to specific datasets. NIH should recommend researchers share data with clear licenses (such as creative commons or other open-source licenses) and data use agreements, as applicable.
- The use of well-curated and certified data repositories, with staff that review and add substantive, machine-readable metadata to submissions, should also be encouraged. Emerging certifications (e.g., the Data Seal of Approval and TRAC) may assist.
- Data should be shared in formats that can ensure the ability to render and interact with the data over the long-term, such as non-proprietary and open source formats. Similarly, for verification and reproducibility purposes, NIH should encourage researchers to think of interoperability broadly, and to document computational resources in the generation of data and data formats.

Submission Date

01/19/2017

Submitter Name

Chuck Cook

Name of Organization

EMBL-European Bioinformatics Institute

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

Bioinformatics

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

4. Any other relevant issues respondents recognize as important for NIH to consider

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

b. Inclusion of a link to the data/software resource with the citation in the report

c. Identification of the authors of the Data/Software products

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

4. Any other relevant issues respondents recognize as important for NIH to consider

Additional Comments

EMBL-EBI_response_NOT-OD-17-015.pdf (679 KB)

EMBL-EBI

EMBL-EBI response to Notice Number: NOT-OD-17-015: NIH Request for Information (RFI) on Strategies for NIH Data Management, Sharing, and Citation

EMBL-EBI is one of the largest providers of bioinformatics data and services in the world. Our core mission is to archive and make freely available scientific data to all facets of the scientific community.

Profile of EMBL-EBI

The European Bioinformatics Institute (EMBL-EBI), part of the European Molecular Biology Laboratory (EMBL), is the premier European centre for data resources and research in bioinformatics. EMBL-EBI data resources cover the entire range of biological sciences from raw DNA sequences to curated proteins, chemicals, structures, systems, pathways, ontologies and literature. EMBL-EBI's mission to provide open and free biological data and tools to the entire scientific community is supported by world-class expert staff and decades of experience in serving open data.

EMBL-EBI resources vary from archival databases that contain research data outputs, such as DNA sequences, to highly dynamic knowledgebases that aggregate, process and visualize research data, often adding layers of value through manual curation by highly qualified personnel. Fundamental tenets of this mission are that all hosted data, tools and infrastructure are freely available worldwide, and that data are represented in and shared in a variety of structured and standard formats for consumption by both people and machines. Free availability, ease of access, and multiple access points are all critically important to maximize the utility and re-use of scientific data. EMBL-EBI's service mission is user- focused, and this is reflected in our approach to assessing the value of our data resources, in which we consider foremost their value to users.

We comment below on the topics identified in the RFI.

Data sharing strategy development

EMBL has supported data sharing for over three decades, and founded the European Bioinformatics Institute to support the sharing and open distribution of scientific research data. Open access data resources, such as those at EMBL-EBI, provide tremendous value to scientists, to funders, and to the public by making available for re-use and analysis research results from scientists worldwide. EMBL-EBI strongly supports the NIH effort to promote sustained and systematic use and sharing of research data.

The highest-priority types of data to be shared and the value in sharing such data

Scientific data are key research outputs that result from investments, both public and private, in research. Public and charitable funders require, and desire, that research results are made publicly accessible and citable to enable other researchers to confirm published results, and to enable re-use of data, which maximizes the value of the original research investments. EMBL-EBI provides a stable, long-term home for research data, and follows FAIR[1] guidelines for storage, findability, and accessibility of data.

EMBL-EBI hosts a very wide range of data resources describing nucleotide sequences, proteins, chemicals, small molecules, metabolomics, proteomics, and literature. The importance of open access resources for researchers has demonstrated through analysis of citations to data within those resources[2], and re-use of scientific data provides tremendous economic benefit for researchers and for society.

In 2015 EMBL-EBI commissioned an independent economic analysis of the value our resources provide to the scientific community[3]. The analysis was based on information gained in a large user survey and used a range of quantitative and qualitative analytical methods to build a picture of both perceived user value and wider economic returns. Direct value was measured in the time, and therefore costs, spent by users accessing EMBL-EBI data resources. Also quantified was the hypothetical value, or contingent value, of what users might be willing to pay for access to the resources. Contingent valuation is a common method used for the valuation public goods.

Wider benefits and impacts were explored by assessing user efficiency gains and assigning economic value to them. These included the value of time saving (productivity) and the avoidance of costs for users that would otherwise be incurred in the creation /collection of the data. To demonstrate the return on investment (ROI) that the resources provide, a macro-economic model (modified Solow-Swan), used indicative R&D economic returns, to calculate the onward value of the use of data resources in the wider economy. The results of our analysis suggest that the access value of EMBL-EBI's data services to our users is well in excess of £272 million ($335 million) per year, that efficiency impacts for our users are in excess of £1 billion ($1.2 billion) per year, and that our annual return on investment is at least £920 million ($1.13 billion) per year.

EMBL-EBI monitors the use of its resources to evaluate their reliability, impact, and utility for the life sciences community. In defining measures of quality it is important to recognize the context in which the service is being provided and to base categorization on a range of criteria. For example, a resource that serves a small community may not have as many page views as a large resource, yet reach 90% of the community it supports, and may be critically important for that community.

---

[1] http://www.nature.com/articles/sdata201618

[2] Patterns of database citation in articles and patents indicate long-term scientific and industry value of biological data resources, https://f1000research.com/articles/5-160/v1.

[3] The Value and Impact of the European Bioinformatics Institute, http://bit.do/EMBL-EBI_Impact.

Monitoring almost always includes regular review by an advisory committee as well a qualitative assessment of community demand for a resource; data provenance; creation and use of community standards; and sharing and collaboration with other resources. Quantitative metrics such as (but not limited to): reliability (percentage of uptime), access speeds, usage statistics (IPs, page views, downloads), citations in the literature, data submission rates, international collaborations, programmatic access, and curational effort are also crucial, and are particularly useful for long-term monitoring of resource usage.

Given these considerations we do not explicitly prioritize some resources over others: prioritization occurs at the stage of developing a new resource. If we identify an unmet need for a data resource we will attempt to fund a pilot resource, then evaluate continuation of the resource based upon feedback from our users.

However, it is clear that some resources are used by wider communities and/or are critical not only for researchers but also for other resources. EMBL-EBI has developed a set of key indicators that recognize the heterogeneous nature of biological data, and the diversity of the supporting data resources, use cases, and communities served. Our approach to evaluating data resources has been developed in coordination with ELIXIR, which uses similar criteria to identify ELIXIR Core Data Resources, as described in a recent publication[4] and in a separate response to this RFI from the ELIXIR Hub.

The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

Data should be made available for re-use for as long as they may have value for users and researchers. In principle EMBL-EBI data resources are managed and supported in perpetuity. At the institutional level this means that we are committed to supporting resources that are still in use and valuable to the scientific community: we have to date never retired a resource solely for budgetary reasons. There have been cases when technological advances have rendered a data resource obsolete, in which case it has been retired or the data are incorporated into a larger resource.

Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

Public data resources benefit from economies of scale. Public data repositories are most efficient when resources are centralized at only one or a few institutions. Highly distributed data resources suffer from inefficiencies in findability and accessibility, are subject to multiple budgetary risks, and require much higher administrative and technical overhead support than centralized institutions.

Some data resources, such as UniProt and the European Nucleotide Archive/GenBank/DDBJ, are jointly hosted with other institutions, including NCBI, and

---

[4] Durinx C, McEntyre J, Appel R *et al.* Identifying ELIXIR Core Data Resources [version 1; referees: 1 approved]. F1000Research 2016, **5**(ELIXIR):2422 (doi: 10.12688/f1000research.9656.1)

this shared hosting reduces running and development costs. With exponential increases in data volumes submitted to public archives these efficiencies are critically important for sustained growth of data archives.

Public access data repositories are an infrastructure, like the electricity grid or public roads, and also like those infrastructures their benefits accrue to the entire society. Infrastructures have high startup costs followed by constant running costs, but yield continuous and long-term benefits to society as a whole. And just as the road system is primarily funded by public monies, funding for biological data resources is also most efficient if supported by centralized funders, usually through the medium of funding bodies, as in the UK and Europe, or national-level research bodies, such as NIH in the U.S.

As open access and scientific knowledge is global, funders may be concerned about 'free riding' (use from outside the funding territory). Although research and knowledge are inherently global, an number of economic studies have established the share of returns to R&D accruing locally to where the knowledge was gained are in excess of 70%.[5]

Inclusion of Data and Software Citation in NIH Research Performance Progress Reports (RPPR) and Grant Applications

Data and software, alongside research articles, are key outputs of research. Often, data and software may be available prior to a paper being published, and frequently provide the foundations for future research. Therefore, it seems quite reasonable to include the opportunity to cite data and software in RPPRs and grant applications. This will be of particular importance in areas of research such as computational biology where research outputs may not always result in high-impact papers, but the contribution is of great value to a wider user community.

The ability to include data and software citations requires that certain technical requirements are met.

To cite and link to a dataset or piece of software could be as straightforward as supplying a brief description and a URL and date of retrieval or accession.

However, for a more robust approach that also ensures both human and machine readablity it is much better if the object has a unique persistent identifier, such as a DOI or an accession number, along with mechanisms to resolve those identifiers to the specific location (or descriptive landing page) where the data or software are archived. For data that use accession numbers rather than DOIs, global uniqueness for resolution purposes can be achieved by using curies: a combination of database name as a prefix, followed by the accession number, which allows services such as identifiers.org to resolve that combination to the correct location (https://www.w3.org/TR/curie/). Furthermore, it is useful for

---

[5] Hall, BH, Mairesse, J and Mohnen, P (2009) Measuring the returns to R&D, NBER Working Paper 15622, NBER, Cambridge MA.
[5]Verspagen, B (2004), The impacts of Academic Knowledge on Macroeconomic Productivity Growth: An Exploratory Study, Eindhoven Centre for Innovation Studies, Eindhoven

human readers to have key metadata supplied for further context where the data or software are cited, in much the same way as an article is cited in a reference list.

There has been progress in implementing mechanisms and practices that support data citation (according to the key references cited in this RFI), although moving pilots increasingly into full production mode will require further developments and engagement with key stakeholders such as data resources, publishers, reference tool developers and funders. Support for software citation is more nascent still (see: https://www.force11.org/software-citation-principles), and also needs a program of community engagement to structure development around implementation.

Both data and software citation have challenges with respect to citation of particular versions, granularity of citation and the requirement to cite collections of files/datasets associated with a study. Approaches such as the BioStudies database at the EBI, which acts as a container of links and files for all data about a study will help simplify data citation (and thereby encourage it) by allowing a single high-level link.

While these requirements will take time and resources to resolve, social adoption of data and software citation will undoubtedly be the greater challenge. A key driver here is promoting the understanding among researchers that credit for data and software will be given. Mechanistically, linking data and software outputs to an ORCID can support this, and projects such as THOR (https://project-thor.eu) are beginning to enable this. However, perhaps more important here is the need for parallel policy-level actions from funders such as NIH to publicly recognize the value of research products other than papers; allowing and encouraging data and software citation in RPPRs and grant applications would be a key indicator of this and an opportunity for NIH to lead in this area.

Submission Date

01/19/2017

Submitter Name

Valerie Jackson, MD

Name of Organization

Radiological Society of North America (RSNA)

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

medical imaging

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

While many different types of data are required for various areas of clinical research, medical imaging provides especially high value for research on a wide range of health issues. Moreover, the expense and effort involved in acquiring, preparing, aggregating and sharing imaging data sets argue for special efforts to ensure their fullest possible use. We therefore recommend that NIH make it a high priority to encourage and support efforts to share imaging research data sets.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

With respect to imaging data, RSNA recommends that NIH endeavor to make publicly funded research data available for secondary use for an extended period of time after it is first published, and that NIH develop policies regarding both unrestricted and controlled access. Because of the size and complexity of imaging data sets and the relatively high expense of maintaining them, it is especially important that data sharing plans address the sustainability and business models for imaging data repositories. Where NIH designates existing, independently operated repositories, researchers depositing data need to be assured of their continued availability. Additionally, it is important that NIH direct researchers, wherever possible, to repositories that store data in standards-based, non-proprietary data formats.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

Imaging data present special challenges because of their size, variety, complexity of their composition (incorporating multiple image types, metadata and other associated textual information) and special liability to exposure of protected health information. Imaging data curation, storage and access require special infrastructure, processes and procedures. NIH should establish guidelines for appropriate methods for image sharing and require detailed description of the methods to be used in research data sharing plans. The particular tools required for image data discovery, annotation, curation and sharing may differ significantly depending on the composition and intended use of the particular data set. A distributed network of federated data warehouses may thus be more effective for image sharing data than a single consolidated repository. NIH should support the development and use of standards, procedure guidelines and open source tools that enable data sharing through federated data repositories. RSNA joins AMIA in advocating dedicated funding from research sponsors for data curation and donation efforts so there are sufficient incentives to share, collaborate, and advance data sharing capabilities. We recommend NIH earmark a percentage of grant funds for such activities as a way to overcome cost barriers. In combination with scoring data sharing plans (DSPs), explicitly setting aside funds to carry-out the DSP will improve data stewardship and sharing. Further, ensuring adherence to FAIR principles – Findability, Accessibility, Interoperability and Reusability – will help demonstrate value to overcome cost concerns.

4. Any other relevant issues respondents recognize as important for NIH to consider

Because medical imaging is at the center of many research trials and is an essential data type for biomedical scientific research, a high priority must be placed on encouraging sharing and reuse of medical imaging data sets. Advances like the Precision Medicine Initiative will rely on quantitative imaging to provide evidence-based measures for the detection, diagnosis and treatment of disease. The development, validation and dissemination of imaging biomarkers will require that extensive, high-quality imaging data sets are made accessible to researchers. RSNA has worked with funding from the National Institute of Biomedical Imaging and Bioengineering (NIBIB) to establish the Quantitative Imaging Biomarkers Alliance® (QIBA®) and the Quantitative Imaging Data Warehouse as well as the RSNA Image Share Network. By providing direction, incentives and support to encourage researchers to utilize such resources for sharing medical imaging data, we believe that it will accelerate progress in critical fields of research.

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

RSNA concurs with AMIA in affirming the value of such requirements provided that research sponsors award dedicated funding for data curation and donation efforts. Additionally, we recommend that reporting requirements be shared across venues (i.e. RPPR, publication in journals, etc.) with common guidance and metadata wherever possible. We further note that multiple approaches to point towards data, such as through data repository URLs, software source code hosting services, and DOIs, should be supported.

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

RSNA joins with AMIA in strongly supporting development of policies to cite data and software developed by grantees. Institutional incentives are needed to encourage development of reusable data and software. Development of persistent unique identifiers, such as DOIs, for data / software citation would be an important contribution to this effort. We note, however, that currently such DOIs for re-use are not well established. We encourage NIH to fund specific projects to improve the use of DOIs for data and software, and encourage NIH to explore how open source code and software containers, which represent snapshots of entire operating system configurations of computers used to develop software, can be leveraged to improve research rigor. Inclusion of data sharing activities in scientists' career assessments is a potentially powerful means of incentivizing data sharing. Recent "Alt-metrics" efforts have begun to build the framework for tracking data reuse and citation as meaningful measurements of researcher contributions. As a means to help develop these policies and to further encourage data sharing, RSNA joins AMIA in recommending that NIH host a roundtable of academic medical leaders, and produce a handbook for integrating this type of "credit" into promotion and tenure decisions.

b. Inclusion of a link to the data/software resource with the citation in the report

c. Identification of the authors of the Data/Software products

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

As discussed above, RSNA concurs with AMIA's recommendation that making data sharing plans scoreable aspects of pertinent grant applications would enable reviewers to assess the mechanisms through which data / software will be shared, and encourage more systematic, robust sharing strategies. One example for how NIH could operationalize the scoring of data sharing plans would be to score both according to priority data types and research targets and according to data quality and usability metrics similar to the 5-star deployment scheme for Open Data of the W3C.

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them

in RPPRs and Competitive Grant Renewals applications

Further, development of policies and templates for Data Management Plans (DMPs) would strengthen data and software sharing, especially towards the goals of traceability and reproducibility of research. We refer you to AMIA's response to the RFI for citation of research in progress that underscores the inconsistencies prevalent in current DMPs and emphasize the need for comprehensive DMPs that address upstream activities that impact data quality, provide traceability or support reproducibility. RSNA also recommends that NIH provide guidelines for Institutional Review Board (IRB) restrictions as well as academic promotions and tenure considerations that promote data sharing and utilization of publicly available data when applicable. These should include making the citation of shared data sets a requirement in any publication or grant activity. Offices of promotions and tenure should also include the development and sharing of data sets as a contribution to the scientific process to be considered in productivity and scientific impact metrics. Moreover, when publicly available data is reused for secondary research a mechanism should be developed to "credit" the contributor(s) of the original data. When research data sets are shared for reuse, it is not possible to anticipate all potential secondary uses of data in initial IRB applications. While approved IRB studies can be modified, re-consenting of patients for data reuse is often not possible. We recommend that NIH provide guidelines to IRB and HIPAA Privacy and Security offices that encourage policies friendly to data sharing by relieving some restrictions currently placed on researchers.

4. Any other relevant issues respondents recognize as important for NIH to consider

Because medical imaging is at the center of many research trials and is an essential data type for biomedical scientific research, a high priority must be placed on encouraging sharing and reuse of medical imaging data sets. Advances like the Precision Medicine Initiative will rely on quantitative imaging to provide evidence-based measures for the detection, diagnosis and treatment of disease. The development, validation and dissemination of imaging biomarkers will require that extensive, high-quality imaging data sets are made accessible to researchers. RSNA has worked with funding from the National Institute of Biomedical Imaging and Bioengineering (NIBIB) to establish the Quantitative Imaging Biomarkers Alliance® (QIBA®) and the Quantitative Imaging Data Warehouse as well as the RSNA Image Share Network. By providing direction, incentives and support to encourage researchers to utilize such resources for sharing medical imaging data, we believe that it will accelerate progress in critical fields of research.

Additional Comments

RSNA Comment Letter re NIH RFI re Research Data Sharing FINAL 2017-01-19.pdf (105 KB)

RADIOLOGICAL SOCIETY OF NORTH AMERICA
820 JORIE BLVD, OAK BROOK, IL 60523
TEL 1-630-571-2670 FAX 1-630-571-7837
RSNA.ORG

RSNA Radiological Society of North America

January 19, 2017

Carrie D. Wolinetz, Ph.D.
Associate Director for Science Policy
Office of Science Policy
National Institutes of Health

Submitted electronically at: http://osp.od.nih.gov/content/nih-request-information-strategies-nih-data-management-sharing-and-citation

Re: Request for Information: Strategies for NIH Data Management, Sharing and Citation

Dear Dr. Wolinetz:

The Radiological Society of North America (RSNA®) is pleased to offer responses to National Institutes of Health's (NIH) request for information (RFI) on Strategies for NIH Data Management, Sharing and Citation. The RSNA is an international society of radiologists, medical physicists and other medical professionals with more than 54,000 members from 136 countries across the globe. The RSNA promotes excellence in patient care and health care delivery through education, research and technologic innovation.

RSNA strongly supports NIH efforts to develop standard policies and introduce incentives for data management, sharing and citation. Expanded access to quality research data will improve the rigor and transparency of scientific research and spur the pace of scientific discovery.

In response to the RFI, we wish to affirm the vital role that medical imaging plays in a wide range of medical research and urge NIH to consider sharing of imaging data for research among its highest priorities. Aided by funding from the National Institute of Biomedical Imaging and Bioengineering (NIBIB), RSNA established the Quantitative Imaging Biomarkers Alliance® (QIBA®) and the Quantitative Imaging Data Warehouse. This effort provides access to a growing and diverse body of high-quality image data sets in order to facilitate development, validation and implementation of quantitative imaging biomarkers. In another NIBIB-funded project, RSNA established the Image Share Network, which enables exchange of medical images and reports for clinical and research purposes. By providing incentives for researchers to share imaging data through such facilities, NIH could accelerate the growth of data sharing resources and multiply their value for science and clinical care.

As to specific policy mechanisms to create incentives for data sharing, we would echo the response to this RFI provided by the American Medical Informatics Association (AMIA) with whom we have consulted in preparing our own response. In particular, we endorse AMIA's recommendations that NIH should:

1) Make data sharing plans a "scoreable" element of grant applications subject to the existing policy,
2) Earmark support for data sharing as of part of applicable grants' direct costs and
3) Develop mechanisms that enable institutional rewards for scholars who create and/or contribute to public datasets and software that other researchers find useful.

We refer you to the AMIA response for detailed rationale and further elaboration of these recommendations. We provide our specific responses to questions in the RFI below.

Thank you for the opportunity to share these views. We hope that this discussion and our collective efforts will bring a higher quality of care and better health outcomes to our patients.

Sincerely,

Valerie Jackson, MD
Chair, RSNA Board of Directors

Curtis P. Langlotz, MD, PhD
Liaison for Information Technology and Annual Meeting, RSNA Board of Directors

RADIOLOGICAL SOCIETY OF NORTH AMERICA
820 JORIE BLVD, OAK BROOK, IL 60523
TEL 1-630-571-2670 FAX 1-630-571-7837
RSNA.ORG

RSNA® Radiological Society of North America

Detailed Recommendations and Comments to NIH Questions

SECTION 1. Data Sharing Strategy Development

High-priority types of data to be shared

While many different types of data are required for various areas of clinical research, medical imaging provides especially high value for research on a wide range of health issues. Moreover, the expense and effort involved in acquiring, preparing, aggregating and sharing imaging data sets argue for special efforts to ensure their fullest possible use. We therefore recommend that NIH make it a high priority to encourage and support efforts to share imaging research data sets.

The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

With respect to imaging data, RSNA recommends that NIH endeavor to make publicly funded research data available for secondary use for an extended period of time after it is first published, and that NIH develop policies regarding both unrestricted and controlled access. Because of the size and complexity of imaging data sets and the relatively high expense of maintaining them, it is especially important that data sharing plans address the sustainability and business models for imaging data repositories. Where NIH designates existing, independently operated repositories, researchers depositing data need to be assured of their continued availability. Additionally, it is important that NIH direct researchers, wherever possible, to repositories that store data in standards-based, non-proprietary data formats.

Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

Imaging data present special challenges because of their size, variety, complexity of their composition (incorporating multiple image types, metadata and other associated textual information) and special liability to exposure of protected health information. Imaging data curation, storage and access require special infrastructure, processes and procedures. NIH should establish guidelines for appropriate methods for image sharing and require detailed description of the methods to be used in research data sharing plans.

The particular tools required for image data discovery, annotation, curation and sharing may differ significantly depending on the composition and intended use of the particular data set. A distributed network of federated data warehouses may thus be more effective for image sharing data than a single consolidated repository. NIH should support the development and use of

standards, procedure guidelines and open source tools that enable data sharing through federated data repositories.

RSNA joins AMIA in advocating dedicated funding from research sponsors for data curation and donation efforts so there are sufficient incentives to share, collaborate, and advance data sharing capabilities.[1] We recommend NIH earmark a percentage of grant funds for such activities as a way to overcome cost barriers. In combination with scoring data sharing plans (DSPs), explicitly setting aside funds to carry-out the DSP will improve data stewardship and sharing. Further, ensuring adherence to FAIR principles – Findability, Accessibility, Interoperability and Reusability – will help demonstrate value to overcome cost concerns.

Any other topics respondents recognize as important for NIH to consider

RSNA concurs with AMIA's response that:

1) NIH should ensure that data sharing policies are clearly articulated to both researchers and patients; that there are mechanisms for consent management; provisions of notification when data is used, and ways to share / return results in appropriate circumstances.

2) NIH may wish to articulate expectations around pre-publication data management, including annotation, metadata, and provenance. For example, NIH should consider what documentation should be shared along with data that is necessary to support its reuse, such as processes for transformation, imputation, coding, mapping standardization, data cleaning, and data quality assessments.

3) NIH should develop guidelines and best practices around data discoverability, including the use of model annotations, metadata schemas focused on a given domain (e.g. imaging) and minimal metadata expectations.

Section II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports (RPPR) and Grant Applications

The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

RSNA concurs with AMIA in affirming the value of such requirements provided that research sponsors award dedicated funding for data curation and donation efforts. Additionally, we recommend that reporting requirements be shared across venues (i.e. RPPR, publication in journals, etc.) with common guidance and metadata wherever possible. We further note that multiple approaches to point towards data, such as through data repository URLs, software source code hosting services, and DOIs, should be supported.

[1] Borne, P., Lorsch, J., Green, E., "Perspective: Sustaining the big-data ecosystem," *Nature*. November 2015. 527, S16–S17

Important features of technical guidance for data and software citation in reports to NIH, which may include:

- *Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI) (https://www.iso.org/obp/ui/#iso:std:iso:26324:ed-1:v1:en)*
- *Inclusion of a link to the data/software resource with the citation in the report*
- *Identification of the authors of the data/software products*
- *Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately*
- *Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed;*

RSNA joins with AMIA in strongly supporting development of policies to cite data and software developed by grantees. Institutional incentives are needed to encourage development of reusable data and software.

Development of persistent unique identifiers, such as DOIs, for data / software citation would be an important contribution to this effort. We note, however, that currently such DOIs for re-use are not well established. We encourage NIH to fund specific projects to improve the use of DOIs for data and software, and encourage NIH to explore how open source code and software containers, which represent snapshots of entire operating system configurations of computers used to develop software, can be leveraged to improve research rigor.

Inclusion of data sharing activities in scientists' career assessments is a potentially powerful means of incentivizing data sharing. Recent "Alt-metrics" efforts have begun to build the framework for tracking data reuse and citation as meaningful measurements of researcher contributions. As a means to help develop these policies and to further encourage data sharing, RSNA joins AMIA in recommending that NIH host a roundtable of academic medical leaders, and produce a handbook for integrating this type of "credit" into promotion and tenure decisions.

Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications;

As discussed above, RSNA concurs with AMIA's recommendation that making data sharing plans scoreable aspects of pertinent grant applications would enable reviewers to assess the mechanisms through which data / software will be shared, and encourage more systematic, robust sharing strategies. One example for how NIH could operationalize the scoring of data sharing plans would be to score both according to priority data types and research targets and according to data quality and usability metrics similar to the 5-star deployment scheme for Open Data[2] of the W3C.

[2] http://5stardata.info/en/

Further, development of policies and templates for Data Management Plans (DMPs) would strengthen data and software sharing, especially towards the goals of traceability and reproducibility of research. We refer you to AMIA's response to the RFI for citation of research in progress that underscores the inconsistencies prevalent in current DMPs and emphasize the need for comprehensive DMPs that address upstream activities that impact data quality, provide traceability or support reproducibility.

RSNA also recommends that NIH provide guidelines for Institutional Review Board (IRB) restrictions as well as academic promotions and tenure considerations that promote data sharing and utilization of publicly available data when applicable. These should include making the citation of shared data sets a requirement in any publication or grant activity. Offices of promotions and tenure should also include the development and sharing of data sets as a contribution to the scientific process to be considered in productivity and scientific impact metrics. Moreover, when publicly available data is reused for secondary research a mechanism should be developed to "credit" the contributor(s) of the original data.

When research data sets are shared for reuse, it is not possible to anticipate all potential secondary uses of data in initial IRB applications. While approved IRB studies can be modified, re-consenting of patients for data reuse is often not possible. We recommend that NIH provide guidelines to IRB and HIPAA Privacy and Security offices that encourage policies friendly to data sharing by relieving some restrictions currently placed on researchers.

Any other topics respondents recognize as important for NIH to consider.

Because medical imaging is at the center of many research trials and is an essential data type for biomedical scientific research, a high priority must be placed on encouraging sharing and reuse of medical imaging data sets. Advances like the Precision Medicine Initiative will rely on quantitative imaging to provide evidence-based measures for the detection, diagnosis and treatment of disease. The development, validation and dissemination of imaging biomarkers will require that extensive, high-quality imaging data sets are made accessible to researchers. RSNA has worked with funding from the National Institute of Biomedical Imaging and Bioengineering (NIBIB) to establish the Quantitative Imaging Biomarkers Alliance® (QIBA®) and the Quantitative Imaging Data Warehouse as well as the RSNA Image Share Network. By providing direction, incentives and support to encourage researchers to utilize such resources for sharing medical imaging data, we believe that it will accelerate progress in critical fields of research.

Submission Date

01/19/2017

Submitter Name

Sarah J. Wright

Name of Organization

Cornell University Research Data Management Service Group

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

Life sciences

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

We provide this guidance to Cornell researchers: Some NSF directorates offer specific guidelines, while most funders are still offering very little specific guidance. NSF's Engineering (ENG) directorate, for example, includes “analyzed” data in that directorate’s policy, meaning those data that are published in articles, dissertations, or supplementary materials. Note that figures within a publication aren’t sufficient – tables of the numbers used to create figures should be made available. In the guidance we’ve seen so far, sharing raw data is not typically required. Where no specific guidance is available, we recommend researchers keep in mind two things when deciding which data to share: What data are necessary to reproduce or validate your results? Note that this may include code. What data have the potential for reuse by others?

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

We provide this guidance to Cornell researchers: Some NSF directorates offer more specific guidelines (ENG specifies data should be kept at least for three years). If you are depositing your data in a data center or archive with a long-term commitment to providing access to the data, then you should simply state this in your plan. If you plan to host the data yourself or pay a service provider to host it for you, then you should specify a time period that is reasonable and that your budget can sustain, and explain that in your data management plan.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

We provide this guidance to Cornell researchers: "NSF does allow for costs associated with data management (typically line G2, with an explanation in the budget justification). If you are depositing your data in a data center or archive, then your data will probably be available for the long term. Most data centers or repositories either accept data free of charge (if it is within their collection scope) or charge a one-time fee at the time of deposit, making budgeting fairly straightforward. Currently, Cornell doesn’t offer any services which allow up-front payment for longer term storage, although the RDMSG is aware of the need for such a service and is considering different options." The question of how to pay for ongoing storage after the end of a grant is one of the biggest barriers, and up-front payment for long-term storage seems to be the best way to attack this problem. Support for developing such a model might also be supported by funding agencies.

4. Any other relevant issues respondents recognize as important for NIH to consider

metadata - even well cited data does not help if it's not properly described for re-use. Support for metadata development and standards that are accessible to researchers, not just metadata experts, need to be more ubiquitous.

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

Increased reporting of data and software allows data producers to demonstrate the impact of their work and helps to establish research data as an important contribution to the scholarly record. For users, citation makes it easier to find datasets, encouraging the reuse of data for new research questions. We also provide this advice on data citation to Cornell researchers: Citing data is very similar to citing publications; there are many "correct" formats to use, but we suggest including the following important information: creator(s) or contributor(s) date of publication title of dataset publisher identifier (e.g. Handle, ARK, DOI) or URL of source version, when appropriate date accessed, when appropriate The order of the information is not as important as having sufficient information to find the data set(s) used. Consider the style guidelines of the research domain or lab group, data source, or preferred publisher. Data publishers may provide a suggested citation that can include additional information such as resource type, retrieval date, and funder or sponsor information. Some repositories will also request citation of related publication(s) along with the data. Follow the most appropriate format while meeting the requirements of the data suppliers.

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

When evaluating repositories, we encourage researchers to seek a repository that can issue a persistent identifier because DOIs (or other identifiers) help to assure reliable, predictable, and unambiguous access to research data.

b. Inclusion of a link to the data/software resource with the citation in the report

c. Identification of the authors of the Data/Software products

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

4. Any other relevant issues respondents recognize as important for NIH to consider

metadata - even well cited data does not help if it's not properly described for re-use. Support for metadata development and standards that are accessible to researchers, not just metadata experts, need to be more ubiquitous.

Additional Comments

Submission Date

01/19/2017

Submitter Name

Matthew Spitzer

Name of Organization

Center for Open Science

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

The Center for Open Science (https://cos.io) supports research across all domains.

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

We recognize the unique challenges for some disciplines in sharing vast amounts of raw, unprocessed data. That being said, every step in which a dataset is "cleaned" represents a set of decisions that are likely to affect the ultimate conclusions for any analysis and also represents a potential loss of information for later scholars. To satisfy these conflicting priorities, we recommend the following standards and principles: At a minimum and as a default, the final, cleaned dataset used in analysis should be openly shared without licensing restrictions on reuse. Exceptions to the above rule should be publicly justified, based solely on ethical or moral constraints, and should follow a reasonable effort to make the data available in an ethical manner. All scripts or instructions provided to data analysts used to clean the raw data should be openly shared.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

The value of a dataset to the research community, like any given publication, is difficult to judge in advance. Given the facts that perpetual storage of data that may have questionable value is not in the public interest and the difficulty in assessing value, we recommend increased investment in data curation experts and long term data storage solutions for content that appears in the scholarly record.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

Preparing the data for public availability does incur some additional cost to the researcher. However, that cost cannot be borne by every requester for the dataset, but rather should be borne by the original publisher of the dataset, who is responsible for ensuring that funding for the research includes resources to prepare a dataset for public sharing.

4. Any other relevant issues respondents recognize as important for NIH to consider

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

Ideally the increase of data and software sharing will not happen via a reporting mechanism, but through a transparent and open workflow that allows funders as stakeholders to evaluate these outputs when needed, or as additional milestones. Funders should consider requiring the sharing of data and software to better enable the extensibility of the research.

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

Use of persistent IDs for citation to any digital object should be gold standard and expected norm of any research output. Knowing that that ideal will take time to achieve and that formation of new habits is slow, we recommend that deviations from the norm be permitted and noted through disclosure statements. These statements should indicate why a particular citation is not provided, without expectation for any particular justification. See, for example, the disclosure statements provided by Nature Publishing Group and others provided by the Center for Open Science.

b. Inclusion of a link to the data/software resource with the citation in the report

Many journals include guidelines for linking to supplemental materials contained within appropriate persistent repositories and this should be required for RPPRs and any eventual publications that result.

c. Identification of the authors of the Data/Software products

Citation and credit are the currency of rewards throughout academia. Data and software products are seen as an increasingly important scholarly output. To the greatest extent possible, and consistent with the spirit of any licensing included in the data or software, the identity of the authors should be made as evident as is the case with other scholarly outputs (i.e. traditional published articles).

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

It is the responsibility of those who generate, post, store, and maintain large datasets to provide sufficient clarity to any end-user to be able to cite the author to the degree that the author desires. When large datasets are created and those who are storing the data are not able to provide credit to the individual data providers, that limitation should be made explicit to the data providers. This is a common occurrence in fields such as ornithology, where “citizen science” generated data are ubiquitous and where such notifications to data generators (i.e. the citizen scientist volunteers) are commonly included. It is then the responsibility of the data user and those reviewing that researcher’s work (either through grant applications or articles) to cite the data to the greatest degree of granularity that they can reasonably discern.

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

The digital repository should be selected from an approved list of verified independent services that meet minimum archival standards, such as http://www.re3data.org/

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

Confirming that Data Management Plans (DMP) are reviewed and meet the data sharing suggestions and requirements could be used to trigger additional awards or supplemental funding and/or acknowledgement.

4. Any other relevant issues respondents recognize as important for NIH to consider

Additional Comments

Submission Date

01/19/2017

Submitter Name

Mary M. Langman

Name of

Organization

Medical Library Association and Association of Academic Health Sciences Libraries

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

health and biosciences

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

The Medical Library Association and Association of Academic Health Sciences Libraries find that raw research data, including human trial data, images, and video, along with their metadata in non-proprietary format - data can be validated/confirmed/reviewed, allowing for reproducibility and transparency of research products

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

● Length of time--MLA and AAHSL recommend a 12-month embargo which enhances exclusivity and matches publication policy, and that; data should be available for 10 years in order to balance the usefulness of data in the field with the logistics of storing data for long periods of time). The associations recommend revisiting these parameters in 10 years. ● Maintaining and Sustaining--MLA and AAHSL recommend utilizing publically available databases including: institutional repositories, discipline-specific repositories (re3data, etc.), third party repositories (Zenodo, etc.) ● The associations recommend asking NLM and NIH to convene a panel of experts to address this issue

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

● Storing/sharing 1. Some repositories charge for storage (i.e. Dryad) 2. Faculty culture against sharing - need for education and reevaluation of how data is weighted when considering tenure and promotion, as well as other academic incentives ● Stewardship 1. Specialized skills required - might require hiring new staff 2. Requires large time commitment to properly describe datasets for sharing 3. De-identifying data while retaining its usefulness 4. Requires knowledge of systems that ensure long-term validity of data – i.e. not corrupted, deleted, changed or destroyed – tracking / audit and validation system ● Mechanisms to overcome barriers 1. Develop standards or consistent ways/places to store and describe data 2. Develop models for systems, and sharing of open source solutions

4. Any other relevant issues respondents recognize as important for NIH to consider

● Assessing impact of single study that generates multiple data sets might be difficult, a clear way to do this is desired ● Emphasize the importance of data catalogs - how are things attributed, where are they located, and how are they accessible ● Many publishers are already doing this, the goals need to be aligned ● Consider what other agencies are doing

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

● Gives another research product to evaluate impact ● Model after publication model - require unique identifier that can then be shared outside of the RPPR (via tools such as NIH Reporter) ● Link ORCID to datasets via unique identifier ● Extract data and place into a centralized system so it can be found and used

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

Crucial if sharing of and access to research outputs is to occur and should be easy to generate given any structured repository as most established repositories already mint DOIs

b. Inclusion of a link to the data/software resource with the citation in the report

N.A.

c. Identification of the authors of the Data/Software products

● Link ORCID to datasets via unique identifier ● Link to eRA Commons accounts (similar to PMCID)

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

● All data sets should stand alone when combined for studies ● Researchers want to track back to the original sources of data and not just the compiled data set

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

● Including a direct link to where the data lives via URL, DOI, etc. - minimum should include metadata that details where data sets are located ● For software - standardized citation rules

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

Including data identifiers inside NIH Reporter or Pubmed (like PMCID)

4. Any other relevant issues respondents recognize as important for NIH to consider

● Assessing impact of single study that generates multiple data sets might be difficult, a clear way to do this is desired ● Emphasize the importance of data catalogs - how are things attributed, where are they located, and how are they accessible ● Many publishers are already doing this, the goals need to be aligned ● Consider what other agencies are doing

Additional Comments

Submission Date

01/19/2017

Submitter Name

Neil Chue Hong

Name of Organization

Software Sustainability Institute

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

We collaborate with researchers across all domains of research.

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

In our opinion, all data that can be used to advance scientific research should be shared. The NIH invests large amounts of funding in scientific research for the common good. Consequently, the produced knowledge should be treated as a public good that will boost innovation and lead progress. Thus, the results of cooperative efforts must be made available to all scientists. In the digital age, this approach is translated as demand for Open Access and Open Science based on Open Standards and Free Software, and articulated in the FAIR principles for data (http://www.datafairport.org/). Given the explosive growth of data generation, the challenge is where effort should be invested. The Institute believes that priority should be given to data which is a) expensive to collect; b) expensive to curate; and c) difficult to recreate or collect. There is immense social value in sharing such data, because it enables discoveries to be made which can significantly impact the entire planet, such as in the development of anti-malarial drugs (DOI: 10.1038/nrd.2016) There is also an economic benefit to the sharing of data. Reports by the European Commission and McKinsey (summarised in https://medium.com/@ODIHQ/the-economic-impact-of-open-data-what-do-we-already-know-1a119c1958a0#.2bxt1ysuf) and a study of the US Landsat dataset, comprising satellite imagery of the Earth's surface, showed the huge annual economic benefit of it being made openly available was $2.19bn in 2011 alone (http://landsat.gsfc.nasa.gov/?p=10949).

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

The Blue Ribbon Task Force on Sustainable Digital Preservation and Access note that bodies like the NIH "should impose preservation mandates, when appropriate. When mandates are imposed, funders should also specify selection criteria, funds to be used, and responsible organizations to provide archiving." Additionally, NIH "should work with grantees and domain researchers to determine archiving needs and consider how economies of scale might be developed to take advantage of specialized use scenarios." Likewise, "Funding agencies should explicitly recognize "data under stewardship" as a core indicator of scientific effort and include this information in standard reporting mechanisms." This provides an incentive to data producers and curators, and importantly the evidence to justify funding which is not based on soft money and project funding. Ideally, data would be available for the foreseeable future. The 4C (Collaboration to Clarify the Costs of Curation) project (http://4cproject.eu/) has developed the Curation Cost Exchange (http://www.curationexchange.org/) to enable understanding and comparison of digital curation costs to support smarter investments as well as collating current understanding of the cost models for digital preservation (http://www.curationexchange.org/read-more/21-other-literature-and-projects). To efficiently manage costs, the NIH needs to provide strong guidance to those it supports with respect to the infrastructure they should use to make their data available. Whilst curation must come from the community, sharing data through many individual sites administered by staff on project funding is fragile and can lead to the loss of resources due to loss of staff or funding.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

There is usually a tradeoff between maximizing access and maximizing funding. A key question to be resolved is how to respect the balance between those who are willing to pay to use the data, those who use the data without paying, and those who cannot pay. If a data resource has too many free-riders, their usage will consume too large a proportion of resources. However, if the barriers to use the data are too onerous, the resource will be underutilised. Models such as direct infrastructure funding and institutional subscriptions may be able to preserve low barriers / costs to access with sufficient funding for curation and community building.

4. Any other relevant issues respondents recognize as important for NIH to consider

The Software Sustainability Institute has played a key part in understanding the role of software citation for research. We have published guidance on citing software (https://www.software.ac.uk/how-cite-and-describe-software) used by UK funders and which have fed into the Software Citation Guidelines. Co-I Prof. Carole Goble has keynoted at workshops organised by the NSF on Data and Software Citation (http://www.software4data.com/) and participated in NIH BD2K workshops. Director Neil Chue Hong was a participant and contributor to the FORCE11 Software Citation Working Group and Future of Software Metadata workshop, and developer of Software Management Plan guidance for UK funders (https://www.software.ac.uk/resources/guides/software-management-plans). Deputy Director Simon Hettrick presented at the 2015 Cyberpractioner Workshop in Washington DC and the Moore Sloan Data Science Summit. We observe that software infrastructure, including that which supports sharing of data, requires a different set of skills, funding and career paths from basic scientific research to maintain. This is also noted in the RCUK Review of e-Science: "Software development is not basic research, but instead requires a critical mass of full-time engineering professionals, paid market wages and supported along a genuine career path." In the UK, this has been acknowledged through the formation of the Research Software Engineers association (http://www.rse.ac.uk/) and award of RSE Fellowships. In the USA, the nascent Advanced Cyberinfrastructure Research & Education Facilitators network (http://www.aciref.org/) is connecting professionals in this area.

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

This reporting provides two useful abilities: - Ability to assess reuse of data and software products - Ability to investigate the "half-life" of products (i.e. whether products become less or more useful through time) Significantly, it sheds light on the true costs of data and software production and use, and suggests more appropriate levels of funding for data sharing and software maintenance. A report by the Software Sustainability Institute (http://www.researchresearch.com/news/article/?articleId=1347583) suggested that at least a third of the total funding on research by the UK research councils was dependent on software. However this vastly underestimates the true cost of data and software, and in particular the wastage when data and software are unnecessarily replicated because they are not shared in a way that makes them reusable.

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

We recommend the adoption of the Joint Declaration of Data Citation Principles (https://www.force11.org/group/joint-declaration-data-citation-principles-final) and implementation as guidance which includes the recommendation of use of ORCIDs for author identification and DOI's for data and software. Existing work has been done to understand how to achieve human and machine accessibility of cited data in scholarly publications (https://doi.org/10.7717/peerj-cs.1). We recommend that the links are in the form of DOIs, as our (to be published) studies of software outputs from UK research projects suggest that data made available through website URLs is often moved or removed after a few years.

b. Inclusion of a link to the data/software resource with the citation in the report

The inclusion of a link to the data or software resource as a citation is vitally important for science for three reasons: 1) It improves reproducibility and transparency, by asking authors of reports to consider where the resources they are using

are 2) It highlights the critical roles that data and software play in modern research 3) It provides a mechanism for giving credit to those that create and maintain data and software, improving the sustainability and reuse of these resources However, the main impact of the inclusion of a link to the resource is that for humans reading the report, it enables them to quickly access the resource, without having to parse the citation - especially important if the report is presented as a webpage or PDF.

c. Identification of the authors of the Data/Software products

Whilst initiatives like Project CRediT (http://docs.casrai.org/CRediT) have created better taxonomies for disambiguating contributor roles and ideas like Transitive Credit (http://openresearchsoftware.metajnl.com/articles/10.5334/jors.be/) provide potential mechanisms for crediting indirect and varying contributions, it is still the case for software that identification and correct attribution of authors is difficult. Almost all revision control systems will record who has contributed what to any particular version of the software, but identified by accounts which are local to the system being used. Many developers of software used in research do not have ORCIDs so cannot be uniquely identified using that system. The challenge for software is that the value of contributions cannot easily be measured automatically, as quantity of contributions does not correlate with their perceived value, and the value may change over time. Therefore, at present, our recommendation is to use a similar approach to papers and define authorship of software products based on the agreed norms of the relevant community, taking into account the principles that an author should have a) made substantial contributions to the conception, design or implementation of the software AND b) been involved in revising or reviewing it critically AND c) agreed to be accountable to questions related to the accuracy of the software. Federal funding agencies including the NIH should support efforts to convene key players to identify and harmonise standards on roles, attribution, value, and credit. This will provide a way forward for identification of authors.

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

We believe that a spectrum of granularity of citations, both in data and software, will be required depending on different use cases. Citation of an aggregation or collection of data (or of a library or application) is useful when one is invoking the reference as a general pointer to background information. Citation of a particular data set or software release is useful when using the reference to provide information on methodology. Finally, although not yet common, citation at granualities lower than this (microattribution) where the reference is to a specific part of the data or software (e.g. a cell in a spreadsheet, record in a database or function in a piece of code) will become useful to identify author contributions in large, long-lived resources (see: http://onlinelibrary.wiley.com/doi/10.1002/humu.22144/abstract).

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

The Software Citation Principles (https://peerj.com/articles/cs-86/) developed by the FORCE11 software citation working group provide a set of use cases justifying the necessity for unambiguously identifying and citing the digital repository.

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

A prize for "Open Science" should be instigated to recognize researchers who share their research results, including data and software with others in a way that is exemplary. This way, as well as honouring scientists for their scientific work, the NIH will raise awareness concerning the benefits of Open Access and Open Science among universities, schools, private entities and other stakeholders, driving the reuse of research outputs by the wider community. Software Management Plans should be mandated for NIH funding, which function in the same way as Data Management Plans, asking proposers to consider who is responsible for the release and sharing of their software. The Software Sustainability Institute has been working with UK research funders to develop guidance (https://www.software.ac.uk/resources/guides/software-management-plans).

4. Any other relevant issues respondents recognize as important for NIH to consider

The Software Sustainability Institute has played a key part in understanding the role of software citation for research. We have published guidance on citing software (https://www.software.ac.uk/how-cite-and-describe-software) used by UK funders and which have fed into the Software Citation Guidelines. Co-I Prof. Carole Goble has keynoted at workshops organised by the NSF on Data and Software Citation (http://www.software4data.com/) and participated in NIH BD2K workshops. Director Neil Chue Hong was a participant and contributor to the FORCE11 Software Citation Working Group and Future of Software Metadata workshop, and developer of Software Management Plan guidance for UK funders (https://www.software.ac.uk/resources/guides/software-management-plans). Deputy Director Simon Hettrick presented at the 2015 Cyberpractioner Workshop in Washington DC and the Moore Sloan Data Science Summit. We observe that software infrastructure, including that which supports sharing of data, requires a different set of skills, funding and career paths from basic scientific research to maintain. This is also noted in the RCUK Review of e-Science: "Software development is not basic research, but instead requires a critical mass of full-time engineering professionals, paid market wages and supported along a genuine career path." In the UK, this has been acknowledged through the formation of the Research Software Engineers association (http://www.rse.ac.uk/) and award of RSE Fellowships. In the USA, the nascent Advanced Cyberinfrastructure Research & Education Facilitators network (http://www.aciref.org/) is connecting professionals in this area.

Additional Comments

Submission Date

01/19/2017

Submitter Name

Ary L. Goldberger

Name of Organization

Beth Israel Deaconess Medical Center/Harvard Medical School

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

cardiovascular and cardiopulmonary science

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

Among the most useful data from our perspective our multi-channel recordings from deidentified subjects where the the data are annotated and some relevant outcome or effect measures are provided.Examples include data from PhysioNet and the National Sleep Resource Resource.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

The longer the time these data are provided the better for research, validation, teaching and other activities.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

Commitment of funds are required for preparation and curation of databases,

4. Any other relevant issues respondents recognize as important for NIH to consider

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

High priority should be given to those who share data especially where altruism leads to new findings.

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

b. Inclusion of a link to the data/software resource with the citation in the report

A definite plus.

c. Identification of the authors of the Data/Software products

Yes,major credit should be given to those who provide data as an essential part of scientific process.

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

4. Any other relevant issues respondents recognize as important for NIH to consider

Additional Comments

Submission Date

01/19/2017

Submitter Name

Andrew SMITH

Name of Organization

ELIXIR

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

Human data, rare disease, plant sciences, marine metagenomics

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

It is clear that there is great value in data sharing, for individual researchers, science in general, innovation and society at large. Studies into the patterns of database citation within research articles and patents indicate major scientific and industrial impact over the long-term (Bousfield et al. 2016). Every effort should be made to ensure that data from research projects are made available as a matter of course, and in a manner adhering to the FAIR Guiding Principles (Wilkinson et al. 2016). Rather than specify in this response which types of data should be highest priority, ELIXIR suggests that where domain-specific repositories exist and are acknowledged by the community as the natural archive for a particular dataset, their use should be encouraged. Within ELIXIR, the process to define indicators for the establishment of 'Core Data Resources' and the current call to select the first set (ELIXIR 2017) acknowledges that some resources are of fundamental importance and play a critical role in supporting whole disciplines.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

ELIXIR considers that data should be made publically available for as long as possible: value constantly emerges from existing data as technology moves forward and new scientific disciplines emerge. The example of protein structures stored within the Protein Data Bank (PDB) demonstrates that existing data generate significant value for academics and industry, decades after their generation. The vast volumes of new data being generated - taken alongside institutional policies to maintain the data publically available for as long as possible - clearly poses budgetary and technical challenges in storing data in perpetuity. On the technical side, compression solutions for alignment files such as CRAM format (Cochrane et al. 2013) and tiered storage for those data used less often can be considered as possible ways of ensuring that data are stored in a cost-effective way. An acknowledgement that databases operate within a life-cycle helps institutions take decisions on whether to retire particular resources. EMBL-EBI, for example, considers resources operate within specific life-cycles. The data within resources in the 'stationary' stage are stored in perpetuity. However, resources can be retired if technological advancement renders it obsolete. On the rare occasions this happens, it follows active consultation with the user community and often the data within that resource are subsumed into another, usually, larger resource. In terms of resource implications, there should be an acknowledgement that the current funding streams for research, based on short term grant cycles, are not necessarily fit for purpose for sustaining databases and components of the data infrastructure.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

For data generators, an important impact of increased reporting and storage of data is the changes to routine and workload that open data mandates have on individual researchers and institutions. ELIXIR suggests governing bodies may reduce the burden for individual researchers by: 1) promoting curated, redundant and maintained collaborative IT and software solutions and supporting services offering economies of scale. 2) incentivize data and software sharing by recognizing the need for funding support of data storage in context with grant applications 3) recognizing the emerging

new role of data stewards (e.g. biobanks and data warehouses) as key building blocks of long-term archiving resources and vital components of infrastructure. 4) creating convincing showcases to highlight the benefit of good data management and good data management planning practice to the project itself, rather than only to re-users of the data. 5) promoting data discovery services, allowing researchers to optimise resources and access (meta)data and services and tools to support data sharing by projects As data and software are increasingly considered first-class products of scientific research, so too should the standards and associated tools and services used to manage, aggregate and describe them. The ELIXIR Interoperability Platform (EIP) is coordinating the development of sustainable capacity around three main paradigms: 1) The FAIR principles (Wilkinson et al. 2016) ; 2) The interoperability provisioning profile. i.e. supporting metadata annotation throughout the lifecycle of data management; and 3) The interoperability value proposition. i.e. shifting the balance of effort / return in data management practices.

4. Any other relevant issues respondents recognize as important for NIH to consider

Most databases - despite storing data generated from research projects, or providing value-added data used by researchers - are funded through a combination of multiple research grants, which have a typically short duration, low success rates and by their nature focus primarily on the development of new research activities rather than service provision. ELIXIR recommends the development of more fit for purpose funding schemes developed with the intention of supporting the operations of services rather than purely new research activities. ELIXIR welcomes the discussions taking place internationally around funding models and their appropriateness and the role the NIH is playing in engaging in these. Global discussions around sustainability hosted through the Human Frontiers Science Programme and the trials taking place within the Big Data to Knowledge Initiative (BD2K) to test the concept of 'cloud coins' for computing are important first steps to understand and develop solutions to the challenge of the long-term resource implication that comes with data infrastructures.

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

It is considered by many, including the Research Data Alliance group on Data Management Plans ("Active Data Management Plans IG" 2014), to be more valuable to update the data management plan continuously over the course of a project rather than developing only one plan at the start of the project. By updating the data management plan it continues to represent the actual processes and decisions that have been taken during the operation of the project. Indeed the final version of the data management plan should be considered an output of the project, and (potentially after removing/masking sensitive information) published. Publication of earlier versions is less important from a data sharing standpoint, but could be valuable to assist resource planning for data services as they can be aware of what kind of projects and associated resource demands may be upcoming. Annual reports providing public summary information about the status of the data management plans and shared research outcomes could incentivize researchers to comply with the requirement to make data available in research projects.

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

ELIXIR is currently working in collaboration with Force11 and other NIH partners to set up a list of recommendations open to more than one identifier scheme and supporting identifiers maintained by reference biomedical repositories. We recommend following three citation guidelines proposed under the umbrella of Force11: - Uniform Resolution of Compact Identifiers for Biomedical Data (Wimalaratne et al. 2017) - A Data Citation Roadmap for Scholarly Data Repositories (Fenner et al. 2016) Software citation principles (Smith et al. 2016)

b. Inclusion of a link to the data/software resource with the citation in the report

c. Identification of the authors of the Data/Software products

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and

when might each distinct data set underlying a study be cited and reported separately

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

Software is a critical part of modern research and yet there is little support across the scholarly ecosystem for its acknowledgement (Smith, Katz, and Niemeyer 2016). Like data and publications, software is a research product that should be shared and cited the same way. To address main challenges in software citation and encourage broad adoption of a consistent policy we recommend: - The adoption of the "Software citation principles" (Smith, Katz, and Niemeyer 2016) - The adoption of minimum information software metadata guidelines like the Bioschemas Tools specification ("BioSchemas - Tools Group" 2016)

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

Citation is an important metric to measure recognition and make quantitative assessments of scientific programs. However it should not be the only metric to assess impact. In software, like with data, the impact can be measured by other indicators like views, downloads and accessibility (Durinx et al. 2016). Thus we recommend funders: To consider other indicators to evaluate the impact of software like the indicators suggested by (Durinx et al. 2016). In addition to citations and impact, the quality and sustainability of software is also a critical factor, as this has repercussions for the reproducibility and quality of the research. Software development best practices contribute to increases quality and sustainability of research software (Leprevost et al. 2014). Thus ELIXIR recommends: compliance with principles like the Open Source Software Recommendations, which promote the discovery and reuse of software as well as the adoption of software development best practices.

4. Any other relevant issues respondents recognize as important for NIH to consider

Most databases - despite storing data generated from research projects, or providing value-added data used by researchers - are funded through a combination of multiple research grants, which have a typically short duration, low success rates and by their nature focus primarily on the development of new research activities rather than service provision. ELIXIR recommends the development of more fit for purpose funding schemes developed with the intention of supporting the operations of services rather than purely new research activities. ELIXIR welcomes the discussions taking place internationally around funding models and their appropriateness and the role the NIH is playing in engaging in these. Global discussions around sustainability hosted through the Human Frontiers Science Programme and the trials taking place within the Big Data to Knowledge Initiative (BD2K) to test the concept of 'cloud coins' for computing are important first steps to understand and develop solutions to the challenge of the long-term resource implication that comes with data infrastructures.

Additional Comments

FULLVERSIONCopyofELIXIRresponsetoNIHRequestforInformation.pdf (223 KB)

ELIXIR response to NIH Request for Information (RFI): Strategies for NIH Data Management, Sharing, and Citation (Notice Number: NOT-OD-17-015)

Intro to ELIXIR submission

This submission represents the response of ELIXIR Europe, the pan-European research infrastructure for biological data. ELIXIR is the initiative to coordinate, sustain and integrate Europe's life science bioinformatics resources, providing a platform for scientific discovery in the life sciences.

ELIXIR is a distributed infrastructure with a central Hub – with the primary function of coordination - located on the Wellcome Genome Campus, Hinxton, and national Nodes – with the primary function of service delivery - in each participating Member State across Europe. The following countries and EMBL are Members of ELIXIR: Belgium, Czech Republic, Denmark, Estonia, France, Finland, Germany, Hungary, Ireland, Israel, Italy, Luxembourg, the Netherlands, Norway, Portugal, Slovenia, Sweden, Switzerland, Spain and the UK. Greece is an Observer.

The submission has been developed in consultation with ELIXIR Nodes and with in-depth input from experts within ELIXIR's Interoperability Platform (EIP) and Data Management Plans Working Group. This response addresses both Section 1 and Section 2.

Section I: Data Sharing Strategy Development

The highest-priority types of data to be shared and value in sharing such data

ELIXIR welcomes the concerted efforts being undertaken by NIH to develop strategic approaches to the topics of data management and sharing and concurs fully with the NIH assessment that 'data sharing is essential for expedited translation of research results into knowledge, products, and procedures to improve human health'.

It is clear that there is great value in data sharing, for individual researchers, science in general, innovation and society at large. Studies into the patterns of database citation within research articles and patents indicate major scientific and industrial impact over the long-term (Bousfield et al. 2016). The Human Protein Atlas, for example, receives on average five citations daily[1] and on average there is one submission to the European Nucleotide Archive every six minutes.

[1] Source: Human Protein Atlas web logs

Every effort should be made to ensure that data from research projects are made available as a matter of course, and in a manner adhering to the FAIR Guiding Principles (Wilkinson et al. 2016). Rather than specify in this response which types of data should be highest priority, ELIXIR suggests that where domain-specific repositories exist and are acknowledged by the community as the natural archive for a particular dataset, their use should be encouraged. Within ELIXIR, the process to define indicators for the establishment of 'Core Data Resources' and the current call to select the first set (ELIXIR 2017) acknowledges that some resources are of fundamental importance and play a critical role in supporting whole disciplines.

Relating to privacy, an impact of increased long-term storage and reuse of data in the biomedicine and healthcare fields is the conflict between the interest for archiving and accessibility and considerations on data protection and personal privacy. Many upcoming and ongoing studies intend to cross reference future health records with biobank data and with lifestyle information through monitoring or questionnaires. This requires an infrastructure setup which allows secure (re)-identification of participants. We recommend that the perspective on patients and participants in medical research is acknowledged as important participants, rather than passive donors and that the landscape of future research and personalised medicine will require patient collaboration and participation. In line with this we recommend increased focus on merging biobank dual responsibilities towards researchers and patients, making it possible for patients to contribute securely to research while at the same time receiving some of the benefits of the outcomes, gain insight into the use of their samples and regain some control of research activities. (Cassa et al. 2012) (Wolf et al. 2012; Jamuar et al. 2016; McGuire et al. 2014; Darnell et al. 2016; Capocasa et al. 2016; Prince et al. 2015)

The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

ELIXIR considers that data should be made publically available for as long as possible: value constantly emerges from existing data as technology moves forward and new scientific disciplines emerge. The example of protein structures stored within the Protein Data Bank (PDB) demonstrates that existing data generate significant value for academics and industry, decades after their generation. The current mining of digitalised electronic health records also demonstrates how new value can be derived from old data.

The vast volumes of new data being generated - taken alongside institutional policies to maintain the data publically available for as long as possible - clearly poses budgetary and technical challenges in storing data in perpetuity. On the technical side, compression solutions for alignment files such as CRAM format (Cochrane et al. 2013) and tiered storage for those data used less often can be considered as possible ways of ensuring that data are stored in a cost-effective way. As file formats and data storage are constantly updated and improved, long-term storage and discovery of data are very real challenges. In addition, new long read sequencing technologies, for example, will add further storage burden in the near future.

An acknowledgement that databases operate within a life-cycle helps institutions take decisions on whether to retire particular resources. EMBL-EBI, for example, considers the following life-cycle stage of a resource: young, established, stationary, retiring and extinct. The data within resources

in the 'stationary' stage are stored in perpetuity. However, resources can be retired if technological advancement renders it obsolete. On the rare occasions this happens, it follows active consultation with the user community and often the data within that resource are subsumed into another, usually, larger resource. See response to this RFI by EMBL-EBI for further information on this aspect.

In terms of resource implications, there should be an acknowledgement that the current funding streams for research, based on short term grant cycles, are not necessarily fit for purpose for sustaining databases and components of the data infrastructure. Here, longer term investments are required and dedicated schemes focussed on infrastructure rather than research could help to ensure a more effective, coordinated landscape that is more effective in making data available.

Most databases - despite storing data generated from research projects, or providing value-added data used by researchers - are funded through a combination of multiple research grants, which have a typically short duration, low success rates and by their nature focus primarily on the development of new research activities rather than service provision. ELIXIR recommends the development of more fit for purpose funding schemes developed with the intention of supporting the operations of services rather than purely new research activities.

ELIXIR welcomes the discussions taking place internationally around funding models and their appropriateness and the role the NIH is playing in engaging in these. Global discussions around sustainability hosted through the Human Frontiers Science Programme and the trials taking place within the Big Data to Knowledge Initiative (BD2K) to text the concept of 'cloud coins' for computing are important first steps to understand and develop solutions to the challenge of the long-term resource implication that comes with data infrastructures.

Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

Many barriers to data stewardship and sharing exist, both for the data generators and the infrastructure operators. For data generators, an important impact of increased reporting and storage of data is the changes to routine and workload that open data mandates have an individual researchers and institutions, creating a barrier between intention and implementation. Another barrier is that individual researchers may fail to see that dealing with a large volume of complex data operated on by a consortium of scientists in multiple locations requires more rigor in procedures than many researchers may be used to.

Rather than focusing on a rigorous bureaucratic standardization framework relating to data storage and citation, ELIXIR suggests governing bodies may reduce the burden for individual researchers by:

1) seeking to promote curated, redundant and maintained collaborative IT and software solutions and supporting services offering economies of scale.
2) considering ways to incentivize data and software sharing by recognizing the need for funding support of data storage in context with grant applications
3) recognizing the emerging new role of data stewards (e.g. biobanks, data warehouses and interoperability services) as key building blocks of long-term archiving resources and vital components of infrastructure.
4) creating convincing showcases to highlight the benefit of good data management and good data management planning practice to the project itself, rather than only to re-users of the data.

As data and software are increasingly considered to be first-class products of scientific research, so too should the standards and associated tools and services used to manage, aggregate and describe them.

The ELIXIR Interoperability Platform (EIP) is coordinating the development of sustainable capacity around three main paradigms: 1) The FAIR principles (Wilkinson et al. 2016) ; 2) The interoperability provisioning profile. i.e. supporting metadata annotation throughout the lifecycle of data management; and 3) The interoperability value proposition. i.e. shifting the balance of effort / return in data management practices.

ELIXIR considers a key driver of knowledge and innovation to be through the provisioning of FAIR data through specific products and services. Finding, sharing and reuse of data between different systems are underpinned by standards. As outlined by Sansone and Rocca-Serra, there are several types of standards of particular importance to the clinical and life sciences community.

- Machine-processable descriptions - e.g., minimum reporting requirements, terminologies, ontologies, file formats or conceptual models for citation, credit and interoperability purpose.
- Identification for discovery, citation and credit;
- Accessibility of the information - e.g., access permission, data protection, patient consent, anonymization and encryption;
- Indicators or metrics to measure performance, use and quality;
- Versioning and documentation practices - e.g., for code, algorithms or tools;
- Tracking provenance of and relationships between digital concepts - e.g., interpretations and conclusions;
- Analysis - e.g., standardized descriptions of the workflow and related software used

From Sansone and Rocca-Serra (Susanna-Assunta and Philippe 2016)

In order to reduce the barriers to data stewardship and sharing further, ELIXIR is developing capacity in the areas of structured metadata, content standards, linked data and identifiers. Specific products, tools and services, include (but are not limited too):

- BioSharing, an ELIXIR Service providing a comprehensive curated resource that maps the landscape of over a thousand standards in the life sciences. The resource also interlinks them with databases and policies (from funders journals and other organisations).
- The Ontology Lookup Service (OLS), a repository for biomedical ontologies that aims to provide a single point of access to the latest third party ontology versions. Users can browse the ontologies through the website as well as programmatically via the OLS API.
- Identifiers.org, an identifier resolving service that enables the referencing of data for the life sciences community in both a location-independent and resource-dependent manner.
- Data FAIRport, a FAIR data interoperability platform that supports the full lifecycle of FAIR, allowing to create, publish, find and annotate data.
- Bioschemas, which coordinates the extension of schema.org for the life sciences domain, and provides a structured semantic markup for web pages' content (including life sciences data) used by the main search engines.
- NIH BD2K bioCADDIE in collaboration with Bioschemas has developed a metadata model called DATS (DatA Tag Suite), which underpins the data discovery index DataMed.

- Cooperation with NIH BD2K bioCADDIE/Force11 Data Citation Implementation Pilot which sets out early adopter recommendations for data repositories and a roadmap for publishers. ELIXIR is working with (Fenner et al. 2016) this group to harmonise our identifier resolution services.

Any other topics respondents recognize as important for NIH to consider

ELIXIR recommends incentivizing not only the creation of new shared data, but also the use of previously shared data (especially other people's data) in new projects. In terms of data release, exceptions to sharing data could be made when the authors can show that the societal value of the data would be higher if sharing is delayed. Delays (embargo), however, should always be time bound and release unconditional.

Incentivizing collaboration between authors of existing software that could include new functionality with those seeking to create new similar software will also produce effective results: the latter usually has to go to a lot of the same development stages and efficiencies can be gained through collaboration.

The ScienceEurope 2015 report on multidisciplinary research sets out recommendations for crediting, rewarding and funding along the whole pipeline of data and software creation, management and use

Section II: Inclusion of Data and Software Citation in NIH Research Performance Progress Reports (RPPR) and Grant Applications

The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

It is considered by many, including the Research Data Alliance group on Data Management Plans (“Active Data Management Plans IG” 2014), to be more valuable to update the data management plan continuously over the course of a project rather than developing only one plan at the start of the project. By updating the data management plan it continues to represent the actual processes and decisions that have been taken during the operation of the project.

Indeed the final version of the data management plan should be considered an output of the project, and (potentially after removing/masking sensitive information) published. Publication of earlier versions is less important from a data sharing standpoint, but could be valuable to assist resource planning for data services as they can be aware of what kind of projects and associated resource demands may be upcoming.

Annual reports providing public summary information about the status of the data management plans and shared research outcomes could incentivize researchers to comply with the requirement to make data available in research projects.

Important features of technical guidance for data and software citation in reports to NIH:

ELIXIR agrees that citation using a Persistent Unique Identifier for both data and software, including appropriate versioning and metadata, is an important feature for citation in reports. However, we believe that any governing recommendations should avoid explicit examples, especially of commercial products and registered trademarks, as several other methods are commonly in use in the biomedical field. Also, for fields where there is a desire to use the identifier for both the digital object and the actual physical sample, as can be the case when describing specific genetic strains of plants or animal tissue.

ELIXIR is currently working in collaboration with Force11 and other NIH partners to up a set of recommendations not restricted to one identifier scheme to provide not just a digital identifier of an object but a persistent and unique link to the physical object. We further stress that the persistence of such a Unique Identifier does not lie within the choice of a particular identifier framework, but in the infrastructure set up by the owner of the data source, their commitment to provide long-term accessibility and the resources available to fulfill this commitment.

We recommend following three citation (Fenner et al. 2016):

- Uniform Resolution of Compact Identifiers for Biomedical Data (Wimalaratne et al. 2017)
- A Data Citation Roadmap for Scholarly Data Repositories (Fenner et al. 2016)
- Software citation principles (Smith et al. 2016)

The Force11 Data Citation Implementation Group supported by NIH BD2K bioCADDIE and ELIXIR are currently working on recommendations for implementing data citation based on the Joint Declaration of Data Citation Principles (JDDCP). This group and specifically the identifiers subgroup are working on supporting harmonization of identifier resolution services to help to identify and cite data involving established resolution services like Identifiers.org, EZID and n2t.

Citation of research software has not been as well attended as other research outcomes as data and publications. However, there is clearly a role for setting standards for citing software. Software is a critical part of modern research and yet there is little support across the scholarly ecosystem for its acknowledgement (Smith, Katz, and Niemeyer 2016). Like data and publications, software is a research product that should be shared and cited the same way. Software is not just a means to an end, but a collective intellectual product, a fundamental asset for building scientific knowledge (Hettrick et al. 2016). Software present challenges similar to data citation. For instance the reference to software should not just include a textual citation, but the identification and description of the specific software version used within the research process. To address main challenges in software citation and encourage broad adoption of a consistent policy we recommend the adoption of the "Software citation principles" (Smith, Katz, and Niemeyer 2016) . We also recommend the adoption of minimum information guidelines like the Bioschemas Tools specification ("BioSchemas - Tools Group" 2016) to complement the principles to consistently describe research software.

Any other relevant issues respondents recognize as important for NIH to consider

Citation is an important metric to measure recognition and make quantitative assessments of scientific programs. However it should not be the only metric to assess impact. In software, like with data, the impact can be measured by other indicators like views, downloads and accessibility

(Durinx et al. 2016). Thus we recommend funders to consider other indicators to evaluate the impact of software like the indicators suggested by (Durinx et al. 2016).

In addition to citations and impact, the quality and sustainability of software is also a critical factor, as this has repercussions for the reproducibility and quality of the research. Software development best practices contribute to increases quality and sustainability of research software (Leprevost et al. 2014). That said, ELIXIR recommends compliance with principles like the Open Source Software Principles ("FAIR Open Source Software Principles - Publication" 2016), which promote the discovery and reuse of software as well as the adoption of software development best practices. We believe by adopting these principles the funders will increase the chances to increase the quality and sustainability of software developed in research grants.

References

"Active Data Management Plans IG." 2014. *RDA*. May 14. https://www.rd-alliance.org/groups/active-data-management-plans.html.

"BioSchemas - Tools Group." 2016. Accessed December 5. http://bioschemas.org/groups/tools/tool.html.

Bousfield, David, Johanna McEntyre, Sameer Velankar, George Papadatos, Alex Bateman, Guy Cochrane, Jee-Hyub Kim, et al. 2016. "Patterns of Database Citation in Articles and Patents Indicate Long-Term Scientific and Industry Value of Biological Data Resources." *F1000Research* 5 (February). doi:10.12688/f1000research.7911.1.

Capocasa, Marco, Paolo Anagnostou, Flavio D'Abramo, Giulia Matteucci, Valentina Dominici, Giovanni Destro Bisol, and Fabrizio Rufo. 2016. "Samples and Data Accessibility in Research Biobanks: An Explorative Survey." *PeerJ* 4 (February): e1613.

Cassa, Christopher A., Sarah K. Savage, Patrick L. Taylor, Robert C. Green, Amy L. McGuire, and Kenneth D. Mandl. 2012. "Disclosing Pathogenic Genetic Variants to Research Participants: Quantifying an Emerging Ethical Responsibility." *Genome Research* 22 (3): 421–28.

Cochrane, Guy, Blaise Alako, Clara Amid, Lawrence Bower, Ana Cerdeño-Tárraga, Iain Cleland, Richard Gibson, et al. 2013. "Facing Growth in the European Nucleotide Archive." *Nucleic Acids Research* 41 (Database issue): D30–35.

Darnell, Andrew J., Howard Austin, David A. Bluemke, Richard O. Cannon 3rd, Kenneth Fischbeck, William Gahl, David Goldman, et al. 2016. "A Clinical Service to Support the Return of Secondary Genomic Findings in Human Research." *American Journal of Human Genetics* 98 (3): 435–41.

"Data Citation Implementation Pilot (DCIP)." 2015. *FORCE11*. September 28. https://www.force11.org/group/dcip.

Durinx, Christine, Jo McEntyre, Ron Appel, Rolf Apweiler, Mary Barlow, Niklas Blomberg, Chuck Cook, et al. 2016. "Identifying ELIXIR Core Data Resources." *F1000Research* 5 (September). doi:10.12688/f1000research.9656.1.

ELIXIR. 2017. "ELIXIR Opens Process for Nominations to Core Data Resources < News < ELIXIR." Accessed January 16. https://www.elixir-europe.org/news/elixir-opens-process-nominations-core-data-resources.

"FAIR Open Source Software Principles - Publication." 2016. *Google Docs*. Accessed December 5. https://docs.google.com/document/d/1r_J1D2Lum1up5XXejCCBVKoPkGuMeYI11qgFAojRBAE/edit?usp=sharing&usp=embed_facebook.

Fenner, Martin, Mercè Crosas, Jeffrey Grethe, David Kennedy, Henning Hermjakob, Philippe Rocca-Serra, Robin Berjon, Sebastian Karcher, Maryann Martone, and Timothy Clark. 2016. "A Data Citation Roadmap for Scholarly Data Repositories." doi:10.1101/097196.

Hettrick, Antonioletti, Carr, Chue Hong, Crouch, De Roure, Emsley, et al. 2016. "UK Research Software Survey 2014." Accessed November 27. doi:10.5281/zenodo.14809.

Jamuar, Saumya Shekhar, Jyn Ling Kuan, Maggie Brett, Zenia Tiang, Wilson Lek Wen Tan, Jiin Ying Lim, Wendy Kein Meng Liew, et al. 2016. "Incidentalome from Genomic Sequencing: A Barrier to Personalized Medicine?" *EBioMedicine* 5 (March): 211–16.

Leprevost, Felipe da Veiga, Felipe da Veiga Leprevost, Valmir C. Barbosa, Eduardo L. Francisco, Yasset Perez-Riverol, and Paulo C. Carvalho. 2014. "On Best Practices in the Development of Bioinformatics Software." *Frontiers in Genetics* 5. doi:10.3389/fgene.2014.00199.

McGuire, Amy L., Bartha Maria Knoppers, Ma'n H. Zawati, and Ellen Wright Clayton. 2014. "Can I Be Sued for That? Liability Risk and the Disclosure of Clinically Significant Genetic Research Findings." *Genome Research* 24 (5): 719–23.

Prince, Anya E. R., John M. Conley, Arlene M. Davis, Gabriel Lázaro-Muñoz, and R. Jean Cadigan. 2015. "Automatic Placement of Genomic Research Results in Medical Records: Do Researchers Have a Duty? Should Participants Have a Choice?" *The Journal of Law, Medicine & Ethics: A Journal of the American Society of Law, Medicine & Ethics* 43 (4): 827–42.

Smith, Arfon M., Daniel S. Katz, and Kyle E. Niemeyer. 2016. "Software Citation Principles." *PeerJ Computer Science* 2 (September). PeerJ Inc.: e86.

"Software Citation Working Group." 2015. *FORCE11*. February 15. https://www.force11.org/group/software-citation-working-group.

Susanna-Assunta, Sansone, and Rocca-Serra Philippe. 2016. "Review: Interoperability Standards," October. doi:10.6084/m9.figshare.4055496.v1.

Wilkinson, Mark D., Michel Dumontier, Ijsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3 (March). Nature Publishing Group: 160018.

Wolf, Susan M., Brittney N. Crock, Brian Van Ness, Frances Lawrenz, Jeffrey P. Kahn, Laura M. Beskow, Mildred K. Cho, et al. 2012. "Managing Incidental Findings and Research Results in Genomic Research Involving Biobanks and Archived Data Sets." *Genetics in Medicine: Official Journal of the American College of Medical Genetics* 14 (4): 361–84

Van den Eynden, Veerle; Knight, Gareth; Vlad, Anca; Radler, Barry; Tenopir, Carol; Leon, David; Manista, Frank; Whitworth, Jimmy; Corti, Louise (2016): Survey of Wellcome researchers and their attitudes to open research. https://dx.doi.org/10.6084/m9.figshare.4055448.v1

Submission Date

01/19/2017

Submitter Name

Di Cross and Nigel Robinson

Name of Organization

Clarivate Analytics, formerly the IP & Science business of Thomson Reuters

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

Science of Science, Program Evaluation

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

While there is no overall priority to particular domains in terms of sharing, life science data is by far the most numerous and is the domain where most data are shared. Astronomy and astrophysics probably produces the highest volume (bytecount) of data which brings challenges for preservation. In some cases, trusted repositories are given a stamp of approval (for example, that which is provided by the World Oceanographic Data Center). The number of seals of approval is increasing and the criteria used for assessment can vary significantly. The DCI approach, however, takes a different view asking: Are the datasets in the repositories being used? What is the workflow in gathering meta-data? The selection process, described here http://wokinfo.com//products_tools/multidisciplinary/dci/selection_essay takes into account several variables in assessing each data repository. We look to ensure that a data repository is well used, is actively receiving data deposits and is being cited (ie, data are being reused). Where data are sensitive (patient records, or the location of endangered species of animals), there may be reasons not to share the data, or the data may need to be anonymized or redacted to enable sharing.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

The standard practice is to stipulate a minimum of 10 years. In our experience not all data repositories commit to 10 years, or have funding in place for that period. Long term preservation has its challenges in changing file formats and in the management of an ever increasing body of work. All of these challenges are well-documented.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

Many individuals contribute to data sharing and reuse: investigators creating data during their research efforts; information science professionals supporting investigators; investigators reusing data; managers and program officers evaluating the impact of data; and information technology specialists creating data repositories. It is essential to consider the perspectives of each group separately in identifying barriers in the process as the incentives and needs for each group may differ, yet the contribution from each group is necessary to achieve the shared goal. For example, reducing barriers for a data curator may not yield increases in the number of datasets shared publicly, since investigators who are the decision-makers in determining whether a dataset is to be shared may have other barriers. The skills needed to enable data sharing are more closely allied to library and information science professionals. We observe that many data sharing initiatives do not include such individuals and are heavily technology-driven. We recommend more extensive subject-matter indexing within and across repositories to facilitate discovery and reuse. Where possible, subject indexing is included in Clarivate Analytics' Data Citation Index. However this is limited by the variable quality and completeness of repository depositions. Today, investigators typically learn about relevant datasets by word of mouth and then search in repositories using the dataset author's name or a specific DOI. As data and articles become more connected, investigators will also be able to identify a data object for re-use through an article citing

those data. DCI enables this by linking data to articles.

4. Any other relevant issues respondents recognize as important for NIH to consider

Many have discussed the precept that data should be designed with sharing in mind, from the start of study conceptualisation, design and data collection through to data sharing, maintenance and preservation. To that end, it would be helpful for investigators to use workflow tools which are embedded within the research data life cycle. Such tools would serve to prompt investigators to do the things necessary to make data as useful as possible and also make these activities easier. We find it instructive to think of GitHub as an example of such an embedded workflow tool. Although GitHub does not itself house data (relying instead on Zenodo), it can be used for code creation, testing, and version control throughout the lifecycle of software development – ie, it is embedded in the processes preceding the act of sharing of code. This interplay between the GitHub platform and the everyday tasks of coders could be envisioned for data producers, thereby promoting the sharing of the resulting dataset. Some domain-specific examples of this for data already exist (ie, GenBank). We recommend increasing awareness of these tools (perhaps introducing them in training programs designed for researchers) and decreasing the barriers to obtaining and using these tools (perhaps subsidizing the cost of purchasing licenses or otherwise providing access).

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

With federal mandates requiring acknowledgement of funding in publications supported by public funds, we have seen a shift in acknowledge of funds in publications supported by the federal government as well as other funders more generally. Today, there is a tendency for investigators to over-report the number of grants involved in supporting work described in a given publication or for the duration of an RPRR. This potential shift into over-reporting may occur with data as well. We have also seen in our evaluation work that data can be reported within the RPPRs are often missing or difficult to access within NIH data systems used for analysis and tracking of grants. Though such data are invaluable for more effective program management and evaluation, program managers must download the PDF copies of RPPRs to get certain information (for example, achieved goals) which cannot be accessed directly in IMPACII or QVR and, even when downloaded, require significant additional processing for analysis. The availability, format and accessibility of these data make it difficult to use for program management and analysis and frequent monitoring.

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

We note that there are costs for obtaining a persistent identifier, though they are minimized by registering through DataCite, with whom we work in partnership to effectively flow the meta-data of deposited datasets into Clarivate Analytics' Data Citation Index. This differs from the model for obtaining a PMID (free to the investigator/author, registration performed by a government agency) and investigators may not be aware that the process for registering a DOI (involving costs to the investigator/author/author institution, registration performed by a non-profit organization sometimes by way of a for-profit entity) is different. In fact, investigators may expect greater similarity between the two models (ie, for data registration to be free). We recommend working with DataCite or other organizations in this arena to leverage existing infrastructure and thereby control costs, standardize meta-data across repositories to facilitate discovery and re-use, and thereby facilitate data sharing and re-use for and by NIH investigators.

b. Inclusion of a link to the data/software resource with the citation in the report

We agree that links to data/software resources are vital information in encouraging reuse of data. Clarivate Analytics' Data Citation Index provides a recommended citation format for every data object.

c. Identification of the authors of the Data/Software products

Authors are a key component for evidence-based attribution and tracking of deposited data – this is particularly relevant from the perspective of an academic institution or a research funder like NIH. Although meta-data standards exist,

funding and author affiliation are optional and therefore often missing. To date we have over 6.5 million datasets in DCI. However, less than 30% have author address and less than 15% contain funding information. We recommend reinforcing the necessity of acknowledging funding and reporting of author addresses with repositories and with NIH grantees.

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

There are numerous data citation standards, many of which are domain-specific. Clarivate has adopted the DataCite recommendations for data citation which are also supported more widely in the community and provide a minimum citation metadata set which can be applied across all domains.

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

This should be part of publisher mandates and is growing in the publishing community, but again there are domain variations and domain-specific practices. The FORCE11 declaration describes some of these issues. The data citation should name the repository as the source and provide a URI to link to the data object in the repository.

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

Requiring the reporting of data products either in RPPR or otherwise does not itself incentivize the sharing of useful data. Indeed, this may become a barrier to sharing, as investigators are increasingly burdened to report on the outputs of their research. Other incentives are required to facilitate the process of sharing data and to recognize those who share data. To encourage data re-use, we recommend the deposition of data in a recognised high quality repository with enough metadata to allow citation and open access for sharing. In our experience with DCI, over 95% of repositories offer open access to the data, some require registration to confirm the data user, but very few charge a fee for data. In addition, NIH may consider creating opportunities to specifically fund – and thereby raise the prestige of – data creation, curation, and sharing. The recognition that funding would bring to contributors whose skill sets are needed to make widespread data sharing and re-use possible will go a long way towards incentivizing these desired behaviors. This may take the form of supplements to existing grants, or as new grant types to support production, curation and sharing of data and software tools. Combined with the use of unique identifiers, this would enable NIH to efficiently track derived value from funded data. By partnering with suitable organizations, NIH can monitor the creation and reuse of data they have funded and also enforce research data mandates for metadata deposition and citation.

4. Any other relevant issues respondents recognize as important for NIH to consider

Many have discussed the precept that data should be designed with sharing in mind, from the start of study conceptualisation, design and data collection through to data sharing, maintenance and preservation. To that end, it would be helpful for investigators to use workflow tools which are embedded within the research data life cycle. Such tools would serve to prompt investigators to do the things necessary to make data as useful as possible and also make these activities easier. We find it instructive to think of GitHub as an example of such an embedded workflow tool. Although GitHub does not itself house data (relying instead on Zenodo), it can be used for code creation, testing, and version control throughout the lifecycle of software development – ie, it is embedded in the processes preceding the act of sharing of code. This interplay between the GitHub platform and the everyday tasks of coders could be envisioned for data producers, thereby promoting the sharing of the resulting dataset. Some domain-specific examples of this for data already exist (ie, GenBank). We recommend increasing awareness of these tools (perhaps introducing them in training programs designed for researchers) and decreasing the barriers to obtaining and using these tools (perhaps subsidizing the cost of purchasing licenses or otherwise providing access).

Additional Comments

Data Citation Best practice whitepaper.pdf (2172 KB)

REUTERS/JUAN CARLOS ULATE

WHITE PAPER

RECOMMENDED PRACTICES TO PROMOTE SCHOLARLY DATA CITATION AND TRACKING

The role of the Data Citation Index

THOMSON REUTERS™

INTRODUCTION

The history of scholarly advancement is closely linked to data re-use. In the spheres of science, social science, and arts and literature, the work and ideas of early scientists and scholars have lead to new and important discoveries in the eras that followed. While in times past, the passing on of scholarly data might have consisted of an inherited laboratory notebook or astronomical observations, today the preservation and dissemination of data increasingly takes place in the digital realm. As the volume of available scholarly data continues to increase at an exponential rate, scholarly societies and academic, private, and government entities look for new ways to disseminate and interpret this vast reservoir of information[1]. Meanwhile, the variety of data used by different disciplines presents unique challenges as entities look to devise standards that accommodate the needs of particular stakeholder groups, as well as the data needs and conventions of scholars in specialized and established areas of study.

Concurrent with these developments has been an increased interest in methods to assess the full impact of scholarly research, including traditional published research products, such as journal publications, as well as nontraditional products such as datasets and software. In this context, the practice of data citation has gained widespread attention in the academic community as a solution to issues of discovery and attribution for nonwtraditional scholarly output.

—

“THE DATA CITATION INDEX … AIMS TO PROVIDE A CLEARER PICTURE OF THE FULL IMPACT OF RESEARCH OUTPUT, AS WELL AS TO ACT AS A SIGNIFICANT TOOL FOR DATA ATTRIBUTION AND DISCOVERY.”

WHY CITE DATA?

Formal data citation has many benefits for the scientific and scholarly community. While open data has been recommended as a means to better instigate scientific discoveries and ensure reproducible results, there have been few demonstrable rewards for data-gathering institutions and individuals to fund programs and facilities for long-term data preservation and access. In many cases scholars and organizations must put into practice requirements as described by governing bodies with an interest in open data. Formal citation of data allows for these research stakeholders to receive proper credit for their work. Researchers may also gain information regarding data reuse, including in cases where data is not necessarily deposited in conjunction with the publication of a journal article, such as with publicly funded research organizations. Also, new metrics on scholarly output may provide benefits for funding and tenure considerations. These desirable outcomes have lead groups such as FORCE11 to develop principles of data citation that advocate data objects as unique citable entities[2].

WHY CITE DATA?

- Enables research conclusions to be verified and validated
- Makes reproducibility of premises and results possible
- Exposes data findings and their value to a wider audience
- Ensures a mechanism for receiving credit for scholarly work and an opportunity for tracking/translating such attribution into rewards

A NEW DATA TOOL

The Data Citation Index℠ (http://wokinfo.com/products_tools/multidisciplinary/dci/) was launched in 2012 by Thomson Reuters as a part of the Web of Science™ suite of indexing resources. In this index, descriptive records are created for data objects and linked to literature articles in the Web of Science. As data citation practices increase, the resource aims to provide a clearer picture of the full impact of research output, as well as to act as a significant tool for data attribution and discovery.

The resource has been developed with attention to the data and metadata needs of various scholarly disciplines, as well as the requirements of publishers and funders in these areas. Thomson Reuters has also entered into partnerships that promote the shared mission of increasing the acceptance of research data as citable contributions to the scholarly record. Through collaborations with data providers and organizations, the Data Citation Index looks to support stakeholders at every step in the data lifecycle, including researchers, data repositories/publishers, and administrators. Such support is enabled by best practices discussed here as they relate to these entities in the context of creation, deposition, and curation of metadata necessary for tracking data citation in the included data sources.

DATA PROVIDER COLLABORATIONS

In order to better address the needs of the data community, Thomson Reuters has partnered with individual data repositories as well as large-scale providers of data and metadata. These partnerships have allowed the Data Citation Index to gain perspective on metadata practices common to various areas of study, in order to develop standard, scalable bibliographic records based on the variable needs of different scholarly communities.

Thomson Reuters is forging a number of partnerships with individual repositories and databases to provide metadata for the creation of bibliographic records for data in the Data Citation Index. Descriptive, structural, and administrative metadata[3] are obtained using a variety of harvesting protocols (including the OAI-PMH XML-based standard exchange protocol). Through close liaison/content negotiation with its various data partners, the Data Citation Index builds upon the various common metadata standards employed to provide a cross-disciplinary data resource. Thomson Reuters can provide support to create the necessary metadata via a simplified, discipline-agnostic XML schema.

Ways to improve citing data

DATACITATION EXAMPLES: RECOMMENDED	DATA CITATION EXAMPLES: NOT RECOMMENDED
Irino, T; Tada, R (2009): Chemical and mineral compositions of sediments from ODP Site 127-797. PANGAEA. http://dx.doi.org/10.1594/PANGAEA.726855	Irino & Tada (2009). Chemical and mineral compositions of sediments from ODP Site 127-797. Published by PANGAEA [www.pangaea.de]
Elliott, Joshua (2013): Simulated county- and state-level maize yields, 1979-2012. Version 1. Figshare. http://dx.doi.org/10.6084/m9.figshare.501263	Elliott's Maize Yield Data (2013). Data accessed from Figshare [June 15, 2015]
Uniprot Consortium (2014): P0DKE6.Uniprot Knowledgebase. http://www.uniprot.org/uniprot/P0DKE6	Uniprot Database. http://www.uniprot.org

When selecting a data repository for inclusion in a data management plan, authors may look to the journals in their discipline; in some cases, journals provide recommendations or require that data be deposited in a specifi repository or in one of a list of recognized data centers.

Thomson Reuters has sought to expand the coverage of the Data Citation Index by partnering with two major data services: DataCite and the Australian National Data Service (ANDS). These organizations aim to encourage the dissemination, discoverability, and citability of research datasets from a diverse array of data repositories and data publishers through the provision of aggregated search and discovery facilities.

DataCite (https://www.datacite.org) is a leading global nonprofit organization dedicated to enabling people to find, share, use, and cite data. Formed in 2009, they aim to provide reliable, easy-to-use persistent data identification services to their partners underpinned by the DOI system (Digital Object Identifier – http://www.doi.org). DataCite engages stakeholders, including researchers, scholars, data centers, libraries, publishers, and funders through advocacy, guidance, and services. Thomson Reuters own recommendation on how to cite a data resource is aligned with DataCite recommendations.

The Australian National Data Service (ANDS) was established in 2008 and aims to enable Australia's research data to become a national strategic resource supporting better, more efficient research and improved policy input. ANDS is partnering with research institutions throughout Australia to help them manage their data more effectively. Their aggregated search and discovery facilities allow researchers on a global level to find and reuse Australian data as part of their ongoing work. Similar to DataCite, ANDS encourages the use of DOIs as persistent identifiers for data objects.

Thomson Reuters has reached an agreement with both DataCite and ANDS to include metadata for their partner repositories in the Data Citation Index[4,5], with editorial staff appraising DataCite and ANDS metadata to verify that they are within scope for inclusion in the resource. Through partnerships of this kind, metadata requirements from a diverse array of sources can be normalized across various platforms and data types to encourage the employment of a standard lexicon to address researcher, data provider, and funder requirements.

Upon data submission, it is important for authors to consider future citation of the data, as well as creating enriched metadata that increase the discoverability of the data set.

CREATION AND DEPOSITION OF DATA AND METADATA

Publicly available research data has been associated with an increased rate of citation for data authors[6]. Scholarly publishers, funding bodies, and government entities increasingly require research authors to make their data publicly available, particularly through repository deposition. Organizations such as the National Science Foundation (NSF) require a discussion of future data management upon submission of grant proposals for project funding[7]. Broad policy requirements put forward by these groups may lack specificity; however the author may often refer to discipline- and data-type specific guidelines for sharing of research results.

WHY DEPOSIT DATA?

When selecting a data repository for inclusion in a data management plan, authors may look to the journals in their discipline; in some cases, journals provide recommendations or require that data be deposited in a specific repository or in one of a list of recognized data centers.

Repositories suggested by journals or publishers are often specialized for the subject area of the publication, as well as for the data and metadata requirements of scholars and scientists in that discipline. It is not uncommon for publishers and funders to recommend or require that the data repositories used be public and/or make the data freely available. Exceptions are made where privacy concerns exist, such as in cases of health data with information about human subjects, data referencing sensitive sites of scientific interest (habitats for endangered species, etc), and commercially sensitive data. Recommended data centers are usually well established, either in their own discipline or as a multidisciplinary data resource.

If the author is affiliated with an academic institution, that institution may have a data repository of its own, often through the institutional library; researchers may wish to consult with data librarians at their institution to discuss the size and format of data sets which are accommodated. Where little guidance for data

Created data sets should cite previous data and traditional literature items (journal articles, books, etc.) where appropriate, and these publications should include data citations in reference sections or other specialist data sections of the paper...

deposition exists, or where no discipline-specific repository exists or meets criteria, authors are encouraged to explore the deposition of their research results in multidisciplinary repositories that meet their criteria for curation of data and metadata.

Thomson Reuters offers a searchable list of data repositories and sources from across the world and across disciplines which have so far been selected for coverage by the Data Citation Index (http://wokinfo.com/products_tools/multidisciplinary/dci/repositories/search/).

Upon data submission, it is important for authors to consider future citation of the data, as well as creating enriched metadata which increase the discoverability of the data set. Metadata elements such as keywords and discipline-specific indexing terms provide avenues for other researchers to discover and re-use the data. Author affiliations and grant and funding agency details provide important information regarding funded research output. This information is highly desirable; however, certain metadata elements are an absolute requirement for the accurate, formal data citation advocated by the Data Citation Index.

ELEMENTS OF A DATA CITATION

Required elements follow the basic, discipline-agnostic data citation guidelines put forward by DataCite (https://www.datacite.org/services/cite-your-data.html). Metadata elements needed for data citation include:

Author/Creator	Individuals or organizations that created or contributed to the data set; this metadata element is vital to guarantee attribution and credit for data contributor, and to provide metrics for their nontraditional scholarly output
Year	The year of "publication" of the data; when it is made publicly available, such as through deposition in a repository
Title	The title of the data object, which may differ from the title of the parent research paper/project
Publisher	The data repository that houses the data and/or the governing organization responsible for publishing, (i.e., making available) the data
Version	Dynamic data sets or those where new editions may be issued (such as with error corrections or new values) must employ proper version control to guarantee accuracy and uniqueness in data citation
Permanent Identifier	A unique and persistent identifier should be assigned; for example, a Digital Object Identifier (DOI); in Data Citation Index citations, this bibliographic element may take the form of a unique URL, databank accession number, or other permanent identifier such as Handle (hdl) (http://www.handle.net/)

A number of data community organizations, including the Research Data Alliance (RDA), DataCite, and Thomson Reuters, encourage authors to practice formal data citation in their work. In the absence of universal standards and guidance, each record in the Data Citation Index includes a recommended formal data citation to use for that data object. Created data sets should cite previous data and traditional literature items (journal articles, books, etc.) where appropriate, and these publications should include data citations in reference sections or other specialist data sections of the paper, as designated by the literature publisher.

RESEARCHER/DATA AUTHOR BEST PRACTICES

- Regard data equally with other citable research output such as journal publications
- Deposit data in an established data repository committed to long-term preservation and use of permanent IDs for data
- Contribute mandatory metadata with deposited data: authors, year, title, publisher, version
- Contribute further metadata to advance discovery: abstract, author affiliations, funding information, keywords, and data-specific information such as data type and methodology
- Practice detailed, formal data citation in data and publications; cite dataset permanent IDs

The submission process should encompass metadata creation to include careful consideration of proper data attribution, where contributing authors and organizations are given credit through author and repository/publisher attribution, or through citation of contributing data and their individual or institutional authors. Elevation of data to an equal footing with citable publications encourages increased citation of researcher output. Providing data to an established data repository may demonstrate agreement with funder and publisher requirements for data access and preservation[8]. Visible output of previously funded research projects provides evidence to funders of positive outcome of funding, helpful in future grant applications, while evidence of data re-use through citation tracking will validate return on funding investment. Discoverability of data through best practices in data deposition and metadata provision increases the likelihood of re-use of the data, ability to reproduce the research, and hence onward data citation and credit for data set authors[9,10], which is gaining increased use in tenure and career decisions in research institutions.

The cross-disciplinary search capabilities of the Data Citation Index enable greater future re-use of research data and new research discoveries through synthesis of data sets from different research areas.

Repositories ... should provide or require suffi t metadata for deposited data to create a formal citation to an identifi e data object with a unique access point.

DATA AND METADATA DISSEMINATION AND CURATION

A number of standards for the accreditation of data repositories have recently been proposed or have come into use. These include The European Framework for Audit and Certification of Digital Repositories[11] and The ICSU World Data System (WDS) Criteria for Membership and Certification[12]. In a recent project, the Peer REview for Publication and Accreditation of Research Data in the Earth science (PREPARDE) group created guidelines for repository accreditation in the context of publishing and peer review of data papers submitted to data-specific journals[13]. The various guidelines put forward share certain common elements, yet may differ where their approach to repository accreditation is dependent upon the requirements and interests of certain stakeholders.

REPOSITORY/PUBLISHER/DATA PROVIDER BEST PRACTICES

- Curate and validate metadata for completeness, accuracy, and consistency
- Issue permanent IDs for data objects
- Provide unique landing pages for data objects
- Maintain detailed update information and practice versioning
- Indicate data resource type in metadata
- Ensure clear attribution for data objects
- Document the repository mission and policies for inclusion

Where PREPARDE has created guidelines for accreditation with a view to repositories as data publication partners in the context of traditional journal publications and data journals, our recommendations here reinforce best practices toward straightforward and unambiguous data citation and discovery. These practices include the use of unique and persistent identifiers, clear attribution, rich metadata, and metadata curation, in addition to sufficient funding and other considerations that show evidence of a commitment to future data preservation and hence continued citation to an extant record or object. Citations and mentions of data repositories and their constituent data objects in published literature provide further evidence that the data are valuable to the academic community and other users of Web of Science.

The repositories included in the Data Citation Index should provide or require sufficient metadata for deposited data to create a formal citation to an identifiable data object with a unique access point. Where dynamic data are concerned, there should be clear practice in place to identify the data used through dates or versioning to assist reproducibility of results. Authors should be clearly defined, and the repository should be committed to providing these individuals and organizations with proper credit; to this end, we recommend the inclusion of the data author's institution as well as any funding information, such as funding organization or grant number. Metadata should be curated by the repository or publisher, with quality checks in place to determine whether required elements have been correctly and consistently supplied by submitting authors.

Further, the repository or publisher needs to have a system in place to issue permanent, unique identifiers to data sets to enable identification, citation, and future retrieval of the specific data object in question. New versions of data objects should be identified, with a clear distinction between version of the data and metadata. DOIs or other unique URIs that accompany metadata should resolve/link to descriptive landing pages

Benefi s of inclusion for repositories include increased visibility due to the use of the Web of Science in more than 6,500 institutions, as well as dedicated links to the repository for each data item.

where the data can be downloaded, or where the user can request access to the data. Metadata should be made available through a programmatic access point where possible, such as an OAI-PMH endpoint. Adding tags to indicate the type of media or data present aids in filtering data sets, code, and other resources in the range of nontraditional scholarly output and enables the identification and filtering of nontraditional repository-curated materials from traditional resources, such as journal publications and theses. Detailed information regarding new, updated, and deleted content records enables clarity and accuracy in future content updates. Through our collaborative partnership with DataCite, metadata submitted by data repositories for the purpose of obtaining DataCite DOIs can be routinely evaluated for harvest and for inclusion in the Data Citation Index.

Other considerations exist when building a repository for long-term data preservation at an academic institution. Currently, many institutional data repositories often consist largely of published literature by authors from that institution. A modern institutional repository that serves researchers from the sciences, social sciences, and arts and humanities will accommodate the need for long-term storage of data, software, images, videos, and more.

Benefits of inclusion for repositories include increased visibility due to the use of the Web of Science in more than 6,500 institutions, as well as dedicated links to the repository for each data item. Increased citation to the repository or publisher aids in future funding requests and metrics for repository use. While currently around 300 data repositories have been included in the Data Citation Index, not all of these achieve all of the criteria described here. In practice, dedication to detail in metadata and data curation, approaches to data set versioning, and reliability and permanence of links to data objects vary widely in this still-emerging landscape. This is also true for many of the repository accreditation criteria put forward by other groups and stakeholders. As data citation and data publishing become more commonplace, these variations will coalesce into a more unified set of guidelines that accommodates the needs of the various stakeholders involved through the work of organizations such as Thomson Reuters, DataCite, and RDA.

RESEARCH FUNDING, PUBLISHING, AND ASSESSMENT

Clarity and specificity in data deposition and publishing guidelines aid research authors as they develop data management plans. Funding organizations and journal publishers are encouraged to make detailed requirements for data sharing available, as well as to have in place well-defined mechanisms to check for compliance with regulations. Recently, a group of journals including *Science and Nature* put forward recommendations for publishing standards to further promote data in the academic community[14]. Other recent suggestions for author incentives include rewarding researchers found to be properly disseminating data; conversely researchers not considering issues related to research data dissemination and long-term preservation could be exposed to a lack of funding or inability to publish[15,16]. Journals should work with scientists, scholars, and funding agencies to establish benchmarks and develop enforcement methods toward fulfilling stated requirements. Thomson Reuters also encourages publishers to adopt policies and procedures to accommodate and enforce the use of formal data citation and to consider the repositories they recommend for the use of best practices in data preservation and discovery.

The Data Citation Index seeks to assist publishers and funders in determining whether researchers are complying with their data requirements by enabling tracking of data re-use and citation in the research literature established in the Web of Science, as well as through data discovery using author, institution, and funding information. Developments such as those described above will enable funders to better use the resource to view research output of previously funded research projects, while institutional administrators may better assess the output of academic departments and individual researchers.

LITERATURE PUBLISHER/FUNDING ORGANIZATION BEST PRACTICES

- Create specific guidelines for data deposition
- Develop formal data citation policies
- Enforce requirements for data sharing and citation
- Establish metadata criteria to allow persistent and unique identification of data in citations

IMPORTANT PRACTICES FOR DATA CITATION

- Cite papers that describe the data in addition to, not as a replacement for, citing the data themselves.
- Cite the data in the formal bibliography or in a specific data acknowledgements section, rather than in footnotes or the methods section.
- Formal citation enables secondary services such as Thomson Reuters to more readily track the impact and value of the research (e.g., through citation counts). Thus, data can receive the same benefits of the management infrastructure for journal articles.
- Formal citation is the most appropriate way of providing the information necessary to locate and access the data.
- When citing data, use a recognised citation style – either as required by the publisher or a suggested standard (e.g., Thomson Reuters, Datacite, etc.).
- Be specific. If there are several versions of the dataset, cite the exact version used in the research.
- Cite data at the finest level of granularity appropriate. Ideally this will be supported by a specific associated identifier, depending on how the data were produced/published. Where necessary, this can be supplemented in the text with more information on the specific subsets or features of the data used.
- Always include dataset identifiers where possible (e.g., DOI or Repository-assigned ID).
- Consider data as primary records of research – cite and be cited.

CONCLUSION

In order to enable increased citation, discovery, and preservation of scholarly data, further development is needed at each stage of the research data lifecycle. These developments will benefit stakeholder groups in the data community, as well as the scholarly community at large. Data authors, curators, disseminators, and funders with a stated commitment to data citation and access will demonstrate this commitment by enacting practices that promote discovery and accountability. Through our experience with research data and metadata from a variety of disciplines, as well as with a wide range of data repositories based upon a variety of different models, Thomson Reuters puts forward these recommendations to increase clarity in the mechanisms needed to uniquely identify submitted data sets and enable new scientific discoveries for scientists and scholars throughout the world.

REFERENCES

1. Stijn Hoorens, Jeff Rothenberg, Constantijn van Oranje, Martijn van der Mandele, Ruth Levitt (2007). RAND Europe Technical Report. Addressing the uncertain future of preserving the past Towards a robust strategy for digital archiving and preservation. RAND Corporation.
2. FORCE11 (2011). Joint Declaration of Data Citation Principles. [online document]. https://www.force11.org/datacitation/
3. Green, A., Macdonald, S., and Rice, R. (2009). Policy-making for Research Data in Repositories: A Guide. Version 1.2. Data Information Specialists Committee-UK. [online document]. http://www.disc-uk.org/docs/guide.pdf
4. Thomson Reuters Collaborates with DataCite to Expand Discovery of Research Data. [online press release]. http://thomsonreuters.com/press-releases/082014/datacite-research-discovery
5. Thomson Reuters Collaborates with Australian National Data Service to Raise the Profile of Research Data. [online press release]. http://thomsonreuters.com/press-releases/112013/Thomson-Reuters-ANDS
6. Piwowar H.A., Day R.S., Fridsma D.B. (2007). Sharing Detailed Research Data Is Associated with Increased Citation Rate. PLoS ONE 2(3): e308. doi:10.1371/journal.pone.0000308
7. National Science Foundation. Press release 10-077 (2010). Scientists seeking NSF funding will soon be required to submit data management plans. http://www.nsf.gov/news/news_summ.jsp?cntn_id=116928
8. ELIXIR et al. (2014). Principles of data management and sharing at European Research Infrastructures. Zenodo. doi:10.5281/zenodo.8304
9. Piwowar H.A, Vision T.J. (2013). Data reuse and the open data citation advantage. PeerJ 1:e175
10. Henneken, E.A., and Accomazzi, A. (2011). Linking to data – Effect on citation rates in astronomy. ASP Conference Series. arxiv.org. http://arxiv.org/pdf/1111.3618v1.pdf
11. European Framework for Audit and Certification of Digital Repositories. Trusted Digital Repository.eu. [online document]. (2010). http://www.trusteddigitalrepository.eu/Trusted%20Digital%20Repository.html
12. ICSU World Data System. Certification of WDS Members. Summary. 11th June 2011. [online document]. https://www.icsu-wds.org/files/wds-certification-summary-11-june-2012.pdf
13. Callaghan et al. Guidelines on Recommending Data Repositories as Partners in Publishing Research Data. International Journal of Digital Curation 2014, Vol. 9, Iss. 1, 152–163, doi:10.2218/ijdc.v9i1.309
14. McNutt, M., Journals unite for reproducibility. Science 7 November 2014: Vol. 346 no. 6210 p. 679, doi: 10.1126/science.aaa1724
15. Van den Eynden, V. and Bishop, L. (2014). Incentives and motivations for sharing research data, A researcher's perspective.
16. Kratz, John. (2015). Making Data Rain. Data Pub. http://datapub.cdlib.org/2015/01/08/make-data-rain/

ABOUT THOMSON REUTERS

Thomson Reuters is the world's leading source of intelligent information for businesses and professionals. We combine industry expertise with innovative technology to deliver critical information to leading decision makers in the financial and risk, legal, tax and accounting, intellectual property and science and media markets, powered by the world's most trusted news organization. With headquarters in New York and major operations in London and Eagan,

Minnesota, Thomson Reuters employs approximately 60,000 people and operates in more than 100 countries.

For more information, go to thomsonreuters.com.

ABOUT THE DATA CITATION INDEX

Data Citation Index on Web of Science provides a single point of access to quality research data from repositories across disciplines and around the world. Research data for this index include data studies, as well as data sets deposited in a recognized repository. Our evaluation process is always underway, with repositories added as often as weekly and existing coverage always under review.

Through linked content and summary information, this index provides researchers with critical perspective and context that is absent when data sets or repositories are viewed in isolation. Updated weekly, this index:

- Includes more than 4 million records from high-quality repositories worldwide
- Features records built from descriptive metadata to create bibliographic records and cited references for digital research
- Provides the scholarly community with standard citation formats for digital research
- Illuminates connections between primary data sets and their context, giving researchers a more complete picture of research output
- Helps users measure the contribution of digital research in specific disciplines and identify potential collaborators
- Enables researchers to discover and provide – or receive – credit for the creation of digital scholarly research data
- Provides data studies from 1900 to present

Data Citation Index is included in Web of Science Citation Connection. You can also subscribe to it on its own, or add it at no extra cost to a Web of Science Core Collection subscription.

To learn more, visit wokinfo.com/products_tools/multidisciplinary/dci/

AMERICAS

Philadelphia +1 800 336 4474
+1 215 386 0100

EUROPE, MIDDLE EAST AND AFRICA

London +44 20 7433 4000

ASIA PACIFIC

Singapore +65 6775 5088
Tokyo +81 3 4589 3102

For a complete office list visit:
ip-science.thomsonreuters.com/contact

S027314
11.2015

© 2015 Thomson Reuters

THOMSON REUTERS™

Submission Date

01/19/2017

Submitter Name

Mary Jo Hoeksema

Name of Organization

Population Association of America

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

Social and behavioral dimensions of population health

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

One of the highest priority types of data to share publicly is survey data collected from scientifically designed samples that are representative of the population and that cover multiple domains (such as health outcomes, socio-economic characteristics, and access to health services). When distributed in user-friendly formats and carefully documented, such data sets can be used by researchers from a variety of disciplines and using a variety of methodological approaches to understand health trends and the determinants of health outcomes. Two excellent examples of large NIH-funded projects that have made their data public since the projects began are the Health and Retirement Study (HRS) and the National Longitudinal Study of Adolescent to Adult Health (Add Health). HRS has over 25,000 registered users and has produced over 3,000 peer-reviewed publications since 1992. Add Health has over 30,000 users and has produced over 2,600 publications. In addition to these large surveys, smaller studies also produce valuable survey data from population-based samples that should be shared. The population field has a strong record of creating public data sets that are distributed by the projects or through widely accessible data repositories. This encourages new uses of the data and facilitates replication.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

Data should be made available forever. There is no reason to have any time limit on the amount of time data are available for secondary research purposes. Many health-related data sets in the population field have been publicly distributed for over 50 years. This has been possible because of the existence of data repositories that can back up and store the data in multiple formats and multiple locations at reasonable cost. Individual projects can rarely take on this responsibility themselves, and it is generally inefficient for them to do so. Large data repositories are able to achieve economies in the creation of technology and the acquisition of hardware necessary to make long-term storage economically feasible. They are also able to develop standardized protocols for archiving and disseminating data that maximize the value of the data to future researchers.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

Individual researchers do not generally have the experience or expertise to document, store, and disseminate the data that is collected in projects. Imposing this burden on individual projects is burdensome and inefficient and leaves data spread across a wide variety of locations in inconsistent formats. Individual researchers may also be inexperienced in how confidentiality can be protected while providing maximum possible access to the data. It is much more effective to use recognized repositories that specialize in documenting, archiving, and disseminating data. In the population field the Inter-university Consortium for Political and Social Research (ICPSR) has been one institution that has played this role since 1962. ICPSR is a consortium of over 750 academic institutions. The ICPSR data archive, based at the University of Michigan, has over 8,000 studies and 68,000 data sets, 37,000 of which are public. These data sets are archived and distributed using state-of-the art methods. Harvard's Dataverse is another important archive, with over 60,000 data

sets. Support for these types of repositories is essential if NIH is to succeed in increasing the stewardship and sharing of data. In the population field it has been standard practice to share data for many years, and PAA believes that this practice should be expanded to other fields. In order to do this it is important that researchers are advised and encouraged to include funds for the preparation of data sets in their budgets, and that reviewers and program officers come to expect that strong data sharing plans will be present in all applications.

4. Any other relevant issues respondents recognize as important for NIH to consider

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

It is important that investigators report on data and software sharing in RPPRs and competing grant applications. This has been a fairly standard practice for population researchers for many years, with reviewers paying close attention to it in competing renewal applications of projects that have collected data. While it is important that researchers should have these incentives to report in RPPRs and competing applications, there should also be incentives in other domains. While not necessarily within the purview of the NIH, institutions and professional associations should reward researchers for sharing data through their performance appraisals, promotion decisions and professional awards. This will be greatly facilitated if there are clear standards for the citation of data, as discussed below. While NIH does not directly control these mechanisms, it can help create an environment in which the production and sharing of data is considered an important and highly valued component of the scientific output of research projects.

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

Use of identifiers such as DOIs is very important and should be strongly encouraged. Emphasizing this in RPPRs and grant applications is very important. Even more important, however, is that these identifiers appear in papers and publications. Encouraging their use in reports to NIH will be one step toward encouraging their use in publications.

b. Inclusion of a link to the data/software resource with the citation in the report

Yes, this is good idea, but also in papers and publications as well as the report. If the DOI is included that should be sufficient to find the link to the data.

c. Identification of the authors of the Data/Software products

Identification of the authors of the Data/Software products is in principle a good idea. It should be noted, however, that many large long-running data collection projects (such as HRS) are produced by large teams with changing leadership and therefore do not include the names of individuals in the recommended citation for the project. It is more important that there be a clear expectation that authors include the data citation recommended by the project, (or recommended by the repository distributing the data) than that there be an expectation that an author be acknowledged.

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

This depends on the individual circumstances. In most cases it will be best to cite the distinct data sets that are being used, following the recommended citations given by the data distributor. For large longitudinal data sets that include multiple years of data and different modules, it would not ordinarily be expected that each year of data or each module would be cited separately. In some cases, however, a special module may be distributed as a unique data set with a unique DOI, in which case it is best to cite it separately.

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

This is generally not required if the information is contained within the DOI, as it is in the case of data distributed by well-established repositories such as ICPSR. For example, the DOI for the Add Health data downloaded from ICPSR is doi.org/10.3886/ICPSR21600.v17. If the repository is not embedded in the DOI, then it should be provided.

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

One of the most important incentives is to make it clear that NIH grants can and should include funds to prepare data sets for sharing. Funding should also be provided to major data repositories to assist researchers in preparing data and to take the burden off of researchers for most of the data sharing burden. NIH should also help establish a norm that data sets are properly cited and that credit is given for the production and sharing of data.

4. Any other relevant issues respondents recognize as important for NIH to consider

Additional Comments

Submission Date

01/19/2017

Submitter Name

David Lam, Director

Name of Organization

Institute for Social Research, University of Michigan

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

Social and behavioral dimensions of health; population dynamics; aging

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

At ISR we believe that all types of data should be shared in order to maximize the value of the investment in collecting the data and in order to facilitate the kinds of replication studies that are essential for the advancement of science. One high priority type of data is survey data collected from samples that are designed to be representative of some population (e.g., children under five, individuals over age 50, women of childbearing age). These surveys are especially valuable when they contain information on a wide range of variables, including rich demographic, social, and economic variables in addition to health inputs and health outcomes. We have a long history of sharing the data collected at ISR, and we built ICPSR in order to assist other researchers in sharing data. For example the NIA-funded Health and Retirement Study (HRS), which began in 1992, has over 25,000 registered users and currently is used for a peer-reviewed publication an average of every other day, with research covering a wide range of topics and methodologies. Smaller studies also have great value and should be shared.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

There is no reason to have a time limit on the period for which data should be shared. At ISR we are still sharing data sets that were collected in the 1950s and 1960s, and many researchers use these data regularly. Once a data set has been properly archived and documented in a well-equipped data repository, the cost of making it available for an indefinite period is relatively small. The big cost is is the initial investment in preparing the data for distribution. It is not realistic for individual researchers to have the responsibility of making their data available forever, however. It needs to be done using data repositories such as ICPSR that have the capacity to maintain data with constant updates in technology, secure backups, and careful attention to issues of confidentiality.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

Efficient archiving and dissemination of data is a specialized skill that is difficult for individual researchers and research projects to take on. Many researchers are appropriately concerned about protecting confidentiality of respondents when data are shared. It is not efficient or effective for individual researchers to develop procedures for protecting confidentiality in shared data. There is a great deal of expertise on these matters in data repositories, and it is best to rely on these repositories for guidance. Similarly, standards have been created for the documentation and dissemination of data. Working with established repositories ensures that these standards will be maintained. The Inter-university Consortium for Political and Social Research (ICPSR), based at the University of Michigan since its creation in 1962, is one important data repository. It is a consortium of universities and other academic institutions, with a current membership of over 750 institutions. ICPSR currently has an archive of 68,000 data sets, 37,000 of which are public. ICPSR and other large data repositories are providing essential public goods to the research community, and it is important that NIH helps support them. This support can come directly through funding for data dissemination activities such as the NICHD-supported Data Sharing for Demographic Research (DSDR) project and the NIA-supported National Archive of

Computerized Data on Aging (NACDA), both based at ICPSR, or through funding included in individual research projects that can be used to pay ICPSR or other repositories for assistance in preparing data sets for archiving and sharing.

4. Any other relevant issues respondents recognize as important for NIH to consider

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

It is very reasonable to expect that investigators report on data and software sharing in RPPRs and competing renewal applications. This should be considered a standard part of reporting on the outcomes of a project. Most NIH-supported researchers in the social and behavioral sciences already do this routinely, and this norm should be extended to other fields.

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

All researchers should be expected to use identifiers such as DOIs in reports and applications to NIH, and also in all papers, reports, and publications for broader dissemination.

b. Inclusion of a link to the data/software resource with the citation in the report

Yes, including a link to the data/software URL is a good idea in reports to NIH, and also in papers and publications. On the other hand, it is generally not essential to have the link if the DOI is included, since the DOI should be sufficient to find a link to the data.

c. Identification of the authors of the Data/Software products

The general guidance should be that the data or software should be cited using the citation that is recommended by the producers or distributors of the data. This may or may not include the name of the authors. If the recommended citation includes the names of the authors, then that should be used in all reports and papers. Producers and distributors of data and software should provide a recommended citation as well as a DOI or other standard identifier.

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

As a general practice it should be recommended that investigators cite the distinct data sets that are being used, following the recommended citations given by the producer or distributor of the data. For large data sets that include multiple years of data and different modules, it would not ordinarily be expected that each year of data or each module would be cited separately.

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

In the case of most large repositories such as ICPSR, the DOI includes a clear reference to the repository and can be used to establish the exact version of the data being used. So use of a DOI will usually mean that an explicit reference to the repository is not necessary. If the DOI does not make this clear, or if a DOI is not available, a direct reference to the repository will be useful.

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

An important way to incentive data sharing is for NIH to give a clear signal that grants can and should

include funds to prepare data sets for archiving and dissemination. As noted above, it is also important that NIH helps fund major data repositories to assist researchers in preparing and disseminating data. Another useful contribution of NIH would be to help support the expectation that data sets are properly cited and that credit is given for the production and sharing of data.

4. Any other relevant issues respondents recognize as important for NIH to consider

Additional Comments

Submission Date

01/19/2017

Submitter Name

Steven M. Girvin

Name of Organization

Yale University

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

Immunobiology, Cancer translational research

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

4. Any other relevant issues respondents recognize as important for NIH to consider

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

b. Inclusion of a link to the data/software resource with the citation in the report

c. Identification of the authors of the Data/Software products

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

4. Any other relevant issues respondents recognize as important for NIH to consider

Additional Comments

NIH-RFI-Jan2017-SMG-YaleUniversity.pdf (270 KB)

Yale University

Office of the Provost
P.O. Box 208333
New Haven, CT 06520-8333

Express delivery address:
2 Whitney Avenue, Suite 400
New Haven, CT 06510-1220

Steven M. Girvin
Deputy Provost for Research
Eugene Higgins Professor of
Physics and Applied Physics

email: steven.girvin@yale.edu
Phone: (203) 432-4448
Fax: (203) 432-0161

January 19, 2017

NIH Request for Information on Strategies for NIH Data Management, Sharing, and Citation
http://osp.od.nih.gov/content/nih-request-information-strategies-nih-data-management-sharing-and-citation

The following comments refer primarily to the sharing of human genomic data and are directed at Section I, Data Sharing Strategy Development Questions 3 and 4 in the RFI, (3) Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers; and (4) Any other topics respondents recognize as important for NIH to consider.

As the NIH and other government funding agencies embrace research data sharing, they must be prepared to provide the financial support for the sharing, access, and long term archiving of research data and code that underlie published results, and the precipitating efforts around standardization and curation during data collection. The NIH should consider the following as it develops a data sharing strategy: Assessment, transparency, and integration.

Establishing criteria to assess the data when developing strategies, policies, and infrastructure for data sharing and grant reporting is essential. A chief criterion is the value of the data. This could be determined in terms of its usefulness for future research (for example, data that are highly standardized) or in terms of cost of generating the data (for example, data from a precious sample or expensive instrumentation). Second, whether data include private or confidential information since that greatly determines whether and how the data can be shared.[1] Clear guidelines about how to handle private data, and the cost of doing so, are useful for researchers and institutions. This would include recommended language for consent forms for human subjects research and consideration of feasibility of withdraw of consent by a study subject. Third, an assessment of whether data support a traditional scientific publication (e.g., journal article), and whether they are linked with any publication(s) as that indicates the extent of the digital corpus associated with the data and the potential complexity of stewarding this corpus.

At present, many PIs and institutions are challenged by data sharing because of the extensive requirements and the expenses entailed in reporting. The NIH in concert with the scientific community should work toward clarifying what constitutes adequate financial support for sharing data and determine a framework to allocate costs fairly among stakeholders. This includes expenses related to the long-term maintenance of deposited datasets to maintain the usability of collected data after the termination of the funding period. The NIH holds information about data sharing activities and cost and should consider making this information public. In particular, the NIH can educate the community about the cost of data annotation and curation which is necessary to ensure their continued usability and

[1] See D Greenbaum, M Gerstein (2013). Proceed with Caution. The Scientist 27:26 (1 Oct.) http://www.the-scientist.com/?articles.view/articleNo/37592/title/Proceed-with-Caution/

long term preservation (in addition to the cost of sequencing, computation, and storage)[2]. This information can inform PIs' data management decisions, increase the accuracy of data sharing cost estimates in data management plans, and overall improve how data sharing is conducted. Furthermore, the NIH can also release information to the community about the cost of *not* sharing data in cases where privacy and confidentiality are a concern or where data generation or maintenance costs cannot be justified.

With respect to integration, we encourage the NIH to consider a model where data submission is more closely integrated with manuscript submission. Some examples of this exist, such as the NIH requirement that a secondary source ID (https://www.nlm.nih.gov/bsd/mms/medlineelements.html#si) is included in the manuscript, or that a study's Gene Expression Omnibus [GEO] accession number is linked to submitted manuscript files. However, the research objects deposited in various distributed data repositories may not be linked back to the publication, and in any case generally require a separate workflow. Establishing one workflow that captures all digital objects related to a study would make it easier for the PI. A single workflow can be implemented while maintaining the current distributed journal and repository system[3] and help advance the community toward a more unified policy.[4] The NIH support of PubMed Central is a step in that direction but more can be done to clarify whether and how NIH Commons and other NIH-approved data repositories are integrated to PubMed Central or other journals. A sophisticated execution of this approach might be an updated journal article which is more like a "mineable dataset"; this would entail working with publishers toward updating the current journal publishing system to allow for computer parsing of papers and machine readable standards.[5] A meaningful and sustainable technical integration of these digital materials could facilitate a coordinated stewardship of traditional scientific publications and the data that underpin such published work, which can help ensure that the data (and code) are made available for as long as the publication is available. Furthermore, to the extent that a single workflow can be technically linked to the digital grant materials (e.g., proposal, report) the NIH can share a more complete view of data management plans and estimated costs vis-à-vis actual activities and costs.

I would like to thank Mark Gerstein, Ruth Montgomery, and Limor Peer of Yale University for their help preparing this comment.

Steven M. Girvin

[2] See P Muir, S Li, S Lou, D Wang, DJ Spakowicz, L Salichos, J Zhang, GM Weinstock, F Isaacs, J Rozowsky, M Gerstein (2016). The real cost of sequencing: scaling computation to keep pace with data generation. Genome Biol 17: 53. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4806511/

[3] For example, Dryad offers integrated manuscript and data submission for several journals (see: http://wiki.datadryad.org/Submission_Integration:_Overview).

[4] An example of the need to work in concert with publishers is the "policy clash" between the Gates Foundation and leading scientific journals around open access (see: http://www.nature.com/news/gates-foundation-research-can-t-be-published-in-top-journals-1.21299).

[5] See, for example, KH Cheung, M Samwald, RK Auerbach, MB Gerstein (2010). Structured digital tables on the Semantic Web: toward a structured digital literature. Mol Syst Biol 6: 403. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2950080/

Submission Date

01/19/2017

Submitter Name

Susan Redline

Name of Organization

Brigham and Women's Hospital

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

Sleep Medicine

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

Priority should be given to sharing data to encourage new discovery--ie, many studies come to completion without fully exploring the potential number of questions or apply newly developed or emerging techniques. This relates to identifying new prediction models, signatures of disease, etc. Data sharing also provides the ability to aggregate data to get greater power, diversity, and to conduct subgroup analyses for "precision medicine" and explore issues such as sex-specific effects. Data sharing also enhances transparency and supports training -attracting new people to the field. Data that can be defined clearly, are from samples under-represented in research, and can be linked to an array of exposure-response data. In < 3 years. the National Sleep Research Resource has shared >2M files (>50TB) --indicating a thirst for physiological signal data. These data hold important information on physiological function with excellent temporal resolution, and support development and testing of new "biomarkers" such as blood pressure dipping from finger pulse data, cardiac health from patterns of HR variability and conduction, and sleep fragmentation. I suggest value be apprised as: a) ease by which users can access and understand data (and harmonized across studies); b) need for large sample sizes to obtain statistical power: c) need for diversity to address precision medicine or disparities research; d) level of validity of the original measurements and their informativeness for disease mechanisms; e) ability to address gaps in literature; f) utility for spurring innovation (new algorithms, methods); g) appropriateness for training and attracting new investigators to a needed field.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

Ideally, unlimited. Data should be stored using well developed standards, processed and supported by open source cloud based tools and storage. Specific repositories should be developed/supported via grants and contracts to meet various goals, and then linked through a suprastructure that allows cross linking of data across individuals common to more than one repository and allow cross talk of tools and synergies of efforts. Grantees should be required to budget for robust data collection/documentation (which ultimately influence the quality of shared data) as well as costs to share data. Grants reviews should require a data sharing plan that is resourced (ie, identify budget for this allocation), which includes both cost of data prep as well as cost to the repository. A cost recovery model can be used, with non-academic users paying fees to offset costs of the platform. Tech companies (google, apple, etc) may be interested in helping support some of these repositories.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

Tools, platforms need to be developed with community input, using open source code and tools. Strict security procedures need to be implemented. Investigators need to be required to share data (after some short "honeymoon") if the collection of data was made with use of tax dollars. Common language to permit data sharing of anonymized data should be required in many consent forms (and standardized templates be provided). The opportunities to increase citations,develop new collaborations, have the community use tools you developed, and participating in a learning

environment are incentives. Journals may require sharing (as some genetics journals do) Some people do not want to share data because their data were not collected in well standardized formats. Better requirements for funded research on the data capture side is needed. It is not acceptable for a major study to report it does not have a well developed data dictionary.

4. Any other relevant issues respondents recognize as important for NIH to consider

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

Developing and sharing such products should be reported. In addition, analytics on how widely these tools/resources are used may be metrics that evaluated in competing applications.

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

This would be of high use.

b. Inclusion of a link to the data/software resource with the citation in the report

Very useful

c. Identification of the authors of the Data/Software products

Appropriate recognition is important.

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

Allocating resources within the budgets of grant proposals-- the costs of high quality data capture and documentation should be supported, as should the costs associated with sharing data. Investigators should be asked to budget for these (just as they may be asked to budget for a DSMB meeting) The R24 mechanism has been invaluable in supporting scientific teams to develop important resources, which by sharing data with 100's to 1000's of individuals markedly enhance "return on investment"

4. Any other relevant issues respondents recognize as important for NIH to consider

Additional Comments

Submission Date

01/19/2017

Submitter Name

Ashok Krishnamurthy

Name of Organization

University of North Carolina at Chapel Hill

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

Data Science

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

High priority types of data to be shared include any human-related data that could be used in interdisciplinary investigations such as environmental health data or cohort studies. The value in sharing these data is to improve upon research that directly impacts human health such as identifying the effects of changes in the environment on human health. Additionally, any data that are costly to re-collect would be valuable to share, as sharing these data will save resources by reducing the potential for duplicative data collection and reduces burden on participants. Other high priority data include simulated and real data targeted specifically for assay/model development and validation; data underlying scientific claims, which allows for verification and extension of results and fosters open science and research transparency; and, data that have high reuse potential (i.e. well-documented quality data sets) that will allow for secondary analysis to answer new and exciting research questions, provide a resource for graduate students who may not have resources to collect new data sets, and act as a gold standard for new researchers.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

Data should be available for as long as possible. Investigators tend to collect more data than is actually used, and sharing these data allows other researchers to make more findings with the original data set. Additionally, there is already evidence of recently published papers (example, see PMID 25693567) using data that was shared from years ago. Allowing for long term data access, however, will require significant funding for setting up and securing database(s), and for staff to run and maintain the data. NIH should plan for delivery, preservation, archiving, and long-term access, which would include networking and computing power. Ideally, there will be more than one trusted repository for this data, and institutes should have the ability to add public mirror sites, as well as create authentication of data providers and synchronization tools. Setting up mirrors and versioning of data instances with the default pointing to the most current version will allow for preservation of the integrity of the original data. Trusted repositories may use multiple strategies to provide reliable long-term access to research data including: membership based models, sponsorship by external stakeholders (including grant funders or institutional units), or charging use fees for data access. These trusted repositories should also develop service level agreements (SLA) that describe up-time, accessibility, disaster recovery, and other policies. These repositories should also have retention policies that stipulate when and why data might be removed (such as data that has become unusable, has not been used for a specified period of time, etc.).

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

There are many costs associated with data stewardship, including investments in cyberinfrastructure, reproducibility and sustainability, as well as expert design and training to support these activities. Overcoming these barriers necessitates establishing and providing best practices for data sharing and management. Costs specific to sustainability could be overcome by encouraging the inclusion of data management and sharing costs in budget justifications. Institutionalizing data citation as a normative practice (i.e. considering data sharing in tenure reviews) could further incentivize

stewardship and sharing. Other barriers include lack of standardized licensing and technical practices to adequately address data-use agreements and privacy concerns. Technical practices employed to overcome data-use and privacy concerns include but are not limited to: authentication, authorization, and encryption. Some of these issues can also be overcome by algorithms to de-identify data, summarize, aggregate, or remotely process sensitive data, or to expose results-only data to researchers not on the IRB or outside of the HIPAA-compliance restrictions. Establishing and publishing guidance for licensing, providing options that consider different intellectual property contexts, may also help overcome some of these barriers.

4. Any other relevant issues respondents recognize as important for NIH to consider

Federal funding agencies should support efforts to convene key players to identify and harmonize standards on roles, attribution, value, and transitive credit in an extensible framework. Federal agencies should also endeavor to develop a workforce of archivists, librarians, data scientists, and other information professionals to support data sharing and preservation infrastructure. Some portions of responses herein excerpted from: Ahalt, S., Carsey, T., Couch, A., Hooper, R., Ibanez, L., Idaszak, R., Jones, M.B., Lin, J., Robinson, E. (2015). NSF Workshop on Supporting Scientific Discovery through Norms and Practices for Software and Data Citation and Attribution, ACM Digital Library, http://dl.acm.org/citation.cfm?id=2795624 (report freely downloadable here: https://softwaredatacitation.renci.org/Workshop%20Report/SoftwareDataCitation_workshop_report_2015_April_20_with_logo.pdf).

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

Major impacts of increased reporting include: PIs and institutions becoming more widely recognized and competitive for the quality of data shared; enabling analysis of shared data sets and software to see how much these products has been looked at, used, cited, etc.; by citing interim data and software releases and intermediate versions, including alpha and beta versions, the PI's visibility increases through expanded product portfolio and allows for credit to researchers involved in intermediate versions; enabling sharing of negative results with a process that may not otherwise be publishable in a peer-reviewed journal, saving time and effort for other NIH researchers. Additionally, increased reporting of data and software sharing could allow for institutionalizing the practice of viewing data and software as "standalone research products," which would incentivize young researchers who are working on expanding their CVs and creating competitive grant applications to increase their visibility and research impact through data sharing. Furthermore, this would legitimize the time and effort it takes to share data through tangible career benefits. Finally, increased reporting of data and software sharing would create added enforcement through other stakeholders (i.e., publishers and journals), because they would be encouraged to standardize the practice of data citation, which will further incentivize data sharing, and further raise the number of researchers who are complying with data sharing policies.

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

If recognized and trusted digital repositories are used, then the operation of assigning and/or using a Persistent Unique Identifier with a DOI to a data or software artifact should be straightforward with very few clicks to achieve. NIH should also focus its efforts in either conversion to a single and universal standard of citation, or develop a cross-reference to easily associate data across different DOIs.

b. Inclusion of a link to the data/software resource with the citation in the report

Authors should be able to cite data and software in their articles at an appropriate level of granularity, whether the whole artifact, a data set, or a specific relevant portion, such as a DOI, (which would point to the source institution of the mirrored data) was used. NIH should request that publishers and repositories interlink their platforms and processes so

that article references, data sets, or software citations all cross-reference each other.

c. Identification of the authors of the Data/Software products

Metadata for Data/Software products should require and use contributor/researcher identifier (e.g. ORCID) to uniquely identify authors of data/software products and associate product identifiers (e.g. DOIs) with author identifiers. Furthermore, NIH should encourage inclusion of data/software citations for authored products in NIH biosketches (and other biosketches managed by NIH's SciENcv tool).

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

Data citations should allow for the ability to cite at multiple levels of granularity, such as a single data point, a data set, a series of data sets, or a data repository. These citations standards need to be able to point as precisely as possible to the digital object(s) that was used, whether an aggregation of data was used or a distinct data set. Some repositories support the creation of unique data citations for individual data sets (example UNF in Dataverse), and could be used as a basis for developing these data citation standards. NIH should also consider that minting DOIs is not cheap, and should be wary of creating DOIs if there is not a use case for individually citing that digital object.

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

To aid in identifying and citing digital repositories, every organization should have a permanent institutional identifier (e.g., like ORCID, but specific for institutions), and/or NIH could request that the research community develop a primary and consistent data and software citation record format (e.g., analogous to BibTex or Reference Manager {RIS} bibliography formats used in journal publishing) to support data and software citation. This primary citation record format should include the repository name, and could be accomplished with a DOI that links to the repository (via a URL to the repository), and institutional ID if available.

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

NIH could strengthen and incentivize data and software sharing in several of the following ways: create additional funding mechanisms such as supplements that will be awarded when data sharing/management plans are successfully enacted; in grant applications, NIH should explicitly request a budget line item for data sharing and management, so that PIs are aware of additional funding that will be provided; and, take part in discussions with other stakeholders on integrating data citation best practices into the publication landscape. NIH should also engage in data citation format efforts and contribute to the development of an authoritative standard, as well as adopt and enforce an authoritative standard in biosketches and reporting mechanisms. The NIH should also incorporate new and more explicit guidelines for how to cite data within NIH Biosketch instructions. Agencies, publishers, societies, and foundations could fund implementation grants to identify and measure data and software impacts in a way that is relevant to stakeholders and research communities. Additionally, professional societies, such as the Association for Computing Machinery or the Institute of Electrical and Electronic Engineers, should lobby university provosts to recognize software and data in tenure processes.

4. Any other relevant issues respondents recognize as important for NIH to consider

Federal funding agencies should support efforts to convene key players to identify and harmonize standards on roles, attribution, value, and transitive credit in an extensible framework. Federal agencies should also endeavor to develop a workforce of archivists, librarians, data scientists, and other information professionals to support data sharing and preservation infrastructure. Some portions of responses herein excerpted from: Ahalt, S., Carsey, T., Couch, A., Hooper, R., Ibanez, L., Idaszak, R., Jones, M.B., Lin, J., Robinson, E. (2015). NSF Workshop on Supporting Scientific Discovery through Norms and Practices for Software and Data Citation and Attribution, ACM Digital Library, http://dl.acm.org/citation.cfm?id=2795624 (report freely downloadable here:

https://softwaredatacitation.renci.org/Workshop%20Report/SoftwareDataCitation_workshop_report_2015_April_20_with_logo.pdf).

Additional Comments

UNC Contributors to Data Management RFI.pdf (19 KB)

RFI Collaborators at the University of North Carolina at Chapel Hill

Renaissance Computing Institute:

- Ashok Krishnamurthy – Deputy Director
- Ray Idaszak – Director, Collaborative Environments
- Jay Aikat – Chief Operating Officer
- Kimberly Robasky – Translational Data Scientist
- Kira Bradford – Postdoctoral Research Associate

Odum Institute for Resaerch in Social Science:

- Thomas M. Carsey – Director
- Jonathan Crabtree – Assistant Director of Cyberinfrastructure
- Thu-Mai Christian – Assistant Director of Digital Archives

Health Sciences Library:

- Barrie Hayes – Bioinformatics and Translational Science Librarian

Submission Date

01/19/2017

Submitter Name

Virginia Steel, Univ. Libraria

Name of Organization

UCLA Library

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

All areas of research at UCLA

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

4. Any other relevant issues respondents recognize as important for NIH to consider

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

b. Inclusion of a link to the data/software resource with the citation in the report

c. Identification of the authors of the Data/Software products

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

4. Any other relevant issues respondents recognize as important for NIH to consider

Additional Comments

Submission Date

01/19/2017

Submitter Name

Kacy Redd

Name of Organization

AAU, APLU, and COGR

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

all research domains

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

As recommended by the National Academies in their report, Sharing Clinical Trial Data, NIH should convene multi-stakeholder groups to determine the infrastructure, policies, and workforce needed for responsible data sharing. The convenings should include committees of faculty within different subfields to help determine what data is truly valuable and for what time limit within their subfield/discipline. In instances where data may have value to multiple disciplines, it will be important to convene multidisciplinary stakeholder groups. High priority should be given to those data sets that are deemed to be most useful to the research community and public. NIH should focus on supporting those communities that have self-organized to share data; they have done so because that data is valuable enough for their efforts. Examples might include epidemiological data critical to public health emergencies such as Zika, genomic data (GeneBank), chemical structures (PubChem), and others. As an initial priority and as a means of quality assurance, high priority data should be defined as data that has undergone peer review, one form of which is data that underlies published articles. For the long-term, the community will need to identify, assess, and deploy best practices and policies to ensure data sets are of high quality, discoverable, accessible, and usable. High priority data should also include data obtained from rare, unique and shared equipment and large research projects which involve multiple investigators (i.e. Hadron collider, telescopes, satellites) and data collected by federal agencies as on Data.gov.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

There is such variability in data types and potential usefulness of data that NIH should allow researchers in their Data Management Plan to set reasonable lengths of time data will be available. As for long-term maintenance of data, there should be consortia of the federal government and universities to manage a few repositories. The costs for managing these repositories should be supported by the research sponsors.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

The monetary burden of first deposit into the repository should be part of the original award and should be direct costs on the grant. These costs are not insignificant; the Royal Society estimated that data sharing could require resources on the order of 1-10% of a funded project. The burden of cost for long term storage past 5-10 years should not be on the primary researcher/institution, but should have a sustainable funding model, either funded direct (charge to the grant or federal repository), recoverable indirects, or to the secondary user. One concern for long-term storage of data is the sustainability of repositories. Having a few repositories that are federally funded and managed, perhaps in partnership with universities and the private sector, may help ensure that deposited data continues to be accessible to the public. It will also be important to ensure there is harmonization in data sharing standards, policies, and practices across federal agencies to reduce administrative burden on research projects. This recommendation is in line with Executive Order 13563 of January 18, 2011 which called for greater coordination across agencies to reduce costs and simplify and harmonize rules; the National Academies report “Optimizing the Nation's Investment in Academic Research: A New

Regulatory Framework for the 21st Century"; and the GAO report entitled "Opportunities Remain for Agencies to Streamline Administrative Requirements" (GAO-16-573: Published: Jun 22, 2016).

4. Any other relevant issues respondents recognize as important for NIH to consider

Faculty will need training and support to share data responsibly, to ensure that data that has national security implications, privacy implications, or which are proprietary are handled appropriately. There is also the concern that while on its own, an individual data set may not have national security or privacy issues, when combined with another data set it becomes a risk (i.e. mosaic effect). There should be some oversight to help ensure that this does not occur. There needs to be investment in the development of standards, governance, metrics, data planning practices, tools, and repositories needed for creating high-quality, reusable data and documentation for the public and scientific community. The Association of American Universities (AAU) and the Association of Public and Land-grant Universities (APLU) have created a working group comprising campus representatives, including provosts, senior research officers, chief information officers, librarians and compliance officers to review issues relating to opportunities and challenges for universities as they move to implement new public access requirements. We look forward to working with NIH, the OSTP, and other Federal research agencies, over the next few months to further clarify data sharing standards and policies. The opinions outlined above are on behalf of the Association of Public and Land-grant Universities (APLU), the Association of American Universities (AAU), and the Council on Governmental Relations (COGR). Any follow-up questions regarding these comments may be directed to: Tobin Smith, AAU (toby_smith@aau.edu); Kacy Redd, APLU (kredd@aplu.org); or Jacquelyn Bendall, COGR (jbendall@COGR.edu).

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

Researchers may not want to share data sets until after accepted publication of their research, which might occur after the official end of the grant. Interim reports would not capture many of these post-publication data sets. Additionally, sharing data as an "interim research product", perhaps released before peer-reviewed publications, may mislead research. Interim data releases could have data whose interpretation and analysis change once the full data set becomes available. We recommend allowing the researcher flexibility in managing interim data releases, including an option to hold the data until publication or some reasonable time period after the end of a project.

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

Requirement of including a DOI is reasonable.

b. Inclusion of a link to the data/software resource with the citation in the report

Requirement of a link to the data set is reasonable.

c. Identification of the authors of the Data/Software products

Acknowledgement of author contributions is important to the scholarly enterprise. If reporting this information to NIH will be made visible to a wider community bestowing benefits and credit to each author, then it is a worthwhile and reasonable request and may outweigh the associated burden of reporting this information to NIH.

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

It seems reasonable that data that underlies a publication should be individually cited and reported.

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

Providing the link to the data set as suggested in #b would seem to be sufficient, but providing the name of the repository and link to the home page would not be overly burdensome.

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

In new applications, there could be a checkbox asking if the researcher has made all applicable data from past awards available. NIH could convene journal editors, who have an important role to play in incentivizing data sharing, and help facilitate conversations around harmonizing their data sharing guidelines.

4. Any other relevant issues respondents recognize as important for NIH to consider

Faculty will need training and support to share data responsibly, to ensure that data that has national security implications, privacy implications, or which are proprietary are handled appropriately. There is also the concern that while on its own, an individual data set may not have national security or privacy issues, when combined with another data set it becomes a risk (i.e. mosaic effect). There should be some oversight to help ensure that this does not occur. There needs to be investment in the development of standards, governance, metrics, data planning practices, tools, and repositories needed for creating high-quality, reusable data and documentation for the public and scientific community. The Association of American Universities (AAU) and the Association of Public and Land-grant Universities (APLU) have created a working group comprising campus representatives, including provosts, senior research officers, chief information officers, librarians and compliance officers to review issues relating to opportunities and challenges for universities as they move to implement new public access requirements. We look forward to working with NIH, the OSTP, and other Federal research agencies, over the next few months to further clarify data sharing standards and policies. The opinions outlined above are on behalf of the Association of Public and Land-grant Universities (APLU), the Association of American Universities (AAU), and the Council on Governmental Relations (COGR). Any follow-up questions regarding these comments may be directed to: Tobin Smith, AAU (toby_smith@aau.edu); Kacy Redd, APLU (kredd@aplu.org); or Jacquelyn Bendall, COGR (jbendall@COGR.edu).

Additional Comments

Submission Date

01/19/2017

Submitter Name

Shiqiang Tao

Name of Organization

University of Kentucky

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

Sleep, Epilepsy, Cancer, Clinical Trials

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

4. Any other relevant issues respondents recognize as important for NIH to consider

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

b. Inclusion of a link to the data/software resource with the citation in the report

c. Identification of the authors of the Data/Software products

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

4. Any other relevant issues respondents recognize as important for NIH to consider

Additional Comments

Submission Date

01/19/2017

Submitter Name

Tim Clark

Name of Organization

FORCE11

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

FORCE11 is broadly concerned with digital scholarship and sharing / citation of research data.

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

See attachment.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

See attachment.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

See attachment.

4. Any other relevant issues respondents recognize as important for NIH to consider

See attachment.

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

See attachment.

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

See Attachment

b. Inclusion of a link to the data/software resource with the citation in the report

See attachment.

c. Identification of the authors of the Data/Software products

See attachment.

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

See attachment.

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

See attachment.

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

See attachment.

4. Any other relevant issues respondents recognize as important for NIH to consider

See attachment.

Additional Comments

FORCE11 response to NIH RFI NOT-OD-17-015 FINAL 20170119.pdf (200 KB)

FORCE11 Submission in Response to NIH Request for Information (RFI)
Strategies for NIH Data Management, Sharing, and Citation, NOT-OD-17-015
URL: https://grants.nih.gov/grants/guide/notice-files/NOT-OD-17-015.html

19 January 2017

Prepared by Tim Clark[1,2], Helena Cousijn[3], Daniel S. Katz[4] and Martin Fenner[5]

[1] Massachusetts General Hospital, Boston MA
[2] Harvard Medical School, Boston MA
[3] Elsevier BV, Amsterdam NL
[4] University of Illinois, Urbana-Champaign IL
[5] DataCite, Hannover DE

FORCE11 (http://force11.org) is an international community of over 2,000 members dedicated to advancing research communications and e-scholarship.

This document summarizes expert views developed through FORCE11's Data Citation Implementation Pilot (DCIP) Expert Groups, FORCE11 Working Groups, and other activities on Data and Software Citation over a period of four years' sustained activity.

It has been reviewed and approved by the FORCE11 Board of Directors.

I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

Concerns of science policy bodies such as CODATA, the Royal Society and the U.S. National Academies about reliability of published scientific findings and reusability of research data [1-3] led to the development of the Joint Declaration of Data Citation Principles (JDDCP) [4, 5], which has been subsequently endorsed by over 100 scholarly organizations.

The JDDCP require archiving of primary research data in persistent stores, and its citation and inclusion in a reference list, wherever it is the basis for a published research finding [4, 5]. This occurs: (1) where findings or claims are based on the authors' primary research data; and (2) where data from other sources is input to the authors' analysis.

In our view, the first use case is critically important because mandating it will force data into persistent archives and thus support validation and verification of many research results. This is already done for certain specialist archives such as sequence, expression and protein structure data. We propose it be expanded, as proposed in the JDDCP, to all research data. Generalist data archives such as Dryad, Figshare, and Dataverse, already exist to handle such data (see re3data for a service describing these and similar data archives [6]). The second use case is important as well, e.g. for meta-analysis studies. However it is currently a less common use case than the first and the second use case presupposes the first already being implemented in practice. Therefore, the first use case is in our view the immediate policy priority. Ultimately, both use cases must be supported.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications.

Maintaining data availability for secondary research imposes both costs and benefits on the research ecosystem. This is ultimately a judgment and cost-benefit determination that varies with the type of data. Like books in libraries, data may be de-accessioned if no longer relevant.

The JDDCP determined that, regardless of the concrete persistence policy for any given dataset, the metadata and its landing page in the data archive should persist. Likewise, a data citation should always resolve by default to such a landing page, from which point further navigation to the underlying data can be requested either manually or in the case of software agents, by content resolution [4, 5] .

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

There are development, implementation, and maintenance burdens for data stewardship and sharing; and these can each be subdivided into policy, infrastructure, operational, and cultural burdens or costs.

FORCE11 and its stakeholder communities have already undertaken very significant policy development work enabling common definition of data citation practices [1, 2, 4, 5, 7-9], along with implementation preparation and pilot studies work, partly funded by NIH and outlined here:

(a) document model revisions to the NISO Journal Article Tag Suite, now standardized in ANSI/NISO Z39.96-2015 [10];
(b) human and machine accessibility guidelines for cited data [5];
(c) a Data Citation Roadmap for Scholarly Data Repositories [11];
(d) a Data Citation Roadmap for Publishers [12];
(e) a model for Uniform Resolution of Compact Identifiers, which are commonly used in biomedical repositories [13];
(f) a set of Software Citation Principles [14].

These studies outline a path to achieve comprehensive research transparency as called for in the FAIR Principles [15].

Additionally, core work undertaken by the bioCADDIE program to build a data discovery index, produced

(g) the DATS vocabulary of data discoverability metadata [16]; and
(h) DataMed, a pilot data discovery index [17].

Additional costs will be associated with rollout of a production version of DataMed. The bioCADDIE team, in coordination with NLM and NCBI staff, can best estimate these costs.

Early-adopter publishers such as Elsevier [9, 18, 19] (https://www.elsevier.com/about/press-releases/science-and-technology/elsevier-implements-data-citation-standards-to-encourage-authors-to-share-research-data) and SpringerNature have helped to define practical requirements at the journal level. This requires supplementation by continued outreach to other publishers and repositories over the next period. FORCE11 has developed a three-year plan to accomplish this, outlined in a recent U13 submission to NIH.

Through NIH policies for those it funds, and working with publishers, authorship practices can be transformed. This is a cultural transformation requiring material incentives similar to those developed to populate PubMed Central.

4. Any other relevant issues respondents recognize as important for NIH to consider

We believe data citation and sharing will have many profound benefits, including significant improvements to our ability to validate and verify results, reuse data, and translatability of research results to successful pharmaceutical development. At the same time, it should be remembered that validity of data depends upon the methods used to obtain, transform and analyze it. Citation, identification and sharing of software [14] and key biological research reagents [20, 21] are essential elements in a larger sharing and joint validation culture that needs the support and sustaining focus of NIH and other research funding and policy bodies.

II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing.

FORCE11 supports increased reporting requirements of data and software sharing in RPPRs and competing grant applications as a means to incentivize data sharing. This includes coordination of standards for data sharing / data management plans with the practices we outline here. Too often such plans are mere decoration and boilerplate. Adherence should be closely verified and this can happen if NIH requires that the data needed for supporting the findings of all grant funding publications is archived in repositories that conform to JDDCP requirements and publisher requirements based on JDDCP. The FORCE11 Publishers Roadmap to Data Citation [12] discusses how publishers can determine what repositories meet archival requirements and provides sources where conformance lists can be found.

We strongly support the use of preprint repositories such as bioRXiv (http://biorxiv.org); in particular where early deposition of preprints is accompanied by coordination of data and software archiving and citation policies with such archives.

2. Important features of technical guidance for data and software citation in reports to NIH, which may include:

a. *Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)*

   The FORCE11 recommendations for data repositories recommends that all datasets and software intended for citation must have a globally unique persistent identifier that resolves to a landing page specific for that resource [11]. For software citations, the identifier should resolve to a landing page referencing both the specific version, and the software project as a whole [14].

b. *Inclusion of a link to the data/software resource with the citation in the report*

   We recommend that data and software citations are included in reports in the same way as in articles, with a persistent unique identifier as described in (a) above. This means that the data and software citations have the same structure as other kinds of citations/references, and include the following elements: author(s), title, repository, year, version and persistent identifier. More information and examples of formatted data references can be found in the FORCE11 recommendations for publishers [12].

c. *Identification of the authors of the Data/Software products*

We recommend that the authors should be included in the metadata for data and software, using ORCID IDs or other appropriate persistent identifiers for authors. We recommend that all contributors to a specific cited version of software be identified in both the landing pages and the citation.

d. *Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately.*

Both use cases are common with data citation and therefore need to be supported. While a study should cite the underlying data as specifically as possible (as recommended in the JDDCP), we also need to be able to cite a collection of data from one or more databases that was used in a study. An example use case for this is a very large number of datasets that is impractical to cite individually. Services such as Biostudies (PMID: 26700850) provide this functionality. As discussed in 2a above, individual software releases and the overall software development project should be able to be cited, with appropriate metadata that can be used to link them together for understanding their interaction and gathering project-wide citation statistics.

e. *Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed*

This information should be part of the required metadata for the data/software resource – the *publisher* property – as described in the FORCE11 recommendations for data repositories [11].

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

NIH can exercise a significant influence on publishers of biomedical research articles, as recent history has demonstrated; and on data archives. For data citation to be comprehensively adopted, an ecosystem without major gaps needs to be developed.

(a) NIH should provide sufficient funding to support required transformations of this ecosystem – including **adequately funding the enhancement of its own data repositories** to comply with the landing page metadata requirements detailed in *A Data Citation Roadmap for Scholarly Data Repositories* [11].
(b) NIH should provide adequate funding for continued outreach to non-NIH repositories, publishers and identifier/metadata providers to fully adopt and support JDDCP compliant data citation practices. This can be done through U13 mechanisms, initially through RFA-CA-16-020 BD2K Support for Meetings of Data Science Related Organizations. Needs for continuation of support and coordination meetings should be continually reassessed.
(c) NIH should work closely with ELIXIR, the EMBL-EBI, and corresponding bodies in Asia, to develop joint funding models for data archiving and citation infrastructure.
(d) NIH should consider expanding its Public Access policy beyond PubMed Central deposition of articles, to include data and software deposition in recognized JDDCP-compliant repositories. Wherever possible it should find ways to strongly incentivize archived data and software citation in primary research articles.

(e) NIH should continue to support JATS related development and reuse activities.
(f) NIH should actively fund extramural activities to further and incentivize software and research resource identification, archiving / cataloguing, and citation.
(g) In addition to requiring Data Management plans, NIH should require a specific management plan for software where it is a major project component of a funded effort. Such plans should include a requirement for reporting on reuse of the datasets and software, thereby incentivizing good data and software management and the sharing of high-quality digital research materials.

4. Any other relevant issues respondents recognize as important for NIH to consider

None at present.

Attachments

1. CODATA/ITSCI Task Force on Data Citation: **Out of cite, out of mind: The Current State of Practice, Policy and Technology for Data Citation**. *Data Science Journal* 2013, **12**:1-75. http://doi.org/10.2481/dsj.OSOM13-043
2. RoyalSociety: **Science as an Open Enterprise**. In. London: The Royal Society Science Policy Center; 2012. https://royalsociety.org/~/media/policy/projects/sape/2012-06-20-saoe.pdf
3. Uhlir Pe: **Developing Data Attribution and Citation Practices and Standards**. In. Washington DC: National Academies; 2012. http://www.nap.edu/download.php?record_id=13564.
4. Data Citation Synthesis Group: **Joint Declaration of Data Citation Principles**. In. Edited by Martone M. San Diego CA: Future of Research Communication and e-Scholarship (FORCE11); 2014 https://http://www.force11.org/datacitation.
5. Starr J, Castro E, Crosas M, Dumontier M, Downs RR, Duerr R, Haak LL, Haendel M, Herman I, Hodson S,´ JH, Kratz JE, Lin J, Nielsen LH, Nurnberger A, Proell S, Rauber A, Sacchi S, Smith A, Taylor M, Clark T: **Achieving human and machine accessibility of cited data in scholarly publications**. *PeerJ* 2015, **1**: **PMID:** 26167542
6. Pampel H, Vierkant P, Scholze F, Bertelmann R, Kindling M, Klump J, Goebelbecker H-J, Gundlach J, Schirmbacher P, Dierolf U: **Making Research Data Repositories Visible: The re3data.org Registry**. *PLOS ONE* 2013, **8**(11):e78080. http://dx.doi.org/10.1371%2Fjournal.pone.0078080.
7. Uhlir P: **For Attribution - Developing Data Attribution and Citation Practices and Standards: Summary of an International Workshop (2012)** In.: The National Academies Press; 2012: 220 http://www.nap.edu/catalog.php?record_id=13564.
8. Altman M, King G: **A Proposed Standard for the Scholarly Citation of Quantitative Data**. *DLib Magazine* 2006, **13**(3/4):march2007-altman. http://www.dlib.org/dlib/march07/altman/03altman.html.
9. Altman M, Borgman C, Crosas M, Martone M: **An introduction to the joint principles for data citation**. *Bulletin of the Association for Information Science and Technology* 2015, **41**(3):43-45. http://doi.org/ 10.1002/bult.2015.1720410313
10. NISO: **JATS: Journal Article Tag Suite, version 1.1**. In., vol. ANSI/NISO Z39.96-2015. Baltimore MD, USA: National Information Standards Organization; 2015 http://www.niso.org/apps/group_public/download.php/15933/z39_96-2015.pdf.
11. Fenner M, Crosas M, Grethe J, Kennedy D, Hermjakob H, Rocca-Serra P, Berjon R, Karcher S, Martone M, Clark T: **A Data Citation Roadmap for Scholarly Data Repositories**. *bioRXiv* 2016. https://doi.org/10.1101/097196

12. Cousijn H, Kenall A, Ganley E, Harrison M, Kernohan D, Murphy F, Polischuk P, Martone M, Clark T: **A Data Citation Roadmap for Scientific Publishers** *bioRXiv* 2017. https://doi.org/10.1101/100784
13. Wimalaratne SM, Juty N, Kunze J, Janée G, McMurry JA, Beard N, Jimenez R, Grethe J, Hermjakob H, Clark T: **Uniform Resolution of Compact Identifiers for Biomedical Data**. *bioRXiv* 2017. https://doi.org/10.1101/101279
14. Smith AM, Katz DS, Niemeyer KE: **Software citation principles**. *PeerJ Computer Science* 2016, **2**:e86. https://doi.org/10.7717/peerj-cs.86.
15. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PAC, Hooft R, Kuhn T, Kok R, Kok J *et al*: **The FAIR Guiding Principles for scientific data management and stewardship**. *Scientific Data* 2016, **3**:160018. http://dx.doi.org/10.1038/sdata.2016.18.
16. Gonzalez-Beltran A, Rocca-Serra P: **WG3-MetadataSpecifications: DataMed DATS specification v2.1 - NIH BD2K bioCADDIE** *Zenodo* 2016: **PMID:**
17. Ohno-Machado L, Sansone S-A, Alter G, Fore I, Grethe J, Xu H, Gonzalez-Beltran A, Rocca-Serra P, Soysal E, Zong N, Kim H-e: **DataMed: Finding useful data across multiple biomedical data repositories**. *bioRXiv* 2016. https://doi.org/10.1101/094888
18. Taylor M: **Data citation is becoming real with FORCE11 and Elsevier**. In: *Research Data.* 2016. http://www.elsevier.com/connect/data-citation-is-becoming-real-with-force11-and-elsevier.
19. Cousijn H, Ash E: **Making data citation a reality**. In: *Elsevier Connect.* Elsevier; 2016. http://www.elsevier.com/connect/making-data-citation-a-reality.
20. Bandrowski A, Brush M, Grethe JS, Haendel MA, Kennedy DN, Hill S, Hof PR, Martone ME, Pols M, Tan SS, Washington N, Zudilova-Seinstra E, Vasilevsky N, Initiative RRI: **The Resource Identification Initiative: A Cultural Shift in Publishing**. *Neuroinformatics* 2016, **14**(2):169-182: **PMID:** 26589523 http://www.ncbi.nlm.nih.gov/pubmed/26589523.
21. Bandrowski A, Tan S, Hof PR: **Promoting research resource identification at JCN**. *The Journal of comparative neurology* 2014, **522**(8):1707: **PMID:** 24723247 http://www.ncbi.nlm.nih.gov/pubmed/24723247.

Submission Date

01/19/2017

Submitter Name

Daniel Valen

Name of Organization

Figshare

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

Biological Sciences

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

Tabular data is one such data type of high value. Essential to this is that the files have sufficient metadata to be understood and for meaning to be taken from it. Something worth explicitly highlighting is that the object should have a descriptive title. All metadata should be both human and machine readable. Machine readable metadata should be available through open APIs or OAI-PMH functionality. It should be noted that by no means should the NIH limit high-priority data to tabular data. Valuable data varies based on discipline, and at Figshare, we host over 600 different file extension types.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

At Figshare, we fall in line with the de facto standard of (at minimum) 10 years retention. We are optimistic that a lot of the most impactful datasets will be maintained longer than this as storage prices fall and value is elucidated through citation counts, alternative metrics showing attention, and reproducibility/reuse metrics. 10 years from last use prior to moving data to another storage source for archiving or preservation is also a feature we've explored with our institutional customers.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

One of the most cited barriers to data stewardship is the cost of curation. Metadata simplicity and light touch curation workflows have provided a solution from the Figshare perspective. Curation workflows are available through our institutional offering where administrators (PIs, postdocs, librarians, repository managers, etc.) can preview, edit, and comment on pre-published research objects as part of the workflow. We also have many controls in place to allow for the release of metadata records, assign embargo periods, or create confidential files. Another barrier linked to curation is metadata association. As a generalist repository, Figshare's core requirements meet the minimum Dublin Core metadata requirements to mint a DataCite Digital Object Identifier (DOI). While we can extend metadata schemas to meet needs of different disciplines, we aim to auto populate as many fields as possible and/or automating upload and metadata assignment via our open API (docs.figshare.com).

4. Any other relevant issues respondents recognize as important for NIH to consider

Data Management Plans (DMPs) should be published alongside data for compliance purposes. Rewarding and promoting best practices for data publication as much if not more than article publication would be ideal. Data should be linked to publications, regardless of whether data is stored in the same repository or is made available from the same publishers.

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

At Figshare, we have seen a rapid increase in the number of citations of research data outputs we host over the last 3 years. The NIH should explicitly reward citation and publishing of all research outputs (including datasets and software) in the same manner. This should be made very clear to new and existing researchers they fund. The NIH should also consider alternative metrics as means of incentivizing and rewarding compliance.

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

Data and software should use DataCite DOIs as the default persistent identifier. Whilst other identifiers are acceptable, the DataCite community is creating standards for interoperability and building open tools to help analyze the impact of data and software.

b. Inclusion of a link to the data/software resource with the citation in the report

All public items should include a standard citation in adherence to the FORCE11 Data Citation Principles (https://www.force11.org/group/joint-declaration-data-citation-principles-final). All public items should also have the ability to export that citation to a suite of reference managers or share the citation via social media outlets.

c. Identification of the authors of the Data/Software products

All public items should be linked to a researcher using ORCiD (industry standard) as well as utilizing a persistent identifier linking them to their organization or institution (such as the Global Research Identifier Database or GRID https://www.digital-science.com/products/grid/). Ideally, all published research objects would also include the grant ID as part of the metadata for compliance and impact tracking.

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

Allowing individual files to be citable as well as grouping files to make a citable collection has proven to be essential as fundamentally the citer will choose the level of granularity they wish to cite.

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

The repository should be a registered repository that is recognized by services such as the NIH and re3data.org index of data repositories with the ability to filter based on features that meet the needs of the researcher(s) (such as RESTful API, ability to version data, open licensing, etc.).

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

Mandated reporting is a first step, followed by attribution and credit, wherein the NIH treats and awards article publications with equal weight to published data. All published data should have metadata fields to enter grant information so that funders such as the NIH can track compliance easily.

4. Any other relevant issues respondents recognize as important for NIH to consider

Data Management Plans (DMPs) should be published alongside data for compliance purposes. Rewarding and promoting best practices for data publication as much if not more than article publication would be ideal. Data should be linked to publications, regardless of whether data is stored in the same repository or is made available from the same publishers.

Additional Comments

Submission Date

01/19/2017

Submitter Name

The YODA Project

Name of

Organization Yale

University

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

Open data and data transparency

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

The Yale University Open Data Access (YODA) Project at the Yale-New Haven Center for Outcomes Research and Evaluation (CORE) fully supports and applauds the development of data sharing strategies by the NIH. To accelerate/maximize knowledge generated through NIH sponsored research, the YODA Project advocates for the sharing of de-identified individual patient-level clinical research data as one of the highest priority types of data to be shared. In addition to summary results, the availability of individual patient-level data from clinical research studies, including clinical trials and cohort studies, provides opportunities for evaluation of secondary endpoints or new research questions, validation of previously conducted effectiveness and/or safety research, and meta analyses. Repetitive data collection is reduced, minimizing study participants' time and effort as well as the cost of undertaking such research, and further maximizing the value of the NIH's research investments.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

The YODA Project recommends that data be shared within 12 months of study completion. Two suitable methods exist for sharing data securely. The first is through an online, publicly accessible data repository that meets accepted security criteria, such as Dryad or another form of cloud-based data storage. Alternatively, data could be shared directly to researchers on a request-by-request basis, such as through Box. In all cases, a Data Use Agreement (DUA) should be executed to ensure that external researchers employ responsible conduct with regard to the data. The DUA should detail any limitations around reuse of the data (i.e., data cannot be used for commercial or litigious purposes) and should require external researchers to commit to making no effort to re-identify patients from the data. Ideally, data would be shared indefinitely. However, this notion is highly constrained by available resources. For sustainability, there needs to be a system in place for investigators to bear some of the cost burden, including explicit and sufficient budgets as part of the clinical research funding to cover the time, effort, and expense required for data de-identification and dissemination through data sharing initiatives, particularly for study teams with fewer discretionary resources. Funding opportunities should also be made available by government agencies and non-profit organizations to support the re-use and analysis of existing data resources.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

There are several barriers to data stewardship and sharing, including: • Incentives to researchers • Preparation of metadata (including data dictionaries, syntax and/or software files) • Version control (deciding which version/portion of the data is shared) • Data storage (and costs – to data holders, to data accessors) • Data security • Data interoperability • Curated access (time and costs) • Informed consent (retrospective) • Sustainability Potential mechanisms to

overcome these barriers include: • Use of DOI for data/software to create value in citation • Publicly report metrics on use of shared data/software • Prepare resources on advanced preparation of metadata • Adoption of common data models, standardized data formats, terminology • Support the sharing of data/software at the same time as publication •

Numerous platforms are emerging that provide secure data platforms • Planning for future expansion of availability of datasets should include consideration of proposed steps for how to include data from retrospective studies. In addition, moving forward, informed consent agreements should explicitly include data reuse as a possible future use of a research participant's data. • Develop meaningful rewards/incentives for data sharing and penalties for not sharing • Develop interventions to change the culture of data sharing

4. Any other relevant issues respondents recognize as important for NIH to consider

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

We support and applaud the NIH in strengthening guidance on the citation of data, databases, and software. Increased reporting of data and software sharing in RPPRs is an effective way to incentivize data sharing, but would be strengthened if this information was also shared publicly.

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

We agree that the utilization of a persistent unique identifier to the data/software source would be valuable. However, version control may be more complex for living products or analytical code. Nevertheless, the ability to save and cite at key points in the project life cycle would allow for appropriate credit, accurate referencing, and reproducibility.

b. Inclusion of a link to the data/software resource with the citation in the report

We support the inclusion of a direct link from the report citation to the data/software resource. This feature would enable expedient and accurate identification of the source data/software and support the integrity of results reported. Accordingly, there would also be a need to ensure the location of the data/software is static. It would also be valuable for the data/software to link back to the original manuscript just as the manuscript links to the data/software.

c. Identification of the authors of the Data/Software products

We fully support the appropriate identification of data/software authors within these new data and software citations. Individuals principally responsible for the preparation and creation of data sets and software may not be the authors of the primary or secondary manuscripts; it is paramount that the appropriate credit and responsibility is attributed. However, if there is no link back to the original manuscript, manuscript authors who are not also authors of the data/software products may not be sufficiently motivated to do so.

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

We recognize the complexity of establishing uniform guidance for a multiplicity of scenarios in addressing the granularity of data citations. However, when an aggregation of diverse data from a single study results in the loss of granularity of each distinct data set, we would recommend that each underlying data set be cited and reported separately. If the aggregated data set is, in fact, simply a collection of the data sets without (significant) modification, then it would seem appropriate to utilize a singular citation of the aggregated data set.

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

We encourage avoiding any ambiguity in citing the digital repository of data/software. Direct links to the data/software are more likely to be helpful than a general link to the data repository. However, considerations should be made to ensure the data repository is still identifiable as well as sustainable.

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

The YODA Project strongly supports that the reporting of data and software sharing in RPPRs be made clearly visible in the research and healthcare policy communities through publicly accessible resources and repositories, such as through ClinicalTrials.gov or NIH RePORTER. Transparency is an important step forward in promoting the responsible and comprehensive dissemination of results of federally-funded research, and should be sufficiently publicized in order to provide a model for other organizations. Through rigorous reporting/citation policies set forth by the NIH, the availability and use of federally-funded clinical research data can be incentivized to generate new knowledge that will benefit society.

4. Any other relevant issues respondents recognize as important for NIH to consider

Additional Comments

Submission Date

01/19/2017

Submitter Name

Dan Mobley

Name of Organization

Brigham and Women's Hospital

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

Sleep, Circadian and Cardiovascular Research

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

Highest priority should be given to data that fosters new scientific discovery. As a research project manager working in the field of sleep medicine, and as the project manager for the NSRR project, I have reviewed access requests and project descriptions from researchers requesting data from the NSRR. We are just into our fourth year of the project, and already we have 2,084 users who have requested accounts. In 2016, 262 users signed Data Access and Use Agreements allowing them to download data to support their own research. The project descriptions and research questions these users submit demonstrate their desire to explore new questions that were not part of the original research projects. The value in sharing the data is that researchers have the opportunity to answer new questions, develop new techniques and algorithms to analyze data, and have the ability to get greater power and diversity to conduct their analyses.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

Data would ideally be made available for an unlimited amount of time for secondary research purposes. The NSRR has noted tremendous interest in data that was collected 20+ years ago. Data repositories could potentially be developed and maintained through grants, contracts, industry support, or access fees. The NSRR has received requests from multiple companies (Google, Verily, Apple and others) who might be willing to pay for access to robust repositories of data or support platforms that can house large amounts of data. Applicants for grants should be required to include data sharing plans and should budget for costs associated with data sharing. When a grant is awarded, investigators should be given a sufficient and predetermined amount of time to analyze the data and should then be required to share that data via a resource such as the NSRR, PhysioNet, bioLINCC, etc. Protocols should be developed to provide guidance for data sharing with the various resources or that provide guidance for establishing new types of resources.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

The cost of maintaining the infrastructure for storing and sharing data is likely the biggest barrier. Maintaining data repositories requires computing resources and technical and other staff. As noted above, a cost recovery model could be considered. Industry might be willing to pay fees to access repositories of data. Some kind of subscription service might also be considered. There are many cloud-based computing platforms that charge subscription fees to access services, software or a knowledgebase.

4. Any other relevant issues respondents recognize as important for NIH to consider

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

b. Inclusion of a link to the data/software resource with the citation in the report

c. Identification of the authors of the Data/Software products

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

The NSRR requests that all users of the resource reference the project by including the name of the resource and the grant number associated with the resource in any publication. I think requiring a citation to any repository would help increase visibility of that repository and would provide NIH with some information on how much data/software resources are being accessed.

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

4. Any other relevant issues respondents recognize as important for NIH to consider

Additional Comments

Submission Date

01/19/2017

Submitter Name

Sayeed Choudhury

Name of

Organization

Johns Hopkins University Sheridan Libraries

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

Data management

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

1) Trans-disciplinary data, especially resulting from collaboration across disciplines and creating unique knowledge constructs that answer questions about specific problems stemming from public need and not domain interest (e.g., socio-behavioral, epidemiologic, environmental, etc.). Sharing such data could have broad applications among disciplines, will be in a discipline agnostic format by design, and facilitates addressing problems from multiple dimensions; 2) Epidemiologic and disease-surveillance data, especially those that have immediate relevance such as studies on Zika, influenza, HIV to identify increase/decrease in transmission patterns to determine intervention strategies; 3) Clinical research data: facilitating iterative and combined efforts to accelerate discovery of new medicines and treatments.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

Librarians and archivists consider preservation requirements in addition to sharing requirements. Medical and health projects when properly documented and associated with publications and projects could have enduring relevance both as reference and potential re-use. We recognize the challenge of front-end preparation of data, and investment of researcher's time, dependent on size and complexity of data, and presence of PHI's in preparation for secondary use. Digital preservation is ultimately less challenging than the structure and workflows that support successful data repositories that researchers can easily populate with quality data collections. Researchers should be aided in making data available with a publication when possible, or within one year of project completion. A centralized data repository registry might help highlight datasets with current articles, but ultimately long-term preservation by approved repositories should be encouraged. In some cases, data cannot be easily de-identified (or cannot be de-identified at all due to removing analytic utility or simply cannot be as is the case with genetic data) and must be kept in secure repositories with appropriate request workflows that outline what must be in place before the data can be accessed (IRB approval, detailed proposal, secure data management plan, data disposition plan, etc.). Monitoring data quality is also a concern, perhaps facilitated by a form of peer review of deposited data. Long term preservation is a return on that investment.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

One significant burden (but one which help overcome barriers) is that researchers will need explicit information about what is required from the very beginning. This information needs to be accessible and easy to read. (Consider basic principles of usability when creating this information.) By creating examples that are easy to find (and perhaps offering options for getting help/training) this will ease the transition to compliance. Additionally, individuals are motivated with

significant evidence of either benefit or cost. Working with data management professionals at academic and research institutions may help ease some of the transition as we can function as liaisons for the researchers by holding workshops or other training and information sessions. This will take a little extra time and organization for the agency at the outset,

but will organize a network of individuals who have the current information and are located in the vicinity of the researcher. De-identifying data for public use is a significant effort as a do-it-yourself activity for researchers, with varying confidence in eliminating disclosure risk. A secure HIPAA-compliant environment is needed for researchers to deposit data with even partial PHI's, with an active vetting process by repository managers to reduce the risk, while not overly burdening appropriate research access. Such repositories may need to be centralized, or at least have highly standardized infrastructure and procedures. They should also provide free or low-cost access to data.

4. Any other relevant issues respondents recognize as important for NIH to consider

The more emphasis toward clear guidance the better, using clear forms and step-by-step guidance whenever possible. Many libraries at universities and larger research institutions have research data services that can support instruction. They are also forming coalitions to share resources with smaller institutions. NIH can continue participating in supporting those resources, as it is doing with the R25 BD2K grants.

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

For the past five years our library data management services have assisted with data management plans for NSF proposals as well as NIH data sharing plans. We found that researchers generally appreciate the results of thinking about/discussing data management planning at either the pre-proposal or post-grant stage, but need strong requirements to motivate their efforts. They also tend to recycle their plans or 'borrow' from colleagues. Thus, encouraging at least one good first effort is important, making the process efficient with good instructions and resulting in content they could implement. Active monitoring of minimal compliance by NIH, not just peer reviewers of proposals, may be necessary. When required early on, they can implement steps throughout the workflow, but may need to report on certain milestones like data release with publications. Emphasizing data security, especially for PHI data, may be a personal-interest hook to thinking about other preparations for sharing. A positive impact from increased reporting at earlier stages throughout the process is that researchers will then be “doing data management” from the beginning - they will have to implement the best practices in order to produce something shareable. On a larger scale this may prove to illuminate some of the benefits of planning that are often overlooked because they are time consuming.

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

Requiring a persistent identifier, DOI's in particular, tends to encourage researchers toward a repository infrastructure, vs. sharing directly by request. Universities operating their own repository infrastructure may need to be prepared to designate a central body to generate DOIs with a clear workflow and infrastructure. Requiring all data and software produced to be referenced in reports to NIH should strive to make the products of research unambiguously identifiable. For software, for example, a reference to a particular version might be adequate for reporting purposes, even if the software is under continuous development. If researchers have to archive their data/code for NIH reporting purposes, when the grant ends, they (a) have already archived their data/code, or (b) have become familiar enough with the process that getting a DOI for their final products shouldn’t be a big burden.

b. Inclusion of a link to the data/software resource with the citation in the report

If the use of DOIs that resolve to resources is enforced as a reporting requirement, then it also handles the issue of including links in the report itself. Again, this tends to drive researchers to data repository choices, but universities and institutions should be ready to facilitate their data sharing workflow. In the case of software, including a link to

e.g., a Github repository in conjunction with an archived version with a DOI seems prudent.

c. Identification of the authors of the Data/Software products

Our library has been providing guidance on use of ORCID ID's, but the university may ultimately encourage more broad adoption as part of centralizing a platform for faculty research and publication profiles. Datasets for NIH and other grant funded projects could ultimately be included in such profile systems, which will help researchers recognize their value.

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

Citing aggregated data is a challenge that the data curation community (e.g. RDAP) is actively working on. Data repositories may ultimately need to take on more features of relational databases, and some, such as Dataverse, have good solutions. In our institution's data archive, we have so far only collections with unambiguous citations to single collections, mainly associated with publications. This may be the norm, but ideally systems should support outlier cases of multiple citations to components of data collections and researchers should have clear instructions or support by repositories on how to accommodate such cases. Cited records should also show the provenance of data components, including toward associated papers or projects where data was first collected/described (this can help disambiguate citations). A similar method can be applied to software citations. A citation to specific instance/snapshot of the dataset/software part(s) used in a particular case may be better for reproducibility purposes. As a rule of thumb, the citation should be at the level that allows others to reproduce/validate the referenced work.

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

This is an important question that might not be well resolved by emerging standards for data citation, which generally rely on the name of the repository in the place of the "publisher" following a standard article citation. Given that registries within a data commons concept require good consistent indexing of repositories for machine readability and citation federation, some standard identifier for repositories may need to be set - perhaps instigated by NIH. This might help encourage data sharing through approved repositories or alternative means that must be vetted in the process of applying for an identifier for adequate citation.

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

NIH can find ways to contribute to researchers receiving academic credit for citable datasets and contributions to repositories, either directly or as altmetrics for impact factors. Repositories can also facilitate the writing of 'data publications' to accompany datasets as standalone short papers that might focus on dataset content and possible reuse, supplementing, but distinct from, research publications that produced the data. Researchers producing datasets that themselves have high impact and reuse could also get special recognition, publicized through professional associations or associated journals, or within the repository using 'badges' or front-end 'highlighted dataset' pages in the repository interface.

4. Any other relevant issues respondents recognize as important for NIH to consider

The more emphasis toward clear guidance the better, using clear forms and step-by-step guidance whenever possible. Many libraries at universities and larger research institutions have research data services that can support instruction. They are also forming coalitions to share resources with smaller institutions. NIH can continue participating in supporting those resources, as it is doing with the R25 BD2K grants.

Additional Comments

Submission Date

01/19/2017

Submitter Name

Sarah Brookhart

Name of Organization

Association for Psychological Science

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

Basic (theoretical) research in scientific psychology and related social and behavioral sciences.

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

All data should be shared, if possible. Open data allows other members of the research community to examine the data and independently analyze it. Open data allows for a rigorous evidence base and increases access for experimental reproducibility. Also, data sharing allows other researchers to use shared data for their own experiments, pool data across experiments, and examine a dataset in the context of the scientific conclusions drawn from the data. This allows researchers to devote resources away from redundant data collection. APS is a signatory of the Transparency and Openness Promotion (TOP) Guidelines, which states that open sharing is a core value of science (https://cos.io/our-services/top-guidelines/).

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

Once made public, data should be made available for perpetuity, unless other circumstances (e.g., unexpected ethical considerations) demand their removal. It is impossible to tell when a dataset may be needed in the future. Data for many studies in psychological science research require minimal resources to maintain; however, for larger datasets (e.g., brain imaging data), storage may become costly.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

University institutional review boards (IRBs) and other ethical guidelines can limit data stewardship and sharing, when these objectives conflict with the important objective of maintaining protections for research participants. Strong partnerships between IRBs and data sharing programs are necessary to ensure that scientists and society can enjoy the benefits of data sharing while important protections for research participants are maintained. In addition, data sharing contributes to enhanced scientific rigor and further supports research that is reproducible, robust, and transparent. Authors may hesitate to share data due to the time and effort required. Recent studies, however, suggest that authors can be encouraged to share data via small “nudges.” APS awards authors who make their data publicly available with an Open Data badge, which appears on the published article in print and online. The introduction of this badge in 2014 coincided with a 10x increase in data sharing in APS journal in which this badge was introduced (see https://www.psychologicalscience.org/publications/observer/obsonline/psychological-science-badge-program-encourages-open-practices-study-shows.html).

4. Any other relevant issues respondents recognize as important for NIH to consider

NIH may be interested to know that APS is launching a new journal dedicated to new developments in research practices, methods, and conduct called Advances in Methodologies and Practices in Psychological Science. This journal will aim to publish innovative findings and tutorials related to data management and sharing, among other issues. APS is selecting a Founding Editor of this journal and intends to publish its first issue in early 2018 (see

https://www.psychologicalscience.org/aamps-call-for-editor-nominations).

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

N/A

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

Use of a Persistent Unique Identifier, such as a DOI, is an important component of data citing. In the same way that DOIs are important for papers cited in other papers, DOIs for data and software resources allow a reliable connection between the published text and the resource. They also imply that the resource will be maintained in perpetuity.

b. Inclusion of a link to the data/software resource with the citation in the report

Inclusion of a link to a data/software resource with a citation in the report is necessary. Such linking and citation allows data/software product authors to gain credit for tools they build and spreads awareness of cutting-edge products that other researchers may find useful for their own work.

c. Identification of the authors of the Data/Software products

See above—identification of authors is necessary for providing credit and directing readers and others to important product resources.

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

It is too soon to make a recommendation on the granularity of data citations, and we recommend that authors use their best judgment in citing data. For example, authors may link to a full dataset, but then, when referencing a set of stimuli, provide a link which connects directly to the stimulus set.

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

Unambiguously identifying and citing the digital repository where a resource is stored is necessary to ease the burden of accessing that resource.

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

See above—APS's Open Badges program has been shown to incentivize data and materials sharing. Prior to 2014, few APS articles were published alongside open data; since the adoption of these badges, however, 43% of articles published in Psychological Science have provided open data, materials, or have preregistered study hypotheses. The Kidwell et al. (2016) study suggested that the introduction of badges for data sharing led to sharing rates an order of magnitude above baseline.

4. Any other relevant issues respondents recognize as important for NIH to consider

NIH may be interested to know that APS is launching a new journal dedicated to new developments in research practices, methods, and conduct called Advances in Methodologies and Practices in Psychological Science. This journal will aim to publish innovative findings and tutorials related to data management and sharing, among other issues. APS is selecting a Founding Editor of this journal and intends to publish its first issue in early 2018 (see https://www.psychologicalscience.org/aamps-call-for-editor-nominations).

Additional Comments

Submission Date

01/19/2017

Submitter Name

Ross McKinney, MD

Name of Organization

Association of American Medical Colleges

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

full spectrum of biomedical research

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

• In our discussions [with senior academic research leaders], there was no real sorting of priorities for the types of data to be shared. Ideas ranged across basic, translational, and clinical research, as well as health services and population data, and were not confined to specific fields or studies. The point most commonly and emphatically made by investigators and research leaders is the necessity to capture the totality of information required to make data useful, including documentation of context, limitations, and other metadata. Data are seldom useful absent such curation. Useful data storage needs to include relevant software or analysis code in the resource as well as the raw data. Imaging studies require documentation regarding acquisition, imaging modalities and patient parameters, in addition to other study information. These types of complex data packages will optimize the utility of the information and facilitate reproducing studies. • Negative data are especially valuable to post in repositories. While many data are not published because they negate or are inconclusive about posited research questions, they become particularly valuable in meta-analysis. The advantages to sharing negative data, especially in clinical studies, have been noted elsewhere, and are consistent with the notion that science advances one failure at a time. • Ultimately, discussants noted, we never know what data will be useful, or how it might be used in future, given unpredictable changes in science, and in the technologies that make use of data.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

• The length of time for making data available for secondary research purposes would be indefinite. It will often exceed the length of time for research projects or grants themselves, and may well exceed time that key personnel remain at an institution. This has implications, noted below, for stewardship of data, and intrinsic cost. • The nature of studies will also affect this calculation. Consider, for example, long-term longitudinal studies, where data may accumulate and be shared through the life of the project. In many other studies, the data will be posted a certain time after initial publication (our constituents preferred a calendar year). Original investigators should have sufficient time for analyzing data before making publicly available, on timeframes that may vary by type of study. • Discussants also noted that we never know what data will be of value in the future, given unpredictable advances in science and technology. An animating vision to guide NIH might be the use of shared data resources to support machine learning and specialized algorithms for searching, synthesizing and analyzing information.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

• Related to the need to curate and document data and relevant software, cost was the central concern raised by researchers. For data sharing to advance, research sponsors and institutions must commit resources. • It is not clear that the public or political leaders, who increasingly support or call for data sharing (and other “transparency”) appreciate the additional burden and cost of creating usable shared data resources. Perhaps this is because of the ease with which

other types of information can be so easily shared. Submitting data and documentation to repositories ensures preservation and accessibility of data for the research team. It also increases citation of work, increases visibility, and opportunities for new scientific collaborations. On the other hand, data sharing reduces investigator advantages when applying for grants and limits protection of publication opportunities for the research team, students and colleagues. The research community is coming to appreciate the opportunity costs and expense to society and science of not sharing data from publicly or privately financed research. That realization is ultimately the impetus for continued progress. • There is a need for both generalist data repositories, for a wide variety of data, and specific repositories. These should continue to be developed and supported by NIH, in addition to institutions or other organizations. The overall utility of these efforts is related to the commitment to standards and providing support. Investigators we spoke with also favored opportunities that facilitate creation of study specific repositories. The best designs and standards emerge from the research communities themselves.

4. Any other relevant issues respondents recognize as important for NIH to consider

• Several constituents have proposed that secondary research conducted with patient-level data should be independently reviewed for scientific merit as a condition of access. This point emphasizes again protection of risk to research subject confidentiality where identifiable data necessary for analysis, or where there is potential for re-identification. • While we tend to describe repositories as centralized resources, they can also be decentralized, federated and structured according to many types of arrangements (consistent with the “bottom-up” approach preferred by most discussants). NIH should examine various existing models, and build incrementally from those models.

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

• Citation and credit for generating and sharing data is fundamentally important as an incentive, to help recognize and advance productive investigators. Use of standard object codes and links for citation will help in recognition. Blockchain, mentioned above, can also be used to track who has accessed or made use of data. For some types of research, data generators may be viewed as research collaborators (or even authors) on a study. But as shared data resources become more routine and commoditized, data may be cited like other sources or references.

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

• While DOI is useful, NIH should encourage biomedical informaticists to develop alternative methods for standard identification (the Biomedical informatics subgroup of CTSA consortia may be helpful.) NIH and the research community should also consider developing a global tracking system for secondary publications from shared data sets. • Effective citation will help improve the rigor and reproducibility of studies, including the increased availability of negative data. • Citation will also have an impact on efforts to improve research integrity. • Discussants noted that it is necessary to change the current system, but urged accepting that such changes will take time, and encouraged the creation of an easy, consistent format that is not up to interpretation.

b. Inclusion of a link to the data/software resource with the citation in the report

[noted above]

c. Identification of the authors of the Data/Software products

N/A

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

N/A

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

N/A

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

Direct funding of repositories and infrastructure, collaborative awards with federal agencies, coordination across federal agencies (already exemplified by EOP OSTP).

4. Any other relevant issues respondents recognize as important for NIH to consider

• Several constituents have proposed that secondary research conducted with patient-level data should be independently reviewed for scientific merit as a condition of access. This point emphasizes again protection of risk to research subject confidentiality where identifiable data necessary for analysis, or where there is potential for re-identification. • While we tend to describe repositories as centralized resources, they can also be decentralized, federated and structured according to many types of arrangements (consistent with the “bottom-up” approach preferred by most discussants). NIH should examine various existing models, and build incrementally from those models.

Additional Comments

AAMC Comment to NIH- Data Management and Sharing final review.pdf (62 KB)

AAMC
Association of
American Medical Colleges
655 K Street, N.W., Suite 100, Washington, D.C. 20001-2399
T 202 828 0400
www.aamc.org

January 19, 2017

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

RE: NIH Request for Information: Strategies for NIH Data Management, Sharing, and Citation (NOT-OD-17-015)

The Association of American Medical Colleges (AAMC) is pleased to have this opportunity to offer comments related to data management and sharing strategies and priorities for the NIH. The AAMC is a not-for-profit association representing all 147 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 160,000 faculty members, 83,000 medical students, 115,000 resident physicians, and thousands of graduate students and postdoctoral trainees in the biomedical sciences.

The AAMC has long supported data sharing in basic and clinical studies, and has embraced efforts to maximize the use of data resources. We appreciate that NIH has asked the research community itself for information as it formulates broad strategies for building data resources. The following suggestions for key elements in development of a data sharing strategy and in data citation are drawn from research leaders at our member institutions. The AAMC has also helped publicize and disseminate the RFI to encourage researchers and organizations to respond directly.

RFI Section 1: Data Sharing Strategy Development:

(1) Highest priority types of data to be shared and the value of sharing such data.

- In our discussions, there was no real sorting of priorities for the types of data to be shared. Ideas ranged across basic, translational, and clinical research, as well as health services and population data, and were not confined to specific fields or studies. The point most commonly and emphatically made by investigators and research leaders is the necessity to capture the totality of information required to make data useful, including documentation of context, limitations, and other metadata. Data are seldom useful absent such curation. Useful data storage needs to include relevant software or analysis code in the resource as well as the raw data. Imaging studies require documentation regarding acquisition, imaging modalities and patient parameters, in addition to other study information. These types of complex data packages will optimize the utility of the information and facilitate reproducing studies.

- Negative data are especially valuable to post in repositories. While many data are not published because they negate or are inconclusive about posited research questions, they become particularly valuable in meta-analysis. The advantages to sharing negative data, especially in clinical studies, have been noted elsewhere, and are consistent with the notion that science advances one failure at a time.
- Ultimately, discussants noted, we never know what data will be useful, or how it might be used in future, given unpredictable changes in science, and in the technologies that make use of data.

(2) The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

- The length of time for making data available for secondary research purposes would be indefinite. It will often exceed the length of time for research projects or grants themselves, and may well exceed time that key personnel remain at an institution. This has implications, noted below, for stewardship of data, and intrinsic cost.
- The nature of studies will also affect this calculation. Consider, for example, long-term longitudinal studies, where data may accumulate and be shared through the life of the project. In many other studies, the data will be posted a certain time after initial publication (our constituents preferred a calendar year). Original investigators should have sufficient time for analyzing data before making publicly available, on timeframes that may vary by type of study.
- Discussants also noted that we never know what data will be of value in the future, given unpredictable advances in science and technology. An animating vision to guide NIH might be the use of shared data resources to support machine learning and specialized algorithms for searching, synthesizing and analyzing information.

(3) Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

- Related to the need to curate and document data and relevant software, cost was the central concern raised by researchers. For data sharing to advance, research sponsors and institutions must commit resources.
- It is not clear that the public or political leaders, who increasingly support or call for data sharing (and other "transparency") appreciate the additional burden and cost of creating usable shared data resources. Perhaps this is because of the ease with which other types of information can be so easily shared. Submitting data and documentation to repositories ensures preservation and accessibility of data for the research team. It also increases citation of work, increases visibility, and opportunities for new scientific collaborations. On the other hand, data sharing reduces investigator advantages when applying for grants and limits protection of publication opportunities for the research team, students and colleagues. The research community is coming to appreciate the opportunity costs and expense to society and science of not sharing data from publicly or

privately financed research. That realization is ultimately the impetus for continued progress.

- There is a need for both generalist data repositories, for a wide variety of data, and specific repositories. These should continue to be developed and supported by NIH, in addition to institutions or other organizations. The overall utility of these efforts is related to the commitment to standards and providing support. Investigators we spoke with also favored opportunities that facilitate creation of study specific repositories. The best designs and standards emerge from the research communities themselves.
- Unfortunately, establishing many free-standing data repositories will likely limit the utility of the data: what you can't find, you can't use. In addition, inconsistent formats may further degrade utility. Thus, there is a conundrum: one central repository with common standards (perhaps for clinical data from studies) might be possible and have real utility, but it would be likely to be too constrained for data obtained in non-standard ways.

(4) Any other relevant issues respondents recognize as important for NIH to consider.

- A principal concern for data repositories is cybersecurity. Not only is security an issue for clinical data—where the privacy of human research participants, patients, etc.,--must be protected, but is also an issue for non-human and other types of data as well, which may be subject to theft, sabotage, or alteration. New policies, such as strong legal protections, standards, and data-use agreements, as well as new technologies will help address concerns. Particular attention is being paid to blockchain technology, a data ledger used for Bitcoin, as a means to enforcing privacy and agreements, for example.
- Lead time for sharing data includes - formatting data; describing scope of consent and data usage; preparing documentation and data dictionaries; and obtaining required institutional approval. Given researcher and institutional commitments that must take place for data sharing to occur. The research community needs to create clear parameters and pathways that make the process easy and consistent. In addition, utilization of these approaches needs to be evaluated and investments should be balanced and proportional to utilization and impact.
- While we tend to describe repositories as centralized resources, they can also be decentralized, federated and structured according to many types of arrangements (consistent with the "bottom-up" approach preferred by most discussants). NIH should examine various existing models, and build incrementally from those models.

Section II: Data and Software Citation in Research Performance Progress Reports (RPPRs) and research grant applications.

(1) The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing.

- Citation and credit for generating and sharing data is fundamentally important as an incentive, to help recognize and advance productive investigators. Use of standard object codes and links for citation will help in recognition. Blockchain, mentioned above, can

also be used to track who has accessed or made use of data. For some types of research, data generators may be viewed as research collaborators (or even authors) on a study. But as shared data resources become more routine and commoditized, data may be cited like other sources or references.

- While DOI is useful, NIH should encourage biomedical informaticists to develop alternative methods for standard identification (the Biomedical informatics subgroup of CTSA consortia may be helpful.) NIH and the research community should also consider developing a global tracking system for secondary publications from shared data sets.
- Effective citation will help improve the rigor and reproducibility of studies, including the increased availability of negative data.
- Citation will also have an impact on efforts to improve research integrity.
- Discussants noted that it is necessary to change the current system, but urged accepting that such changes will take time, and encouraged the creation of an easy, consistent format that is not up to interpretation.

Other topics important for NIH to consider:

- Several constituents have proposed that secondary research conducted with patient-level data should be independently reviewed for scientific merit as a condition of access. This point emphasizes again protection of risk to research subject confidentiality where identifiable data necessary for analysis, or where there is potential for re-identification.

The AAMC appreciates the opportunity to comment to the NIH on this issue and would be happy to provide any further information moving forward. Please contact me or my colleague, Stephen Heinig, Director of Science Policy, (sheinig@aamc.org, 202-828-0488) with any questions about these comments.

Sincerely,

Ross E. McKinney, Jr, MD
Chief Scientific Officer

Submission Date

01/19/2017

Submitter Name

James A. Bryant, Ph.D.

Name of Organization

Ishpi Information Technologies, Inc.

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

Healthcare Information Technology

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

The data type defines which operations can safely be performed to create, transform and use the variable in another computation. When a programming language requires a variable to only be used in ways that respect its data type, that language is said to be strongly typed. This prevents errors because while it is logical to ask the computer to multiply a float by an integer (1.5 x 5), it is illogical to ask the computer to multiply a float by a string (1.5 x Alice). When a programming language allows a variable of one data type to be used as if it were a value of another data type, the language is said to be weakly typed. Technically, the concept of a strongly typed or weakly typed programming language is a fallacy. In every programming language, all variable values have a static type, but the type might be one whose values are classified into one or more classes. And while some classes specify how the data type's value will be compiled or interpreted, there are other classes whose values are not marked with their class until run-time. The extent to which a programming language discourages or prevents type error is known as type safety.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

Retention Period: A principal investigator (PI) is the holder of an independent grant administered by a university and the lead researcher for the grant project. PIs are responsible for determining the length of time data should be made available; the appropriate means for maintaining and sustaining the data; and analyzing the long-term resource implications of storage. Retention Periods are dictated by data relevancy, as opposed to storage requirements. The amount of time data should be stored is determined by information type. Occasionally, retention periods are at the discretion of the PIs; however, many sponsor institutions require data be retained for a minimum number of years. For instance, HHS requires project data be retained for at least 3 years after funding ends. Continued Storage: Once the minimum retention period is met, the PI decides whether to continue with data storage. PIs evaluate the benefits and risks of extended storage. PIs never know when data might be needed, and continued storage of confidential data increases the risk of possible violation, not to mention the monetary cost associated with storage. Destroying Data: When the decision has been made to end data storage, data should be thoroughly destroyed. Effective data destruction ensures that information cannot be extracted or reconstructed. Many document storage companies now offer onsite shredding and secure destruction of written and electronic records. For electronic data, software products, such as Eraser or CyberScrub, are available. All pointers related to the purged data should also be removed to avoid retrieval errors.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

Data stewardship is a collection of data management methods covering acquisition, storage, aggregation, deidentification, and procedures for data release and use. A data steward is intended to convey a fiduciary (or trust) relationship with data that requires a loyal data manager vested to the interests of the entities whose data is stored.

Barriers to data stewardship for NIH may include privacy concerns and how to securely gain access to personal identifiable data. This includes the information essential to understanding healthcare quality, safety, efficiency, and outcomes. Data ownership, versus data access, require two distinct resolution approaches. The first, consistent with the concept of health information ownership, would be to incentivize data access through payment for information. A data stewardship entity purchases the required health information considered to be essential to patient safety, quality, comparative effectiveness, and population health. With data monetized, the government could negotiate over the scope and terms of access and use. The strength of this model is the recognition of data ownership rights. The chief limitations could be the overall cost and the uncertainties that surround market negotiations. An alternative approach is to treat data as a public good. Stewardship entities could be federally chartered, with broad authority to collect, prepare, and support the use of health information in research. This model would achieve the broad goals set by advocates of evidence-driven care. In many respects, his model has taken precedence in health reform, although it is still subject to limitations on usage.

4. Any other relevant issues respondents recognize as important for NIH to consider

We believe that a Continuity of Operations Plan (COOP), complete with a back-up and restore schedule, is critical for the digital repositories focused on a data management strategy at NIH. Distributed data repositories that are replicated on a regular basis provide a high availability of access while providing disaster recovery in real time. Additionally, the COOP should include protection from ransomware so that NIH and adjoining institutions are not held hostage by rogue actors. Ransomware protection should consider and include: • Ransomware continues to proliferate and evolve because of the success of attacks securing ransom payments from high-value targets – expect more in 2017 • The growth of new sophisticated malware variants and zero-day exploits will continue and fortifying cyber defenses is critical to preventing breaches • Traditional antivirus is ineffective at stopping ransomware because polymorphic variants and obfuscation techniques easily bypass signature-based defenses • Next-generation antivirus is now a requirement to eliminate ransomware from your environment

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

Grant recipients use the RPPR, a federally mandated format for required interim or annual reporting, to complete progress reports on NIH grant awards. As a grant recipient, progress reports are required annually to document accomplishments and compliance with the terms of the award. Recipients must describe their scientific progress, identify significant changes, report on personnel, and describe plans for the subsequent budget period or year in these annual reports. NIH requires recipients to submit an RPPR for non-competing NIH awards. Funding for non-competing years of the grant can only be awarded after the NIH program and grants management staff review and approve the progress report. The review of the RPPR, by NIH staff, is a key element in NIH's monitoring of the grant award. The impact of increasing reporting of data and software sharing in RPPRs, competing grant applications to enrich reporting of productivity of research projects, and incentivizing data sharing will improve effectiveness. Although the primary incentive to input the data seems to be for continued funding, the downstream benefit of data availability to thousands of users provides a significant benefit for continued research. The veracity of data and software sharing in RPPRs will improve and provide current data that enhances research performance.

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

Technical guidance for data and software citations in NIH reports will be greatly enhanced by persistent unique Identifiers, such as the Digital Object Identifier (DOI). A DOI is a type of persistent identifier used to uniquely identify objects. The DOI system is particularly used for electronic documents, such as journal articles. DOI means "digital identifier of an object" rather than "identifier of a digital object". Thus, the term DOI stands for "digital object-identifier" rather than "digital-object identifier". Metadata about the object is stored in association with the DOI name. It may include, the object title or a location such as a URL indicating where the object can be found. The DOI for a document

remains fixed over the lifetime of the document, whereas its location and other metadata may change. Referring to an online document by its DOI provides more stable linking than simply using its URL, because if the URL changes, the publisher only needs to update the metadata for the DOI to link to the new URL. A DOI name differs from standard identifier registries such as the ISBN and ISRC. The purpose of an identifier registry is to manage a given collection of identifiers, whereas the primary purpose of the DOI system is to make a collection of identifiers actionable and interoperable. The DOI is commonly used in academia to universally identify articles, journals, and magazines. The use of a Persistent Unique Identifier that resolves resources would be a huge benefit for NIH.

b. Inclusion of a link to the data/software resource with the citation in the report

The inclusion of a link to the data/software resource with the citation is critical. The Journal of Statistical Software states software providers may offer a recommended or required, citation format for any software user. This ensures the provider's research contribution is acknowledged. For example, the R open source statistical programming language and environment provide a BibTeX entry in their FAQ that can be used if citing R: @Manual{, title = {R: A Language and Environment for Statistical Computing}, author = {{R Development Core Team}}, organization = {R Foundation for Statistical Computing}, address = {Vienna, Austria}, year = 2011, url = {http://www.R-project.org} Some providers make the citation part of the license. For example, SAS's mandated citation: The [output/code/data analysis] for this paper was generated using [SAS/STAT] software, Version [9.1] of the SAS System for [Unix]. Copyright © [year of copyright] SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA. HSL library requires the following citation to be included when describing results achieved: HSL(2011). A collection of Fortran codes for large scale scientific computation. http://www.hsl.rl.ac.uk The P-Genome bioinformatics software provides a citation for their software and publication: Sarkar IN, Planet PJ, Thornton J, DeSalle R, Figurski DH. Phylogenomenclature. Available at: http://www.dbmi.columbia.edu/~ins7001/research/CAOS/P-Gnome/PGdownload Sarkar IN, Thornton J, Planet PJ, Figurski DH, Schierwater B, DeSalle R. An Automated Phylogenetic Key for Classifying Homeoboxes. Molecular Phylogenetics and Evolution 2002 Sep; 24(3):388

c. Identification of the authors of the Data/Software products

Identification of the authors of the Data/Software products is equally important. Describing the contribution of data/software to research closely relates to several other issues around the role of data/software in research. This includes publishing research data/software in a persistent and citable way; ensuring the availability of research software (and data, online services, and other artifacts) for the long-term; promoting the recognition of software as a valuable research output; and ensuring that research software developers have their contributions recognized and rewarded. These are concerns that affect researchers using the software, as well as those who develop or modify research software; those who release research software; paper reviewers; program committees; publishers; and funders. All these agencies have a part to play and the Software Sustainability Institute is working with many different experts and groups to explore and resolve these issues. In addition, it allows data consumers to look for additional related data/software from the same author/developer, which could provide a macro perspective of the research topic and be relevant to their data usage.

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

The fact that some software consists of discrete well-defined components imposes additional requirements when describing software. As we saw with HSL, not only is a researcher's use of HSL significant, the names of the sub-routines within HSL are also required, to deliver an accurate description of the research undertaken. This is analogous to the situation faced by data publishers and consumers, as part of their research. In response to these, The Digital Curation Centre (DCC) have produced a guide on data citation and linking that discusses these problems and provides advice: Ball, A. & Duke, M. (2011). "How to Cite Datasets and Link to Publications?". DCC How-to Guides. Edinburgh: DCC. Available online: http://www.dcc.ac.uk/resources/how-guides/cite-datasets As the authors describe, data can be hierarchically structured into various elements including in databases, tables, columns, rows, directories, files, records, data points, collections, and documents. Each of these elements may or may not have a specific identifier associated with them, depending upon how the data producer has published them. The DCC guide authors recommend that authors should cite data sources at the finest level of granularity that was adopted when an identifier, or for our purposes, a citation,

was assigned. This can be supplemented in the text with the information to find any specific subset of the data that was used in the research. As researchers using HSL show, this advice can be readily adopted to describe the use of research software.

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

The data/software should be unambiguously identified and cited in the digital repository where the resource is stored and can be found and accessed. It is important to catalog in a standard format so that the data/software resource is retrievable by citation, author, data category, and Persistent Unique Identifier. Critical to the digital repository is having an online-backup (hot) that concurrently saves the data/software and is protected from ransomware. An effective data management strategy at NIH must include replicating digital repositories in an effort to protect the data/software so that NIH cannot be held hostage over the valuable data/software. Improper categorization and assignment of metadata could potentially produce an overabundance of data where relevant data is mixed with not so relevant data frustrating the user community. Once data is stored, methods to search and retrieve should be widely publicized to the user community, to provide a consistent method of search, allowing only relevant results to be displayed. This improves ease of data access and minimizes the overabundance of data that is displayed in response to searches.

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

The additional routes that NIH could strengthen and incentivize data and software sharing are through the National Science Foundation and Research Universities across the United States. NIH can join with these organizations in maintaining a distributed data repository that can be accessed from any location.

4. Any other relevant issues respondents recognize as important for NIH to consider

We believe that a Continuity of Operations Plan (COOP), complete with a back-up and restore schedule, is critical for the digital repositories focused on a data management strategy at NIH. Distributed data repositories that are replicated on a regular basis provide a high availability of access while providing disaster recovery in real time. Additionally, the COOP should include protection from ransomware so that NIH and adjoining institutions are not held hostage by rogue actors. Ransomware protection should consider and include: • Ransomware continues to proliferate and evolve because of the success of attacks securing ransom payments from high-value targets – expect more in 2017 • The growth of new sophisticated malware variants and zero-day exploits will continue and fortifying cyber defenses is critical to preventing breaches • Traditional antivirus is ineffective at stopping ransomware because polymorphic variants and obfuscation techniques easily bypass signature-based defenses • Next-generation antivirus is now a requirement to eliminate ransomware from your environment

Additional Comments

KIJV - RFI Response - NIH Data Management .pdf (199 KB)

Data Management, Sharing, and Citation

Request for Information Response

Department of Health and Human Services

National Institutes of Health

Notice Number: NOT-OD-17-015

19 January 2017

Submitted to:	Submitted by:
Department of Health and Human Services National Institutes of Health Office of Science Policy Division of Scientific Data Sharing Policy Telephone: (301) 496-9839 Email: SciencePolicy@mail.nih.gov	Mr. Earl D. Bowers Ishpi Information Technologies, Inc. 7240 Parkway Drive, Suite 200, Hanover, MD 21076 Phone: (757) 646-8568 Fax: (888) 388-5152 Email: Earl.Bowers@ishpi.net Cage Code: 4HUR7 DUNS: 620244264 TIN: 20-5353838

This Request for Information Response 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 or quotation. If, however, a contract is awarded to this offeror or quoter 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. This restriction applies to all pages of this proposal. Source Selection Information – See FAR 2.101 and FAR 3.104.

Capability Statement

To support in this NIH effort, Ishpi Information Technologies, Inc. (DBA *ISHPI*) has chosen to team with our trusted partner Konark Software Solutions, LLC (DBA Konark Software). Our two organizations have entered a Small Business Administration (SBA) approved 8(a) Mentor-Protégé relationship where *ISHPI* (Mentor) and Konark Software (Protégé) have the unique ability to provide services in the standard prime contractor/subcontractor relationship or provide services under our Joint-Venture (JV) arrangement as Konark-Ishpi JV LLC (DBA KIJV) – DUNS: 080412023. Through our unique Mentor-Protégé relationship, KIJV inherits the socioeconomic status and size standard of Konark Software, and all the capabilities of both organizations (including corporate experience, past performance, and CMMI capabilities). KIJV is operated using the management methodologies of *ISHPI*, as the Mentor organization.

Konark Software provides technology consulting services to state, local, federal, and commercial entities. Konark Software is an SBA certified 8(a) minority owned small business with offices in Virginia Beach, VA and North Charleston, SC. Konark Software specializes in business process reengineering; technology consulting; email and web security; program and project management; disaster recovery and business continuity; security information and event management (SIEM); cloud computing solutions; software design and application development; document imaging and workflow strategy; knowledge management; and network and data security services. Konark Software can provide expertise to NIH Data Strategy efforts in defining metadata, storage mechanisms, retention criteria, keyword searches etc. As a full lifecycle IT operation support organization, Konark is focused on enhancing service levels and optimizing resources to provide quality support services at affordable pricing. Konark recognizes the importance of security and the impact it has on day to day customer operations. To enhance security, Konark automates IT Asset Discovery and desktops/servers patch management tasks. In addition, Konark performs event log analysis to identify operational and security issues.

ISHPI is a Capability Maturity Model Integration (CMMI) Maturity Level (ML) 5 and six-time Inc. 500 | 5000 and Washington Technology Fast 50 award winning Service Disabled Veteran Owned (SDVO) Business specializing in Program Management; Cybersecurity; Software Development and Maintenance; and Information Technology (IT) services for the Federal Government. *ISHPI* is now one of only two small businesses in the United States with an organization appraised at CMMI ML 5, making *ISHPI* among the Top 5.6% of CMMI-appraised organizations in the U.S. and the Top 7.4% CMMI-appraised organizations in the world, according to the CMMI Institute's Published Appraisal Results. As such, *ISHPI* is able to offer NIH a proven, high-performance Small Business organization that uses statistical methods and quantitative data to understand risk, manage processes, and predict outcomes based on past performance.

ISHPI provides powerful software and IT solutions, and has managed to thrive in a highly competitive and rapidly changing business environment because our solutions are efficient, effective, and uses the latest technology. *ISHPI* consists of professionals and experts who are exceptionally qualified to fulfill all contract tasks and deliverables. As a premier provider of highly specialized IT and Cyber services, *ISHPI*'s corporate experience includes a full range of lifecycle services supporting the types of systems, applications and technologies that integrate and combine to deliver actionable information to our customers. *ISHPI* is a customer-oriented organization that provides a variety of critical state-of-the-art information and technology solutions. Realizing the

Use or disclosure of the information contained on this sheet is subject to the restrictions on the title page of this proposal.
Source Selection Information - See FAR 2.101 and FAR 3.104

dynamic nature of the technological age has equipped *ISHPI* with the right tools, expertise, methodologies and teaming arrangements that provide the flexibility to respond to trends and unique customer requirements. *ISHPI*'s workforce of more than 406 highly trained professionals and subject matter experts (SMEs) currently provide support in Program Management; Strategic Planning; Healthcare IT support; Software Development, Cybersecurity; Helpdesk Support; Systems Life Cycle Design Engineering; Earned Value Management (EVM); Planning and Requirements Documentation Development; and Logistics Engineering.

Moreover, over the past eight years, *ISHPI* has won, managed, and successfully executed more than $196 million worth of Government contracting work, including 95 Government contracts, 63 competitive task orders, and 10 competitive Indefinite Delivery, Indefinite Quantity (IDIQ)/Government Wide Acquisition Contract (GWAC) vehicles. *ISHPI* currently manages and executes 70 active independent contracts with ceiling value of more than $445 million. During this time, *ISHPI* has grown exponentially, as illustrated in **Table 1.** *ISHPI's* phenomenal growth in both revenue and employees attests to the capability of KIJV to manage and successfully perform work for NIH.

Table 1: *ISHPI*'s Growth

	2008	2009	2010	2011	2012	2013	2014	2015	2016
Revenue	$1.9M	$5.8M	$6.2M	$7.3M	$17.2M	$22.8M	$37.1M	$50.3M	$60.9M
Workforce	18	42	62	68	109	174	234	423	405
Growth %	N/A	205%	7%	14%	135%	32.5%	62.7%	345.6%	23%

ISHPI is also an International Information Systems Security Certifications Consortium (ISC)²® Official Training Provider, one of only 17 companies in the United States with a credential backed by (ISC)²®, the globally recognized Gold Standard in information security certifications. *ISHPI* is part of a select global network of authorized training organizations committed to delivering the highest standard in cyber security training and is authorized to provide training seminars, courses, and certifications including Certified Information Systems SecuriAty Professional (CISSP®), Systems Security Certified Practitioner (SSCP®), Certified Secure Software Lifecycle Professional (CSSLP®), Certified Cyber Forensics Professional (CCFP®), and Certified Authorization Professional (CAP®).

There are multiple compelling reasons for the NIH to select our Team, one of which is our highly-experienced staffing profiles currently delivering critical program management, IT operations support, and software engineering services like those tasks outlined in this RFI notice. Because we bring the experience to support the data management needs of NIH, we can deliver an effective and efficient solution with a minimal learning curve and offer lower risk, accelerated implementation, lower cost and increased value. The extensive knowledge and experience with providing these services and using similar technology, products and IT standards outlined by NIH, are crucial to the future success of these initiatives because of their complexity. The unique relationship between *ISHPI* and Konark Software enables the team to adapt and refine our approach in accordance with the approved NIH acquisition strategy.

Use or disclosure of the information contained on this sheet is subject to the restrictions on the title page of this proposal.
Source Selection Information - See FAR 2.101 and FAR 3.104

Section I. Data Sharing Strategy Development

Submitter Name: James A. Bryant, Ph.D.
Name of Organization: Ishpi Information Technologies, Inc.
Type of Organization: Service Disabled Veteran Owned Business, Veteran Owned Business, Minority Owned Business, Native American Owned
Role: Associate VP | Senior Health Care Solutions Architect
Domain of Research: Healthcare Information Technology

Type of Data: Structured, Semi-Structured and Unstructured; String, Integer, Float (floating point), Character and Boolean; Primary and Secondary Data.

Repositories You or Your Organization Primarily Utilize. National Institutes of Health (NIH) Office of Extramural Research (OER), Agency for Healthcare Research and Quality (AHRQ), Centers for Disease Control and Prevention (CDC), Food & Drug Administration (FDA), Substance Abuse and Mental Health Services Administration (SAMHSA) and Veterans Administration (VA).

1. The highest-priority types of data to be shared and value in sharing such data

The data type defines which operations can safely be performed to create, transform and use the variable in another computation. When a programming language requires a variable to only be used in ways that respect its data type, that language is said to be strongly typed. This prevents errors because while it is logical to ask the computer to multiply a float by an integer (1.5 x 5), it is illogical to ask the computer to multiply a float by a string (1.5 x Alice). When a programming language allows a variable of one data type to be used as if it were a value of another data type, the language is said to be weakly typed.

Data Type	Used for	Example
String	Alphanumeric characters	Bob123
Integer	Whole numbers	7, 12
Float (floating point)	Number with a decimal point	3.15, 9.06
Character	Encoding text numerically	97
Boolean	Representing logical values	TRUE, FALSE

Technically, the concept of a strongly typed or weakly typed programming language is a fallacy. In every programming language, all variable values have a static type, but the type might be one whose values are classified into one or more classes. And while some classes specify how the data type's value will be compiled or interpreted, there are other classes whose values are not marked with their class until run-time. The extent to which a programming language discourages or prevents type error is known as type safety.

Use or disclosure of the information contained on this sheet is subject to the restrictions on the title page of this proposal.
Source Selection Information - See FAR 2.101 and FAR 3.104

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

Retention Period: A principal investigator (PI) is the holder of an independent grant administered by a university and the lead researcher for the grant project. PIs are responsible for determining the length of time data should be made available; the appropriate means for maintaining and sustaining the data; and analyzing the long-term resource implications of storage.

Retention Periods are dictated by data relevancy, as opposed to storage requirements. The amount of time data should be stored is determined by information type. Occasionally, retention periods are at the discretion of the PIs; however, many sponsor institutions require data be retained for a minimum number of years. For instance, HHS requires project data be retained for at least 3 years after funding ends.

Continued Storage: Once the minimum retention period is met, the PI decides whether to continue with data storage. PIs evaluate the benefits and risks of extended storage. PIs never know when data might be needed, and continued storage of confidential data increases the risk of possible violation, not to mention the monetary cost associated with storage.

Destroying Data: When the decision has been made to end data storage, data should be thoroughly destroyed. Effective data destruction ensures that information cannot be extracted or reconstructed. Many document storage companies now offer onsite shredding and secure destruction of written and electronic records. For electronic data, software products, such as Eraser or CyberScrub, are available. All pointers related to the purged data should also be removed to avoid retrieval errors.

3. Barriers to data stewardship and sharing, and mechanisms to overcome these barriers

Data stewardship is a collection of data management methods covering acquisition, storage, aggregation, deidentification, and procedures for data release and use. A data steward is intended to convey a fiduciary (or trust) relationship with data that requires a loyal data manager vested to the interests of the entities whose data is stored. Barriers to data stewardship for NIH may include privacy concerns and how to securely gain access to personal identifiable data. This includes the information essential to understanding healthcare quality, safety, efficiency, and outcomes.

Data ownership, versus data access, require two distinct resolution approaches. The first, consistent with the concept of health information ownership, would be to incentivize data access through payment for information. A data stewardship entity purchases the required health information considered to be essential to patient safety, quality, comparative effectiveness, and population health. With data monetized, the government could negotiate over the scope and terms of access and use. The strength of this model is the recognition of data ownership rights. The chief limitations could be the overall cost and the uncertainties that surround market negotiations.

An alternative approach is to treat data as a public good. Stewardship entities could be federally chartered, with broad authority to collect, prepare, and support the use of health information in research. This model would achieve the broad goals set by advocates of evidence-driven care. In many respects, his model has taken precedence in health reform, although it is still subject to limitations on usage.

Use or disclosure of the information contained on this sheet is subject to the restrictions on the title page of this proposal.
Source Selection Information - See FAR 2.101 and FAR 3.104

4. Any other relevant issues

Another possible mechanism to overcome these barriers would be to designate certain data use as being in the public interest and to designate certain data as falling within a required data submission category, as a condition of participation in federal health programs. In this context, the term "federal health programs" might also include Social Security Act programs (e.g., Medicare and Medicaid) and Public Health Service Act health programs (public health; health resources development and health professions; health research; and other direct population health investments). Congress' Article I Constitutional powers are sufficiently broad enough that the reach of such a data submission requirement could encompass not only patient-level data emanating from provision of care directly under federal programs but also include data resulting from the provision of care to all payers. Health data governed by submission requirements could be aggregated, managed, and prepared for use by data stewardship entities, which in turn could freely license the data for use by researchers who demonstrate compliance with data stewardship responsibilities. This approach leaves data at the provider level available for subsequent uses, while also allowing for a flow of relevant scientific data into stewardship entities capable of supporting the type of research enterprise viewed as essential to healthcare system reform.

Section II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports (RPPR) and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

Grant recipients use the RPPR, a federally mandated format for required interim or annual reporting, to complete progress reports on NIH grant awards. As a grant recipient, progress reports are required annually to document accomplishments and compliance with the terms of the award. Recipients must describe their scientific progress, identify significant changes, report on personnel, and describe plans for the subsequent budget period or year in these annual reports.

NIH requires recipients to submit an RPPR for non-competing NIH awards. Funding for non-competing years of the grant can only be awarded after the NIH program and grants management staff review and approve the progress report. The review of the RPPR, by NIH staff, is a key element in NIH's monitoring of the grant award.

The impact of increasing reporting of data and software sharing in RPPRs, competing grant applications to enrich reporting of productivity of research projects, and incentivizing data sharing will improve effectiveness. Although the primary incentive to input the data seems to be for continued funding, the downstream benefit of data availability to thousands of users provides a significant benefit for continued research. The veracity of data and software sharing in RPPRs will improve and provide current data that enhances research performance.

2. Important features of technical guidance for data and software citation in NIH reports

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

Technical guidance for data and software citations in NIH reports will be greatly enhanced by persistent unique Identifiers, such as the Digital Object Identifier (DOI). A DOI is a type of persistent identifier used to uniquely identify objects. The DOI system is particularly used for electronic documents, such as journal articles.

Use or disclosure of the information contained on this sheet is subject to the restrictions on the title page of this proposal.
Source Selection Information - See FAR 2.101 and FAR 3.104

DOI means "digital identifier of an object" rather than "identifier of a digital object". Thus, the term DOI stands for "digital object-identifier" rather than "digital-object identifier".

Metadata about the object is stored in association with the DOI name. It may include, the object title or a location such as a URL indicating where the object can be found. The DOI for a document remains fixed over the lifetime of the document, whereas its location and other metadata may change. Referring to an online document by its DOI provides more stable linking than simply using its URL because if the URL changes, the publisher only needs to update the metadata for the DOI to link to the new URL. A DOI name differs from standard identifier registries such as the ISBN and ISRC. The purpose of an identifier registry is to manage a given collection of identifiers, whereas the primary purpose of the DOI system is to make a collection of identifiers actionable and interoperable.

The DOI is commonly used in academia to universally identify articles, journals, and magazines. The use of a Persistent Unique Identifier that resolves resources would be a huge benefit for NIH.

b. Inclusion of a link to the data/software resource with the citation in the report

The inclusion of a link to the data/software resource with the citation is critical. The Journal of Statistical Software states software providers may offer a recommended or required, citation format for any software user. This ensures the provider's research contribution is acknowledged. For example, the R open source statistical programming language and environment provide a BibTeX entry in their FAQ that can be used if citing R:

> @Manual{,
> title = {R: A Language and Environment for Statistical Computing},
> author = {{R Development Core Team}},
> organization = {R Foundation for Statistical Computing},
> address = {Vienna, Austria},
> year = 2011,
> url = {http://www.R-project.org}

Some providers make the citation part of the license. For example, SAS's mandated citation:

> The [output/code/data analysis] for this paper was generated using [SAS/STAT] software, Version [9.1] of the SAS System for [Unix]. Copyright © [year of copyright] SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA.

HSL library requires the following citation to be included when describing results achieved:

> HSL(2011). A collection of Fortran codes for large scale scientific computation. http://www.hsl.rl.ac.uk

The P-Genome bioinformatics software provides a citation for their software and publication:

> Sarkar IN, Planet PJ, Thornton J, DeSalle R, Figurski DH. Phylogenomenclature.

Use or disclosure of the information contained on this sheet is subject to the restrictions on the title page of this proposal.
Source Selection Information - See FAR 2.101 and FAR 3.104

Available at: http://www.dbmi.columbia.edu/~ins7001/research/CAOS/P-Gnome/PGdownload

Sarkar IN, Thornton J, Planet PJ, Figurski DH, Schierwater B, DeSalle R. An Automated Phylogenetic Key for Classifying Homeoboxes. Molecular Phylogenetics and Evolution 2002 Sep; 24(3):388

c. Identification of the authors of the Data/Software products

Identification of the authors of the Data/Software products is equally important. Describing the contribution of data/software to research closely relates to several other issues around the role of data/software in research. This includes publishing research data/software in a persistent and citable way; ensuring the availability of research software (and data, online services, and other artifacts) for the long-term; promoting the recognition of software as a valuable research output; and ensuring that research software developers have their contributions recognized and rewarded.

These are concerns that affect researchers using the software, as well as those who develop or modify research software; those who release research software; paper reviewers; program committees; publishers; and funders. All these agencies have a part to play and the Software Sustainability Institute is working with many different experts and groups to explore and resolve these issues. In addition, it allows data consumers to look for additional related data/software from the same author/developer, which could provide a macro perspective of the research topic and be relevant to their data usage.

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

The fact that some software consists of discrete well-defined components imposes additional requirements when describing software. As we saw with HSL, not only is a researcher's use of HSL significant, the names of the sub-routines within HSL are also required, to deliver an accurate description of the research undertaken.

This is analogous to the situation faced by data publishers and consumers, as part of their research. In response to these, The Digital Curation Centre (DCC) have produced a guide on data citation and linking that discusses these problems and provides advice:

Ball, A. & Duke, M. (2011). "How to Cite Datasets and Link to Publications?". DCC How-to Guides. Edinburgh: DCC. Available online: http://www.dcc.ac.uk/resources/how-guides/cite-datasets

As the authors describe, data can be hierarchically structured into various elements including in databases, tables, columns, rows, directories, files, records, data points, collections, and documents. Each of these elements may or may not have a specific identifier associated with them, depending upon how the data producer has published them. The DCC guide authors recommend that authors should cite data sources at the finest level of granularity that was adopted when an identifier, or for our purposes, a citation, was assigned. This can be supplemented in the text with the information to find any specific subset of the data that was used in the research. As researchers using HSL show, this advice can be readily adopted to describe the use of research software.

Use or disclosure of the information contained on this sheet is subject to the restrictions on the title page of this proposal.
Source Selection Information - See FAR 2.101 and FAR 3.104

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

The data/software should be unambiguously identified and cited in the digital repository where the resource is stored and can be found and accessed. It is important to catalog in a standard format so that the data/software resource is retrievable by citation, author, data category, and Persistent Unique Identifier. Critical to the digital repository is having an online-backup (hot) that concurrently saves the data/software and is protected from ransomware. An effective data management strategy at NIH must include replicating digital repositories in an effort to protect the data/software so that NIH cannot be held hostage over the valuable data/software. Improper categorization and assignment of metadata could potentially produce an overabundance of data where relevant data is mixed with not so relevant data frustrating the user community. Once data is stored, methods to search and retrieve should be widely publicized to the user community, to provide a consistent method of search, allowing only relevant results to be displayed. This improves ease of data access and minimizes the overabundance of data that is displayed in response to searches.

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

The additional routes that NIH could strengthen and incentivize data and software sharing are through the National Science Foundation and Research Universities across the United States. NIH can join with these organizations in maintaining a distributed data repository that can be accessed from any location.

4. Any other relevant issues respondents recognize as important for NIH to consider

We believe that a Continuity of Operations Plan (COOP), complete with a back-up and restore schedule, is critical for the digital repositories focused on a data management strategy at NIH. Distributed data repositories that are replicated on a regular basis provide a high availability of access while providing disaster recovery in real time. Additionally, the COOP should include protection from ransomware so that NIH and adjoining institutions are not held hostage by rogue actors. Ransomware protection should consider and include:

- Ransomware continues to proliferate and evolve because of the success of attacks securing ransom payments from high-value targets – expect more in 2017
- The growth of new sophisticated malware variants and zero-day exploits will continue and fortifying cyber defenses is critical to preventing breaches
- Traditional antivirus is ineffective at stopping ransomware because polymorphic variants and obfuscation techniques easily bypass signature-based defenses
- Next-generation antivirus is now a requirement to eliminate ransomware from your environment

Use or disclosure of the information contained on this sheet is subject to the restrictions on the title page of this proposal.
Source Selection Information - See FAR 2.101 and FAR 3.104

Submission Date

01/19/2017

Submitter Name

James D. Luther

Name of Organization

Duke Univeristy

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

Biomedical research

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

See attached memo that is focused on the potential for increased faculty burden and further forced cost-sharing by universities.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

4. Any other relevant issues respondents recognize as important for NIH to consider

See attached memo that is focused on the potential for increased faculty burden and further forced cost-sharing by universities.

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

b. Inclusion of a link to the data/software resource with the citation in the report

c. Identification of the authors of the Data/Software products

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

4. Any other relevant issues respondents recognize as important for NIH to consider

See attached memo that is focused on the potential for increased faculty burden and further forced cost-sharing by universities.

Additional Comments

Duke Response to RFI NIH-Open Access 1-19-17 FINAL.pdf (143 KB)

Duke University

DURHAM
NORTH CAROLINA
27708

FINANCIAL SERVICES
COST AND REIMBURSEMENT ACCOUNTING

BOX 104137
TELEPHONE (919) 684-5723
FACSIMILE (919) 684-8547

January 19, 2017

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

Subject: NIH Request for Information: Strategies for NIH Data Management, Sharing, and Citation

Comments in response to notice number NOT-OD-17-015

On behalf of Duke University, I am pleased to provide comments on NIH's "*Request for Information: Strategies for NIH Data Management, Sharing, and Citation".* 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 inherent in OSTP's and NIH's plan for data stewardship and long-term storage. Duke appreciates this opportunity to offer recommendations on how these issues might be resolved.

We would like to thank you for your continued interest and willingness to accept input from universities and the associations and organizations that represent us. In particular, Duke University would like to take this opportunity to endorse and support the AAU/APLU/COGR submission dated January 19, 2017. We would also like to affirm the comments submitted by members of the Duke University Libraries dated December 21, 2016. The intent of this submission is to provide additional context for items of critical importance.

We support the overall approach to data management, sharing, and citation and the plan developed by NIH to increase public access to the results of scientifically generated data by NIH-sponsored projects. We recognize the real potential behind these practices, and as the comments by the Duke University Libraries demonstrate, Duke already commits significant financial and human resources to data stewardship and administrative costs incurred therein. If NIH implemented the proposed plan without modification and clarification, the increase in these costs and the administrative burden, particularly for faculty, could be significant.

We recognize the challenge of developing requirements that achieve the stated goal of data that is publicly accessible and searchable, while attempting to make them as least burdensome as possible. As referenced in the AAU/APLU/COGR memo, a recent study by Royal Society estimated the requirements of data sharing could demand resources on the order of 1-10% of a funded project. For faculty members and PIs, the effort required to build, maintain and coordinate the storage of metadata generated through their projects amounts to a significant amount of hours, added administrative burden and increased costs for the University in the form of increased demand for administrative support and information technology resources and infrastructure.

We therefore feel it is of critical importance that the regulations provide universities the ability to recover costs associated with this new and expanding regulatory requirement. Further, requiring that agencies implement in a consistent and harmonized fashion will provide faculty and institutions a more effective path with improved economies of scale to support the more effective and efficient implementation of these regulations.

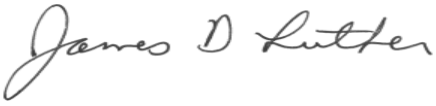
The concept of reducing burden and harmonizing across agencies is in direct alignment with Executive Order 13563 of January 18, 2011 which called for greater coordination across agencies to reduce costs and simplify and harmonize rules, and for agencies to identify and consider regulatory approaches that reduce burdens and maintain flexibility. This has further been a prominent aspect of both the National Academies report "Optimizing the Nation's Investment in Academic Research: A New Regulatory Framework for the 21st Century" and the GAO report entitled "Opportunities Remain for Agencies to Streamline Administrative Requirements" (GAO-16-573: Published: Jun 22, 2016).

The cost of this new and potentially expensive requirement should be shared with secondary users and/or federal sponsors either as a direct charge to the sponsor or via recoverable (e.g. non-capped) indirect costs. Institutions already experience capped cost recoveries and cost-sharing which results in universities contributing billions of dollars to support the research mission. Here at Duke, this contribution approximates $150 million. New regulatory requirements like this could contribute and further dilute the institution's ability to support the research mission and the faculty members.

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 most 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,



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. Raphael Valdivia, Vice Dean for Research
Tim Walsh, Vice President Finance

Submission Date

01/19/2017

Submitter Name

John Michael DeCarlo

Name of Organization

IBM

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

Major Chronic Diseases

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

4. Any other relevant issues respondents recognize as important for NIH to consider

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

b. Inclusion of a link to the data/software resource with the citation in the report

c. Identification of the authors of the Data/Software products

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

4. Any other relevant issues respondents recognize as important for NIH to consider

Additional Comments

IBM Final Comment NIH Data Sharing 1_19_17_0.pdf (243 KB)

January 19, 2017

NIH - Office of Science Policy
Division of Scientific Data Sharing Policy
Rockledge 1, Suite 750
6705 Rockledge Drive
Bethesda, MD 20817

RE: NOT-OD-17-015, NIH Request for Information (RFI): Strategies for NIH Data Management, Sharing, and Citation

Dear Office of Science Policy:

IBM appreciates the opportunity to comment on NOT-OD-17-015; *Strategies for NIH Data Management, Sharing, and Citation*, released November 14, 2016. We support the agency's ongoing efforts to ensure peer-reviewed publications and digital scientific data resulting from federally-funded scientific research are accessible to the public, industry, and the scientific community to the extent feasible and consistent with applicable laws and policies.

IBM offers two key considerations as NIH develops its policies pertaining to the sharing of digital scientific data generated from NIH-supported research: (1) Prioritize and ensure genomic data sharing to address major diseases, such as cancer; and, (2) Address long-standing barriers to sharing of health data by consideration of blockchain technology.

Prioritize genomic data for sharing based on the value of such sharing

In 2017, there will be an estimated 1,688,780 new cancer cases diagnosed and 600,920 cancer deaths alone in the U.S.[1] As the fight against cancer continues, both researchers and oncologists are becoming increasingly aware that speeding up the quest for cures hinges, in part, on the ability to access, analyze, and draw insights from genomic data.

Following the initial mapping of the Human Genome in the early 1990s, NIH funded researchers across the U.S. continue to make groundbreaking discoveries that advance our molecular understanding of cancer and lead to the development of targeted therapies that can ultimately save lives. Indeed, researchers and clinicians alike are placing increasing value on the importance of genomic data to better understand and treat disease.

At IBM, we're also committed to the fight against cancer. We are applying the cognitive computing power of Watson to enable clinicians to quickly translate DNA insights derived from genomic data into personalized treatment options for patients. Known as "Watson for

[1] American Cancer Society, *Cancer Facts & Figures 2017*.

Genomics," IBM is collaborating across 16 leading cancer institutes, such as Duke, Yale, University of North Carolina among others, to help doctors identify cancer- causing mutations and mapping those mutations to evidence-based therapeutic options. As part of this process, Watson is reducing the time associated with DNA analysis from weeks to minutes in some cases. For cancer patients, research has demonstrated that increased timeliness to diagnosis and treatment is associated with better outcomes.[2]

As current and future NIH funded research yields critical insights and scientific knowledge related to the genomic basis of the disease, we urge NIH to prioritize genomic data and ensure its ability to be shared. Indeed, we were pleased to see a renewed effort on behalf of NIH with the release of the Genomic Data Commons (GDC) in 2016, which aims to increase the availability and sharing of genomic data from research funded by the National Cancer Institute (NCI).

While genomic data is critically important for cancer research, it's equally as critical to the understanding of many other diseases and delivering more broadly on the promise of precision medicine. Type I and Type II Diabetes, neurological disorders (e.g. Alzheimer's and Parkinson's), and CHF and COPD are just a few of the serious chronic diseases in which genetic characterizations will be essential to full understanding of disease pathways.

Currently, there is very active research in disease onset and progression modelling, as well as treatment optimization. This work has the potential to contribute to earlier detection, primary and secondary prevention strategies, and personalized treatment regimens. In addition, even diseases with known genetic underpinnings, such as Huntington's, will benefit greatly from an expanded understanding of genomic and exomic data. As a result, future work on drug safety and drug repositioning will benefit enormously from such data and support a reduction in adverse drug reactions and improve our ability to identify new uses for existing medications.

We recommend that NIH continues to pursue additional efforts that ease and encourage the sharing of genomic data so that IBM and others can further accelerate precision medicine.

Consider using blockchain technologies to provide a highly secure, authenticated framework for data sharing

While IBM understands the importance of data sharing to accelerate precision medicine in disease areas such as cancer, as a company that acts as a trusted steward for both identified and de-identified data, we are very cognizant of the concerns surrounding potential breaches and mishandling of patient health data. That being said, IBM believes that large scale sharing of health data has been unduly limited in reaction to these concerns and the data exchange

[2] Neal, R. D., Tharmanathan, P., France, B., Din, N. U., Cotton, S., Fallon-Ferguson, J., … Emery, J. (2015). Is increased time to diagnosis and treatment in symptomatic cancer associated with poorer outcomes? Systematic review. *British Journal of Cancer*, *112*(Suppl 1), S92–S107. http://doi.org/10.1038/bjc.2015.48

process can be greatly enhanced through a more proactive approach of turning the focus to technological advancements and more secure methods of exchange.

We discourage NIH from enacting future policies that stymie innovation by placing unnecessary limits on data sharing. Rather, as a potential solution to overcome privacy and security challenges, IBM encourages NIH to explore emerging industry led approaches to facilitate secure exchange of patient level health data across the health care ecosystem.

For example, to enable the protected exchange of clinical, genomic, electronic medical records (EMR) and other types of health data, IBM is currently piloting new technology referred to as "blockchain." Unlike traditional data exchange, blockchain establishes accountability and transparency in the data exchange process by keeping an audit trail of all transactions on an unalterable distributed ledger.

In early 2017, IBM and the FDA partnered to explore how blockchain can provide benefits to public health by enabling the secure, efficient, and scalable data exchange of patient-level data from several sources, including EMRs, clinical trials, genomic data, and health data from mobile devices, wearables, and the "Internet of Things."[3] Although the initial focus will be on oncology-related information, ultimately, the collaboration will shed new light on ways to leverage the large volumes of diverse data in today's biomedical and healthcare industries to yield new biomedical discoveries by combining data across the healthcare ecosystem.

IBM encourages NIH to study this pilot and blockchain technology more broadly for its potential applicability to facilitate secure use and exchange of sensitive data generated by federally funded research.

Conclusion

IBM appreciates the opportunity to comment on the RFI. We welcome the opportunity to discuss our comments in further detail as the agency develops its data sharing strategy. If you have any have questions or concerns, please contact John Michael De Carlo, Executive, Watson Health Policy, IBM at john.michael.decarlo@ibm.com

Sincerely,

Roslyn Docktor,
Director, Watson Health Policy

[3] For more information regarding IBM's partnership with FDA on blockhain technology, please see: http://www-03.ibm.com/press/us/en/pressrelease/51394.wss

Submission Date

01/19/2017

Submitter Name

Letisha R. Wyatt, PhD

Name of Organization

Oregon Health & Science University

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

Basic and clinical biomedical research

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

Our group believes that the highest priorities for types of data to be shared are (i) experimental data that was used to generate figures for a published manuscript. This would include the raw data from data collection, information about how the data was transformed/analyzed, and the analyzed data set; (ii) negative data, pre-registration/preliminary data, data from large scale screens or -omics evaluations that isn't publishable in the context of a manuscript; and (iii) all supplemental material (including data and identifying information about the reagents and resources used to generate the underlying data) in manuscripts, as most journals do not promise long-term archiving of these components.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

This depends on the sustainability of the repository infrastructure (organizational, resources, and technology) and collection management policies. However, a retention schedule should exist for raw, archival, and analyzed data, along with other appropriate publication materials. We think that individual researchers should store the data as long as possible. One can't anticipate how data will change in the future, so perhaps one way of culling the "keep everything for forever" impossibility would be to make sure that datasets that are maintained past a 5- or 10-year period would be those that are most available to meaningful reuse. In either case, metadata should be available forever, so that if data is not available, at least it is possible to have some information about it. As well, considerations should be made about the longevity of storage with regard to the usefulness of the associated metadata and formats. And our final point relates to discoverability, whereby more precise metadata can help ensure that a resource has accurate ranking in search engines such as Google as well as more targeted applications such as BioCaddie's DataMed or Re3Data. After all, what good is the data if it cannot be found for reuse?

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

Data stewardship/sharing barriers: - Lack of man-power and training: human-power and expertise in digital assets management, digital archival management, and digital preservation will always be necessary and add to cost - Few considerations about effectively capturing metadata through the research lifecycle, which should involve administrative (important for data stewardship and this is where lab managers come in - not individuals actively engaged in the research), technical, and descriptive elements to support provenance - Inconsistent data sharing policies between Institutions, funding agencies, and publishers - Researcher confusion about sharing requirements and/or value - Researcher fear of being scooped - Limited available funds - this should be considered in grant application budgets - Researchers don't know of or are overwhelmed by all of their repository options - Minimal rewards for researchers who share their data - Limitations of repository subject and institutional functionalities and bad user interfaces - Licensing and data use restrictions are inconsistent at best and nonexistent at worst Ways to overcome these obstacles might include embargo options to help mitigate fear of being scooped; promoting tools that can integrate into workflows that

aid in data sharing (e.g., the Open Science Framework, Github, Figshare); sustainable and robust funding mechanisms to address costs associated with repositories, publishing in open access journals, educating researchers; taking some of the work out of the process by creating standards and universal funder/publisher/institutional policies; and more rewards for data sharing (i.e., metric for promotion and tenure).

4. Any other relevant issues respondents recognize as important for NIH to consider

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

b. Inclusion of a link to the data/software resource with the citation in the report

c. Identification of the authors of the Data/Software products

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

4. Any other relevant issues respondents recognize as important for NIH to consider

Additional Comments

NIHRFIondatamanagement_OHSU_OSU_submitted_2017-01-19.pdf (273 KB)

On behalf of individuals at Oregon Health & Science University and Oregon State University, we provide the following response to the **Request for Information: "Strategies for NIH Data Management, Sharing, and Citation".**

Contributors to this RFI are working directly with data via the integration of data at scale from numerous public data repositories, support for websites and services that provide access to integrated data, and/or assisting library patrons who have a diverse set of uses and guidance needs for digital information. Thus, we are familiar with the challenges related to data management, sharing, and citation. Collectively, our roles include biomedical science researcher, standards developer, data specialists, repository library/information scientist, and data curator. We have chosen to focus exclusively on data preservation and sharing (Section I only) as we feel that these are central to many of the issues related to successful research rigor and reproducibility which is a current focus of the NIH.

Contributors:

Letisha R. Wyatt, PhD - Basic Science Liaison/RDM Librarian, Oregon Health & Science University
Nicole Vasilevsky, PhD - Biocurator/Ontologist, Oregon Health & Science University
Kate Thornhill, MLIS - Repository Community Librarian, Oregon Health & Science University
Jackie Wirz, PhD - Research Data Specialist (ODG), Oregon Health & Science University; Assistant Dean, Graduate Studies, Oregon Health & Science University
Clara Llebot Lorente, PhD - Data Management Specialist, Oregon State University
Melissa Haendel, PhD - Director of the Ontology Development Group (ODG), Oregon Health & Science University
Julie McMurry - Software Program Manager, Oregon Health & Science University

Section I: Data Sharing Strategy Development

The NIH recognizes that many factors must be considered when determining what, when, and how data should be managed and shared. These factors include, for example, the purpose for sharing, supporting data reuse and reproducibility, maturity of the science, the infrastructure uniqueness of the data, and ethical considerations. The NIH seeks comment **on any or all** of the following topics to help formulate strategic approaches to prioritizing its data management and sharing activities:

1. The highest-priority types of data to be shared and value in sharing such data (Maximum: 250 words)

Our group believes that the highest priorities for types of data to be shared are (i) experimental data that was used to generate figures for a published manuscript. This would include the raw data from data collection, information about how the data was transformed/analyzed, and the analyzed data set; (ii) negative data, pre-registration/preliminary data, data from large scale screens or -omics evaluations that isn't publishable in the context of a manuscript; and (iii) all supplemental material (including data and identifying information about the reagents and resources used to generate the underlying data) in manuscripts, as most journals do not promise long term archiving of these components.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications (Maximum: 250 words)

This depends on the sustainability of the repository infrastructure (organizational, resources, and technology) and collection management policies. However, a retention schedule should exist for raw, archival, and analyzed data, along with other appropriate publication materials. We think that individual researchers should store the data as long as possible. One can't anticipate how data will change in the future, so perhaps one way of culling the "keep everything for forever" impossibility would be to make sure that datasets that are maintained past a 5- or 10-year period would be those that are most available to meaningful reuse. In either case, metadata should be available forever, so that if data is not available, at least it is possible to have some information about it. As well, considerations should be made about the longevity of storage with regard to the usefulness of the associated metadata and formats. And our final point relates to discoverability, whereby more precise metadata can help ensure that a resource has accurate ranking in search engines such as Google as well as more targeted applications such as BioCaddie's DataMed or Re3Data. After all, what good is the data if it cannot be found for reuse?

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers (Maximum: 250 words)

Our (non-exhaustive) list of barriers includes:

- Man-power and training: human-power and expertise in digital assets management, digital archival management, and digital preservation will always be necessary and add to cost
- Few considerations about effectively capturing metadata through the research lifecycle, which should involve administrative (important for data stewardship and this is where lab managers come in - not individuals actively engaged in the research), technical, and descriptive elements to support provenance.
- Inconsistent data sharing policies between Institutions, funding agencies, and publishers
- Researcher confusion about sharing requirements and/or value
- Researcher fear of being scooped
- Limited available funds - this should be considered in grant application budgets
- Researchers don't know of or are overwhelmed by all of their repository options
- Minimal rewards for researchers who share their data
- Limitations of repository subject and institutional functionalities and bad user interfaces
- Licensing and data use restrictions are inconsistent at best and nonexistent at worst

Ways to overcome these obstacles might include providing an embargo option to help mitigate fear of being scooped; promoting tools that can integrate into workflows that aid in data sharing (e.g., the Open Science Framework, Github, Figshare); providing

sustainable and robust funding mechanisms to address costs associated with repositories, publishing in open access journals, educating researchers; taking some of the work out of the process by creating standards and universal funder/publisher/institutional policies; and creating more rewards for data sharing (i.e., metric for advancement in promotion and tenure).

4. Any other relevant issues respondents recognize as important for NIH to consider (Maximum words: 250)

Our group unanimously agrees that training is a high priority. There is an increased need for training and education for information and data management. What we are trying to achieve is a paradigm shift in how researchers really see and use data. While many Institutions are now requiring short trainings on data management, new trainees would benefit from more intensive and long-term training combined with practice, throughout their training period. This could be assisted by additional funding for sustainable efforts to educate and train all researchers and promote a culture that facilitates discovery, reuse, and sharing. We need to incentivize senior researchers to make this a priority in their work, we need to incentivize Institutions to support these efforts, and we need to motivate researchers to participate through genuine engagement rather than policy or requirement. Education of both early career and established researchers might include such topics as (but not limited to):

- Repository systems: when should you put data in an Institutional repository versus a domain specific one;
- Organizational structures: who should you decide to deposit your data with for archival management;
- Proper data standards used for deposition descriptions (which should be standardized for sharing across generic and Institutional repositories);
- Digital preservation; and
- Evaluation of trustworthiness of data reproducibility

Submission Date

01/19/2017

Submitter Name

Shaun Purcell

Name of Organization

Brigham & Women's Hospital

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

Genetic epidemiology, cognitive neuroscience of sleep

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

Priority should be given to data that is difficult/expensive to obtain, and ideally that which has been collected on individuals/models for which other datatypes already exist, or are likely to be collected in the future. Facilitating “cross-modal” analyses (perhaps not originally anticipated at the time of data collection) is a central goal. For example, augmenting a longitudinal, richly-phenotyped dataset with genetic or imaging data, etc; different omics on post-mortem brain samples, etc. Important to be able to link individuals across databases. One exemplar of data sharing: the National Sleep Research Resource (NSRR, sleepdata.org), which collates, curates and distributes sleep signal, medical and demographic data on thousands of individuals. This resource (which has facilitated my own work greatly) has enabled analyses (e.g. of brain activity during sleep) which were presumably not envisioned as goals when many of the original studies were performed for very different purposes (e.g. focusing on sleep apnea and cardiovascular outcomes). This type of 'repurposing' scenario seems like the real 'win' of data archiving. Individuals in domain-specific databases such as the NSRR should be explicitly linked to other relevant databases (e.g. genetics), and a PubMed-like tool would exist that can query on study (who and what is in this study?), phenotype (what studies have this phenotype?) and/or individual (what other phenotypes/databases are available for this individual).

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

Retain all key data as long as possible (contingent upon use, and/or whether other commonly-used databases contain information on that individual/assay/phenotype, etc i.e. to predict whether a given resource is likely to be used/useful in the future). Realistically, the model of projects being driven by 3-4 year grants means that it will not be sustainable for the original investigators to maintain and distribute data indefinitely after that period. Even if typical R01s include a plan and funds for maintaining datasets as an investigator-centric activity, people move, files get lost, pipeline dependencies break down, etc. Two complementary routes: 1) that NIH centrally hosts storage for projects, even if this necessitates some pre-specified portion of the grant budget to be put aside for this purpose; 2) more funding (e.g. via R24 or similar mechanisms) for community efforts to develop appropriate (but still centrally indexed) databases, similar to NSRR. Naturally, having a modest-sized network of inter-connected, individually-rich resources is preferable to individual investigators simply throwing files up on the web. This does still place a burden on investigators to ensure that data are distributed in an appropriately well-documented format: but knowledge of this will mean that applicants should formulate data sharing plans that explicitly budget for these activities.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

As above, stewardship is realistically outside of the hands of individual investigators. From the perspective of the investigator obtaining the data, the process of sharing can be opaque, tedious, or restrictive. As more resources become available, one wants to avoid potential users being tied up in endless application processes. While still maintaining

security and privacy of data as important goals, perhaps a two-tier system of expedited access for sufficiently qualified investigators is warranted, cf. Global Entry / TSA precheck. If that approach works for national security, presumably parallels for data access in research may be effective too.

4. Any other relevant issues respondents recognize as important for NIH to consider

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

b. Inclusion of a link to the data/software resource with the citation in the report

c. Identification of the authors of the Data/Software products

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

A possibility is that every funded grant has an associated “Resources/deliverables Page” (NIH-hosted), on which the investigator is able to document the direct deliverables from that work. As well as a links to citations, this could include links to datasets, software, tutorials or presentations, etc. To make achievable for all investigators and easy to maintain, one could adopt a wiki-like interface so it is easy to upload/generate material without centralized oversight and maintainence. Could include a brief list of key specific aims and key results. Limited to two pages or so to make it manageable, and it gets “locked down” after some period. A complement of the "Specific Aims" page, if you like. One could imagine that such “project-centric reporting” could become a useful part of peer-review, etc, downstream, in the context of grant reviews, i.e. to complement individual-centric (biosketch) data and manuscripts. It would be important to avoid this becoming solely a reporting burden: ideally it would be optional, simple/relatively free-form, but it could be a useful platform for investigators to showcase all deliverables from a given project. The supposition is that, given sufficient visibility, peer-feedback drives a virtuous circle of reporting/sharing via this mechanism.

4. Any other relevant issues respondents recognize as important for NIH to consider

Additional Comments

Submission Date

01/19/2017

Submitter Name

John Tagler & Michael Mabe

Name of Organization

AAP Professional and Scholarly Publishing Division & International Association of STM Publishers

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

Our members publish journals and other scholarly material in all domains of research

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

The highest-priority types of data to be shared should be those based in or connected to publications, as these have the greatest potential to improve users' understanding of research findings and enable the research community and practitioners to further validate and replicate findings in published articles. These are also data that are already shared in many communities of practice, and therefore are the lowest hanging fruit for encouraging their use and broader dissemination. In looking beyond such data, it is critical to distinguish between data 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 respect intellectual property protection and copyright laws. We refer you to 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). The report, written by a coalition representing researchers, publishers, libraries and data centers, is a comprehensive look at research data and sharing. The need to expand incentives for providing broad and timely access to new data must be balanced with the need to preserve incentives for researchers to interpret and analyze their results through curation and peer-reviewed publication. NIH should also be mindful of precedents it may set regarding data management and preservation, including with respect to privacy. While the NIH's focus is on biomedical data, NIH's policies are likely to influence approaches taken in other research and funding communities.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

This is a complex issue with significant resource implications, which should be addressed by researchers themselves within their communities of practice. Publishers are already working with the communities we serve to develop standards for linking and sharing data where authors choose to do so, and to provide persistent links to external data collections. Federal policies should take into account the differences between information products at different levels of the pyramid mentioned in our response to question 1. Information products at the top of the pyramid, those connected to publications, should be persistently preserved in perpetuity for the integrity of the scholarly record, whereas data at lower levels of the pyramid (e.g. some raw data) might not need to be preserved for as long. It will be key for NIH to work with all stakeholders, including primary researchers, secondary researchers, publishers, libraries and data centers, to create clear rules and protocols for the management and sharing of data. A collaborative approach will ensure that the needs of each stakeholder group are addressed and that the progress of science is not impeded.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

Unlike publications which can be counted in pages, datasets can be measured in mega-, giga- or even terabytes. Unlike publications, which undergo peer-review and editing, extensive formatting and tagging, data files come in a variety of forms, formats and levels of quality. It is difficult to control for quality, display and consistency, and there are no

standards for which data should be preserved and how or where it should be managed. The compliance costs of sharing data are significant, especially when compared with current practices. In order to maximize its usefulness, data should be tagged, metadata added and the data must be reviewed to determine what can be shared and where. In addition, there are significant costs associated with storage, distribution bandwidth and overall management and curation. Initiatives must be carefully developed to support storage, dissemination, tagging, and validation. Success will depend on a collaborative approach that elicits buy-in from all communities and includes consultation and contributions by key stakeholders to develop robust, sustainable and flexible standards. NIH must carefully consider how best to create incentives for data management and sharing, and provide support for such activities. Publishers stand ready to lend their expertise to such a collaborative process to provide value to the research community and to the taxpayer. NIH should not invest resources to recreate what is already being achieved by the private sector, but should leverage public-private collaborations to ensure continued innovations that contribute to the progress of science and innovation and help grow the American economy.

4. Any other relevant issues respondents recognize as important for NIH to consider

Community-based policies: AAP/PSP and STM agree with the OSTP's Interagency Working Group on Digital Data that "data stewardship is best accomplished in a system that includes distributed collections and repositories maintained where the custodian has trusted community-proxy status with the relevant communities of practice." A critical component of any policy needs to be collaboration with researchers, publishers, librarians, universities and research institutions in an interconnected system based on community needs, standards and best practices. Each stakeholder community can contribute its expertise and ensure the creation of data management policies that reflect the different practices of individual research communities. Confidential data: Research communities have gone to great efforts to develop standards that ensure research subjects are treated ethically and that confidentiality of data is preserved. NIH must make sure that it does not undermine these protections as it works to expand access to data. Fraud: NIH will need to consider how it can work to detect data fraud. Tools can be deployed to analyze data that is "too perfect" from a statistical standpoint or to analyze images for manipulation, but next generation tools need to be developed to stay ahead of any efforts to mislead the public. Validating data: Capabilities should be developed for validating data in terms of both quality and utility, especially when considering some of the questions raised in the technical section with respect to long-term support for and relevance of the data.

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

Such increased reporting may create additional burdens on researchers. Publishers would be interested in taking part in discussions with funders and supported researchers to explore if there might be opportunities to reduce these burdens through ongoing projects or new initiatives and standards.

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

The Digital Object Identifier (DOI), a unique code for each piece of content in a scholarly publication, was developed and adopted through a multi-stakeholder, community-driven approach. It is successful because the standard evolved in response to a real problem in scholarly communication and is providing practical benefits to users of published articles about research. Digital data standards are newer and still evolving. Publishers have worked throughout the digital era to develop appropriate standards, persistent identifiers and protocols to enable seamless interlinking between publications through the development of the CrossRef organization and the use of standardized digital object identifiers (DOIs). CrossRef and DataCite have already been hard at work to extend these practices to data, but challenges remain. In order to minimize costs and maximize accessibility and usability of data, NIH should work with these existing initiatives and standards organizations like NISO to ensure the widespread adoption of both standardized DOIs and standard metadata protocols for data. Potential exemplars include DataCite (http://datacite.org/), APARSEN (http://www.alliancepermanentaccess.org/index.php/current-projects/aparsen/), Opportunities for Data Exchange

(ODE, www.ode-project.eu), CoData (http://www.codata.org/), NISO/NFAIS Supplemental Journal Article Materials Project (http://www.niso.org/workrooms/supplemental), PARSE.insight (http://www.parse-insight.eu/), and ORCID (http://orcid.org/).

b. Inclusion of a link to the data/software resource with the citation in the report

Any guidance as to the format and location of links, as well as acceptable repositories or locations to which a link might be directed, should be developed in consultation with stakeholders and consistent with the development of broadly-accepted community standards, as discussed above.

c. Identification of the authors of the Data/Software products

The scholarly community already has a robust attribution and credit system with respect to peer-reviewed publication, including disambiguation tools like ORCID and systems to identify various types of contributions to a work. Existing systems and tools could be leveraged in a bi-directional manner by linking between datasets and publications on the one hand, and exploring a requirement with key stakeholders that all data which informs the analysis and conclusions of a peer-reviewed publication be cited and attributed according to community standards on the other. The federal government's role could be to help by promoting those standards and provide clear rules for the citation of datasets and acknowledgement of modifications to source data. Such standards should also promote unique and persistent identifiers for data and disambiguate researcher, institution and funder information in metadata. Over the past decade, publishers developed the Digital Object Identifier (DOI), a unique code for each piece of content in a scholarly publication, and similar identifiers are being developed by DataCite for data (www.datacite.org). The work of DataCite, CrossRef, ORCID, and DOE's Data ID Service should be leveraged to ensure data is appropriately archived and recognized as a primary research output.

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

For the persistence and integrity of the scholarly record, it is important that digital repositories be identified alongside the deposit of the information resource cited. Automated updates to limit the burden on researchers and preserve the integrity of the information would be helpful. Unambiguous identification could support succession planning and location services should the repository cease operation or the resource be moved at a future time.

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

Many communities of practice are investigating the question of how to support researchers in sharing data and what kinds of incentives work. These studies should inform the development of NIH policies. Empirical evidence should be used to make evidence-based decisions about what efforts should move forward and how best to develop policy. Involving a broad array of stakeholders in policy development and implementation will ensure the preservation of incentives for innovation and help improve information sharing and training within each stakeholder community, as well as incentivize the sharing itself. Stakeholders can help develop clear standards and guidelines for the availability of research data, certification and auditing of data repositories and metadata standards, which respect each community's standards and practices, working together to create universal policies that work for all communities. Stakeholder input is also important for the integrity of the scholarly record, including the creation of links between datasets and the scholarly publications that analyze and interpret the data. Supporting such standards will improve researcher buy-in and compliance with requests for sharing. Developing guidance, in consultation with key stakeholders, to minimize the administrative burden on key stakeholders, would also improve compliance.

4. Any other relevant issues respondents recognize as important for NIH to consider

Community-based policies: AAP/PSP and STM agree with the OSTP's Interagency Working Group on Digital Data that "data stewardship is best accomplished in a system that includes distributed collections and repositories maintained where the custodian has trusted community-proxy status with the relevant communities of practice." A critical component of any policy needs to be collaboration with researchers, publishers, librarians, universities and research institutions in an interconnected system based on community needs, standards and best practices. Each stakeholder community can contribute its expertise and ensure the creation of data management policies that reflect the different practices of individual research communities. Confidential data: Research communities have gone to great efforts to develop standards that ensure research subjects are treated ethically and that confidentiality of data is preserved. NIH must make sure that it does not undermine these protections as it works to expand access to data. Fraud: NIH will need to consider how it can work to detect data fraud. Tools can be deployed to analyze data that is "too perfect" from a statistical standpoint or to analyze images for manipulation, but next generation tools need to be developed to stay ahead of any efforts to mislead the public. Validating data: Capabilities should be developed for validating data in terms of both quality and utility, especially when considering some of the questions raised in the technical section with respect to long-term support for and relevance of the data.

Additional Comments

Submission Date

01/19/2017

Submitter Name Alexander

Sherman **Name of**

Organization

Massachusetts General Hospital

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

Data Science, Data Sharing, Bioinformatics

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

- Define shareable data types/granularity: o Clinical trials: *De-facto standard CDISC data model shall be extended to NIH-sponsored trials, as recommendation with compulsory compliance timeline; o Biomarker/disease-specific natural history studies, including clinical trials readiness networks: Mostly disease-specific, certain common CDEs and outcomes shall be introduced across studies * NIH (NINDS) completed CDE projects for several diseases (without maintenance/enforcement those CDEs were not widely adopted) o Images - standards for Anonymization * Headers' information: study ID, subject ID, disease ID, timepoint, etc. * Raw versus annotated * Image acquisition platform Information o WGS Genetic files * Well-annotated metadata linking to other patients' datasources * Raw files, .bam files, etc. * Sequencing platform Information o Biospecimen * Nomenclature of biospecimens across NIH institutions * Tissue nomenclatures are disease-specific, requires reconciliation across disease areas for disease-controls * Label standardization (GUIDs and Study ID) * Some prominent examples: www.alsconsortium.org o Patient-reported outcomes o Omics files o Secondary data sources (data analyses files) - Develop requirement/use cases including: - Data Sources - Consent language for sharing (disease areas/data recipients) - Discrete/Continuous - Finalized/Updateable - Embargo/Expiration dates - Obfuscation - Audit trails/Provenance - Develop mechanisms, standards and services for linking and merging patient data originating from multiple efforts/sources - Use cases for data sharing may be determined by the data origins - Provide specific guidelines and practical data sharing examples - Inform IRBs on how to best handle requests for re-analyses

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

Storage length data-type dependent: - Clinical trials and biomarker/studies/registries: indefinitely (Pooled- resources Open-Access ALS Clinical Trials database (PRO-ACT) contains harmonized data from 23 clinical trials spanned over 20+ years (www.alsdatabase.org) - Images, GWS files: indefinitely - Biospecimens: determined by nomenclature type; digitized omics analyses results: indefinitely Embargo period should expire after data publication, or a year after grant's end. NIH shall: - Develop or utilize existing searchable inventory management system for individual datasets upload, maintenance, management, download and tracking; - maintain registry of validated and compliant with to-be-developed SOPs institutional/consortia repositories For either (centralized or federated) model consider: - Data contributors' obligations for data quality/completeness - NIH's obligations when hosting institutions seize to exist - Utilization and support of existing ETL tools and groups that develop/utilize them (NCRI@MGH) - Utilization of groups supporting disease-specific or consortia-driven repositories; require grant recipients contributing to these repositories. (www.NeuroBANK.org) - Requirements for linking clinical data to bio-, genetic-, image banks - Development of permissive, ontology-driven data representation strategy and dynamically expandable data modeling structure for

indexing and data interchange and retrieval though web services for following advantages: o Adoptable new data types; o Specialized data standards for specific consortia; o Ontology-driven approach for data mapping and management o Granular data indexing, from large “blocks”, to segmented, semantically specified for certain subpopulations and
variants; o Develop metadata layer’s elements for:

- Data

Audit trail

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

For both Centralized and Federated models cost of data sharing is determined by costs of individual components including: - Storage; - Bandwidth of data upload/download operations; - Data preparation, management, maintenance, transformation, anonymization, support of data dictionaries, ontologies and other descriptors; - Development of legal and regulatory framework and maintaining and tracking legal documents including DUAs
and MTAs for both data contributors and data requestors; - Consent language management and inventory; - Time to request execution - Software development/customization and maintenance. - Incentives for sharing - Reporting and tracking requirements - Burden for original data transformation to NIH-promoted formats and models - Burden for maintaining data in its original formats required by the projects and in the NIH-promoted one. Continuously flowing data: - Require more sophisticated micro-transactional management system for accepting the updating transactions, reconciliation, and failure recovery. - A “query on demand” approach through web services may be used.

4. Any other relevant issues respondents recognize as important for NIH to consider

Essential for NIH research awardees to systematically create and preserve research project data and data documentation to facilitate data sharing and/or reproduction. • Facilitate reproduction of original analyses to increase the integrity of NIH- funded research findings; and • Promote data sharing to enable conduct of additional analyses using data from NIH-funded studies, thereby augmenting the knowledge generated from the original study. • Encourage participation in Data Sharing consistent with the NIH principles for all non-NIH-funded research projects. •
Demonstrated willingness to support open science THE NIH SHALL: • Determine Data Maintenance and Support period • Define standards for data storage, maintenance, curation, pooling and versioning • Create user-friendly portal for data source registration • Create policies and standards for - inclusion of data contributors into publications - data quality, anonymization and harmonization for inclusion into pooled data resources • Create/enforce common consent language for data sharing, including distribution and utilization for secondary analyses • Define anticipated responsibilities for data sharing team, including response time and associated costs • Create hybrid Data Sharing model to include both federated and pooled data resources • Charter Data Access/Sharing Committee to develop policies and procedures for servicing data sharing requests, including requests from third-parties and commercial entities, and for policing/enforcing publication and acknowledgements • Expand "Data" definition to biospecimen, images, GWS files and their derivatives • Consider becoming an arbiter in disputes and conflicts arising from data sharing/utilization/citation process, especially when newly-developed indices determine higher/lower scores on grant applications.

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

- Systematic data/software products reporting in RPPR and Competitive Grant Renewal applications will strengthen evidence of productivity; - Grant proposals should formally mention contributions to data repositories;
- Standard for software sharing shall be developed. Several language-dependent software repositories/archives,

with existing standards on performance and documentation, allow uploading software programs. Citation system for such contributions may be developed to us in CVs and grant proposals (CRAN for R programs.) - Necessary to nurture several academic teams experienced in data collaboration/aggregation/sharing and in software platforms' for data capture, management, curation, transformation and sharing development/maintenance (www.NeuroBANK.org, www.alsdatabase.org, www.NeuroGUID.org) - Web-based data analytical tool may serve as playground for investigators to analyze external data when data sharing is legally impossible or to plug in their data into outside algorithms. Site's URL may be used in citation. The caveat is that the site shall be maintained indefinitely. Such sites shall be acknowledged and encouraged to continue (sample size selection website run by David Schoenfeld (MGH) gets 60,000+ hits/year) - Suggested incentives (and encouragements) may include: o Utilizing previous PI's history on data sharing for future considerations/continuation o Tracking shared datasets utilization in other studies and publications at par with citation index o Co-authorship of data contributors in publications that utilize their shared data (need to establish a minimum

dataset and minimum number of records to be considered) o Preferential period for data analyses on merged datasets to large data contributors

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

Utilization of DOI for the data/software citation is overdue. Challenges and solutions may include: - Development of SOPs to identify/enlist such resources and what to consider a validated and well-maintained resource - Development of life cycle requirements for such resources and responsibilities and incentives to third parties who may be taking over the resource if the original owner decides to step down - Shall the same DOI be considered if o The content of the existing resource was significantly changed (in which case could new contributors be treated at par with original ones in the resource citations) o Other data types were added by new contributors - Consider the following situations to resolve: o when to retire a DOI and start a new one even if some/all original data will migrate to the new resource o how to reference several DOIs if both contain the same data

b. Inclusion of a link to the data/software resource with the citation in the report

Yes

c. Identification of the authors of the Data/Software products

As part of the resource registration, it is necessary to enlist data contributors or PIs of the project(s) that originated the data (perhaps with names of sponsoring organizations). For shared/merged resources originated from consortia projects or multiple natural history studies, contributing effort may be acknowledged as well (e.g. considering number of records or data quality of such records, etc.)

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

As the DOI system will be developed, each resource will have a separate index that may be cited separately. The SOPs that spell out resource registration shall take care of such ambiguity. Same authors may be listed in several registered resources.

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

The challenge is to create standards, guidelines and SOPs to address the following: - what is a resource - how to identify it - who may register a resource and how to avoid duplication of resource registration - who will be policing the resource registry - how to address conflicts and complains - how to setup an arbitration panel to resolve issues on authorship and contributions - who can retire a resource - shall resource hosting site be acknowledged

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

NIH shall work with all stakeholders in clinical research continuum including foundations, industry, non-profits and patients. There is a wealth of resources and experiences/empirical data on sharing. Data ownership shall be addressed and agreed upon by all parties. Consider a situation when patients participating in a research projects require their data from the project's PI and donate the data to a data repository. Both the original PI's dataset and the data aggregator (say a disease consortium) got assigned DOIs. Which resource shall be cited, the original or both?

4. Any other relevant issues respondents recognize as important for NIH to consider

Essential for NIH research awardees to systematically create and preserve research project data and data documentation to facilitate data sharing and/or reproduction. • Facilitate reproduction of original analyses to increase the integrity of NIH- funded research findings; and • Promote data sharing to enable conduct of additional analyses using data from NIH-funded studies, thereby augmenting the knowledge generated from the original study. • Encourage participation in Data Sharing consistent with the NIH principles for all non-NIH-funded research projects. • Demonstrated willingness to support open science THE NIH SHALL: • Determine Data Maintenance and Support period • Define standards for data storage, maintenance, curation, pooling and versioning • Create user-friendly portal for data source registration • Create policies and standards for - inclusion of data contributors into publications - data quality, anonymization and harmonization for inclusion into pooled data resources • Create/enforce common consent language for data sharing, including distribution and utilization for secondary analyses • Define anticipated responsibilities for data sharing team, including response time and associated costs • Create hybrid Data Sharing model to include both federated and pooled data resources • Charter Data Access/Sharing Committee to develop policies and procedures for servicing data sharing requests, including requests from third-parties and commercial entities, and for policing/enforcing publication and acknowledgements • Expand "Data" definition to biospecimen, images, GWS files and their derivatives • Consider becoming an arbiter in disputes and conflicts arising from data sharing/utilization/citation process, especially when newly-developed indices determine higher/lower scores on grant applications.

Additional Comments

Submission Date

01/19/2017

Submitter Name

Abigail Goben

Name of Organization

University of Illinois at Chicago

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

No Domain Specified

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

• The highest priority data for data to be shared recommended by our researchers focused on raw experimental data (with documentation), physical samples which cannot be reproduced, and electronic health record data. Experimental data with accompanying metadata was described as of high value as it would allow for aggregation and extended secondary analysis. The emphasis from the researchers was the raw data particularly, not the summary data. As samples may be challenging to obtain and costly to keep, these were also identified as a priority for sharing. Finally, electronic health data, because of its myriad uses to obtain retrospective, cross-sectional, and cohort analyses as well as having specific value to individual patients was perceived as highly important. • These answers come from discussions between health science library faculty and a few biomedical researchers, both clinical and basic sciences, across our institution.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

• These answers come from discussions between health science library faculty and a few biomedical researchers, both clinical and basic sciences, across our institution. • The automatic answer for data retention is that it should be kept and made available forever or at least 100 years. Because of the long delay between initial discovery and acceptance into clinical practice, retention and sharing policies of 10-20 years are not unimaginable. The greatest issue identified at present was that funding is not available for research data storage after the duration of the grant. The researcher is not allowed to allocate funds in order to store and upgrade that storage after the end of the award period, which impedes researchers who are interested in storing, retention, and preservation of their data. • Basic science researchers indicated that in some instances, retention times of as little as 3 years may be appropriate as that could be the time from experiment to publication.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

• These answers come from discussions between health science library faculty and a few biomedical researchers, both clinical and basic sciences, across our institution. • Metadata, taxonomy, ontology, and documentation standards were repeatedly identified by researchers as an area that created a barrier to data sharing. By setting national standards for various types of data in collaboration with academic disciplines through societies, data sharing could be greatly improved. • Additional rules are seen as an unfunded burden on the Primary Investigator – the onus of compliance falls entirely to the researchers with minimal institutional support. • Not all data needs to be retained nor shared. There does not appear to be planning given to the need to selectively curate and weed data. • One researcher felt that no bench research data should ever be shared because of the cost of curation/conversion. • A lack of institutionally funded and supported solutions for even the data only used for publications. A solution would be to have one that was discipline agnostic and held storage capacity, had clear templates and had data curation assistance. • The need to protect sensitive data such as higher level mammal research, climate change data, etc from potentially hostile outsiders. • In an

era of bioinformatics, requiring data to be shared in short time periods leads to potential for scooping.

4. Any other relevant issues respondents recognize as important for NIH to consider

University Libraries have been and continue to be leaders in research data management education and support. While the informationist grants have provided some opportunities for collaboration and partnership, they are very narrowly focus and do not apply even to many large institutions. Further funding through the NLM to create education, tools, and to facilitate staffing in the libraries would enhance not only individual research projects, but meet needs across the institution outside of discipline specific lines. Journals are also placing increasing requirements on researchers. Coordination between the NIH and the Journals to assist in preventing mixed messages where researchers must try to negotiate between conflicting obligations would be welcome. Please see Chapter 2 http://www.ala.org/acrl/sites/ala.org.acrl/files/content/publications/booksanddigitalresources/digital/9780838988596_crd_v1_OA.pdf for more details.

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

• This wasn't seen as a particularly new obligation perceived with this, with the understanding that data were likely still to be raw or in process and access should be restricted to prevent misuse/misinterpretation. • There was some question of the value of data or software as an object at this stage. RPPRs do not contribute to promotion and tenure dossiers in any meaningful fashion, and are considered an accountability step in demonstrating that the work being contracted for is actually producing some products.

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

i. There was some hesitancy to requiring a DOI in an RPPR as the perception was that these are assigned to a final data set, something that is finished rather than ongoing. This comes partially from the limited number of sources from which one can obtain a DOI presently—how they accept data, etc. This could also put an unnecessary burden on researchers at smaller institutions who are not minting DOIs for their own repository. Questions were raised about managing version control. A further question was raised about how to manage increased data citation needs with page limitations.

b. Inclusion of a link to the data/software resource with the citation in the report

i. This was generally agreed as useful, again with limited circulation. Three links were identified as potentially useful: Data No Longer Being Used; Data Being Published; Data In Progress.

c. Identification of the authors of the Data/Software products

i. We encourage this and also including the institution(s)

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

i. It was generally agreed that this will vary between disciplines and one proscriptive answer is likely to run into immediate difficulties. Guidance on when this might be appropriate would be welcome.

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

i. This only should apply to final data, published data, or data no longer being used as many repositories do not allow ongoing data collection and updates to a record. This seems more useful for grant applications as opposed to RPPRs. ii. One suggestion was that there should also be a living will or other end of life plan for repositories for them to be considered so that if a repository can no longer host data there is a plan for migration to prevent data loss. This should come with institutional or NIH commitment

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

• Reward mechanisms currently do not encourage data sharing. New mechanisms need to be in place in order to facilitate a culture change. This must also have influence at study sections to encourage uptake. • Make things easy – add a checkbox on a form that says "put in your Data Citations here" • Engage research communities to develop data repositories and standards for data, one example was reporting Exercise Testing Data. Promote enhanced communication. • Create penalties for not sharing data. • Provide startup grants for institutions of all sizes. While sustainability is an option for many institutions, the FTE, initial infrastructure outlay, etc is prohibiting many institutions from engaging in research data management support in a truly effective way, leaving researchers to fend for themselves. • Provide grants for the curation of Dark Data, data that is no longer being used in the lab but which may be unique or unusual. • Facilitate development of scalable and easily customizable open source lab management and electronic notebook systems • Funding or research to promote interoperability with core services that provide data analysis or processing, sample analysis or processing, or other. Many institutions do not have data workflows or pipelines that are continuous or supported at an institutional level. For example, there is an opportunity with bioinformatics or analysis cores to create a pipeline from their work-for-hire projects and a long term digital preservation/archiving solution. The University Library has been recognized as a potential partner for this type of work.

4. Any other relevant issues respondents recognize as important for NIH to consider

University Libraries have been and continue to be leaders in research data management education and support. While the informationist grants have provided some opportunities for collaboration and partnership, they are very narrowly focus and do not apply even to many large institutions. Further funding through the NLM to create education, tools, and to facilitate staffing in the libraries would enhance not only individual research projects, but meet needs across the institution outside of discipline specific lines. Journals are also placing increasing requirements on researchers. Coordination between the NIH and the Journals to assist in preventing mixed messages where researchers must try to negotiate between conflicting obligations would be welcome. Please see Chapter 2 http://www.ala.org/acrl/sites/ala.org.acrl/files/content/publications/booksanddigitalresources/digital/9780838988596_crd_v1_OA.pdf for more details.

Additional Comments

Submission Date

01/19/2017

Submitter Name

niels volkmann

Name of Organization

sanford burnham prebys

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

structural biology

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

Primary data, including movie frames where applicable, should be highest priority. This is especially true for cellular tomography where the content of the tomograms is very rich and has a lot of information that is not likely to be fully mined in one study. Reasons are (1) if protocols are stored with the data, subsequent results can be reproduced; (2) quality of results may dramatically improve with new software tools; (3) this would provide an adequate workbench for method development; (4) in case of cellular tomography, where the information content is unlikely to be mined in a single study, simple re-mining data with a different view-point may lead to very interesting new results.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

For all practical purposes, the primary data should be stored permanently. There should ideally be a two-tier process. In the first tier NIH should provide sufficient funds for researchers to build up and maintain local storage capacity with adequate space allocation and fast access for day-to-day research. There should be a second tier where researchers can deposit all primary data for the long haul, without cost to them and with proper intellectual property rules in place. The model can be informed by the PDB or similar types of data depositories.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

Storage technology is a fast moving field. NIH needs to provide adequate mechanisms for researchers not only to build up adequate local storage capacity but also to update the systems to current standards on a regular basis. This should not only include support for hardware but also for software layers that improve storage performance.

4. Any other relevant issues respondents recognize as important for NIH to consider

Quality control could be a potentially issue. Publications are peer-reviewed, data and software, unless they are published, do not go through a peer-review process.

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

Reporting is not enough. Recognition of some sort is essential to encourage sharing. Encouraging citability of software and data sets by journals as well as NIH would be an important step forward.

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource,

such as a Digital Object Identifier (DOI)

This would be a preferred mechanism because DOIs are persistent and would not change when researchers move or website names change etc.

b. Inclusion of a link to the data/software resource with the citation in the report

Using DOIs that are linked to the resource would be preferable.

c. Identification of the authors of the Data/Software products

Authorship of data and software should be treated the same way authorship is treated for research publications.

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

That should depend on the nature of the study that uses the citation. Studies that focus on a small number of data sets should cite the specific data sets they are using, studies that perform large-scale studies using multiple data sets should use aggregated citations.

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

This is essential and can also be handled through the DOI mechanism.

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

Encourage the use of pre-print servers such as arXiv and bioRxiv. Treat data and software in similar ways as publications.

4. Any other relevant issues respondents recognize as important for NIH to consider

Quality control could be a potentially issue. Publications are peer-reviewed, data and software, unless they are published, do not go through a peer-review process.

Additional Comments

Submission Date

01/19/2017

Submitter Name

Rebecca Reznik-Zellen

Name of Organization

University of Massachusetts Medical School

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

Biomedical

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

4. Any other relevant issues respondents recognize as important for NIH to consider

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

b. Inclusion of a link to the data/software resource with the citation in the report

c. Identification of the authors of the Data/Software products

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

4. Any other relevant issues respondents recognize as important for NIH to consider

Additional Comments

NIHRFIonStrategiesforDataManagementSharingandCitation.docx (21 KB)

Response to NIH RFI on Strategies for Data Management, Sharing, and Citation

Rebecca Reznik-Zellen, Regina Raboin for Lamar Soutter Library
Diego Vazquez for Office of Sponsored Programs
University of Massachusetts Medical School

Section I: Data Sharing Strategy Development

With exceptions for sensitive and personally identifiable information, digital research products, including complete datasets[1] (particularly of basic science research), should be preserved indefinitely. The life of data has been shown to extend beyond a given project grant cycle, beyond the employment of a principle investigator or even the changing priorities of a field.[2] Data feed historical inquiry, validate published claims, support new lines of investigation, and enable large-scale global questions to be probed. Deposit into managed repositories will ensure that these assets remain viable for the longest term possible.

Managed repositories affiliated with research institutions or with national research bodies are driven by their mission to ensure the ongoing access to and viability of assets under their care. These repositories follow international information science standards and best practices, such as TRAC[3] and FAIR[4], and are advocates for open access to research in a broad sense. In contrast, commercial publishers offer platforms for data hosting as a secondary or tertiary service. There is no expectation or necessity for these platforms to be freely accessible, comprehensive, or preservation-minded. In addition, as private entities, publishers of different types (e.g. commercial, society) have different capacities to build and offer these services. Therefore, libraries, managed repositories, and data centers (including third-party data hosting platforms such as Figshare and Zenodo) should be given priority as infrastructure for research data preservation and sharing.

The long-term resource requirements for managed repositories include the costs of physical infrastructure such as storage, networking, and security; the provision of ongoing metrics, education and training materials, use cases, and other supplementary documentation; and a skilled and distributed workforce to maintain infrastructure and mediate data deposition. The responsibility for subsidizing this infrastructure should fall to the organizations that fund research, rather than to those who conduct the research. In learning from the implementation of the NIH Public Access Policy, funders should provide mechanisms to enable libraries, managed repositories, and data centers to ensure ongoing support for data preservation and sharing. These mechanisms could build on proven strategies, such as: budget allowances for data

---

[1] A complete dataset is fully described, has been assigned a permanent unique identifier, and is in its final form. The dataset does not need to be linked to a published paper, but is complete in itself and is available for secondary analysis.
[2] National Academy of Sciences. Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age (2009). Summary: Promoting the Stewardship of Research Data, p.7-9.
[3] https://www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying/trac
[4] https://www.force11.org/group/fairgroup/fairprinciples

preservation and sharing for individual grant projects; block grants or center awards to institutions for the development and evolution of centralized services and programming[5]; or a credit model for access to data sharing and preservation infrastructure[6].

Before any broad policy for data sharing and preservation can be enforced, the infrastructure must be in place and accessible to those that will be expected to use it. In addition, minimum criteria for shared data should be outlined and include: basic metadata, a permanent unique identifier, and discoverability. Data management plans play a special role in articulating the data sharing and preservation strategies for a given project; they assist the investigator in understanding the value of their data, preserving its meaning, and respecting its future use. They document the institutional support services and resources needed to successfully complete a grant and be in compliance with federal and other funders. The requirements of the data management plan, particularly with respect to data sharing and preservation, must reflect what is being asked of the investigator. Integrating data management planning, sharing, and preservation into the primary project application and project workflow--rather than relegating these issues to a supplemental document--could raise the status of these components of the research lifecycle and the services that support them, as well as mitigate the perception of the data management plan as a burdensome "add-on" in writing the grant proposal.[7]

Section II: Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

Our institution is concerned about the impact of requiring reporting of data and software sharing in annual Research Performance Progress Reports (RPPRs), as submitting data and software prior to the end of the grant cycle may have an unintended chilling effect on data sharing behaviors. While we recognize the importance of demonstrating compliance during the course of a project, requiring formal citation to data products that are under development could be counterproductive to efforts at making data a first class research product.

Initiatives for data citation[8] encourage people to share data in a way that uses a persistent, unique identifier and that points to well-described and documented, ideally finished data. Electronic laboratory notebook systems have made the minting of dois fast and easy, and could facilitate the provision of dois for mid-project data. However, some of these systems do not provide guidance on best practices for the assignment of dois, they do not explicitly account for versioning in their metadata, and they may apply automatic licenses that are not modifiable to content with a doi. Minting dois for the sake of a progress report may diffuse initiatives for best practices in data citation.

What other approaches could be used by NIH-funded investigators to ensure appropriate steps are taken to share and preserve data?

---

[5] Such as BD2K (https://datascience.nih.gov/bd2k)
[6] For example, Commons Credits (https://datascience.nih.gov/BlogCommonsCreditsModelPilot)
[7] See for example: https://www.nsf.gov/pubs/policydocs/pappguide/nsf15001/gpg_2.jsp#dmp
[8] For example: https://www.datacite.org/ and https://www.force11.org/

- Utilize popular data versioning systems, such as Github, to allow researchers to locate and report early versions and prototypes of deposited data and or code. This path would allow for data sharing and re-use during the active life of the grant, without proliferating dois.
- Demonstrate when and where grant funding has been allocated to data sharing and preservation through the reporting mechanism.
- Leverage the role of the data management plan and submit protocols and training materials, which should be in place at the beginning stages of a project and reflect the workflow changes over time, as reportable items for the progress report.
- Additionally, the NIH could incentivize data and software sharing and preservation by implementing best practices in sharing and preservation for repositories and data centers. An NIH 'imprimatur' or badge could be given to those institutional, government, and independent repositories and data centers that adopt these standards.[9]

We recommend that NIH only require the submission of a data reporting/sharing report during the final year of a competitive segment using the interim or Final-RPPR. This is a logical data capture point that will minimize the frequency and annual burden of reporting on investigators while capturing a much broader span of data than the annual requirement would capture. If the NIH does choose to require or recommend data reporting during the course of the grant cycle, the reporting instructions should be constructed in such a way that a PI can successfully comply with data at various stages of completion for that reporting period. A PI should have the opportunity to report changes to data gathering and sharing protocols without fear of penalty. Such instruction would also accommodate and encourage PIs who are able to report data of "considerable importance" or completed data that may be ready for sharing.

[9] See for example: https://doaj.org/publishers#seal

Submission Date

01/19/2017

Submitter Name

Melissa Haendel

Name of Organization

The Monarch Initiative

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

Data integration, rare disease research

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

Primary experimental data, especially that used to generate figures. Value: Improves transparency and the ability to assess whether a repeated experiment is consistent with the original. Important is to include negative experimental data and pre-analytical data (such as clinical trial registration data). Curated knowledge bases, whether they are curated from the primary literature or directly from data. Value: Improves findability and accessibility of data that would otherwise require vast review and retention of the literature. Ontologies, vocabularies, standards, models, vocabularies, APIs, ID schemas, docker containers, etc. Value: Anything that makes data move more effectively, efficiently, and reproducibly.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

In order to make things useful and available for use long term, they must be made available in formats and access structures that openly utilizable, eg. not in proprietary formats such as MS word. Given this caveat, we recommend the following according to the data types described in the above question: Primary experimental data: Depends on size, complexity, and cost as well as to the degree of lossiness relative to more compact forms of the knowledge such as curated knowledge bases. If cost or size are barriers, then permanent metadata records should be considered. Curated knowledge bases: Should be persistently sustained forever (subsumed / archived is OK; versioning critical). Ontologies, vocabularies, standards, models, vocabularies, APIs, ID schemas, docker containers, etc.: Should be persistently sustained forever (subsumed / archived is OK; versioning critical).

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

Bandwidth. Dedicating people. Solutions: Embed data stewards or librarians into all research projects. Require collaborations with knowledgebases and experts in data stewardship. Funding. Solution: require data sharing costs (especially FTE) be included in grant budgets; fund/sustain centralized resources (such as repositories and curators). Inconsistent Policies. Solution: data sharing policies for journals should be standardized [1]. Poorly identified data. Solution: Require consistent data citation and an identifier policy for all DMPs. Improper management of identifiers is a key problem[2] and the community has made extensive recommendations for data providers[3][4]. Lack of Incentives. Solution: The most important factor is creating a culture wherein it is cool to share [5]; this is especially hard without making the metrics of sharing public and better tracked. Other potential carrots are summarized in a recent study.[6] Confusion. Researcher confusion about how to share/where to share. Solution: Portals or helpdesk function that guide researchers through the process. Lack of Attribution. Changing the culture so that data reuse benefits rather than competes with the sharer. Solution: Attribute all contributions. Lack of examples. Many researchers don't know what a good data sharing plan looks like; there are three examples on the NIH website[7] but these pertain to clinical data and are therefore less relevant for the open access infrastructure. Solution: Several exemplary DMPs should be made available for different research (eg. biochemical, genomic, clinical case study, basic, translational, clinical trial). DMPs for

EVERY funded grant should be publicly accessible within NIH reporter.

4. Any other relevant issues respondents recognize as important for NIH to consider

- Require data sharing for all grants - not just grants for $500,000 or greater (which is the current policy) -Genomics data is often singled out, it would be nice to note other data types to also be shared, for example, phenomics, metabolomics, etc. -Include data stewards and biocurators on grant review panels. Work with the Biocuration society to determine expertise. Not all knowledgeable experts are primary grant writers or primary authors of manuscripts- this is a fundamental flaw in evaluating the informatics and data stewardship in the grant review process. - Build in collaborative funding to work together rather than compete. The new NCI ITCR program does this. Conversely, the NCATS OT3 mechanism created a collaborative consortium to build infrastructure based on expertise rather than traditional competitive mechanisms. - Licensure and availability of data reuse information whenever referencing any data, so that it can be reused legally and ethically. Licenses should not require negotiation and licenses themselves should be legally re-distributable without engaging legal counsel. Not all data resources are free to use, derive, and redistribute, even if they are publicly funded and seemingly publicly available. This is true for almost all existing NIH-funded resources. We note that considerations for data are significantly different than those for software and they must be considered separately (see this blog for example[31]). For licensing metrics see also the other RFI response[8].

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

The requirement to roll out data sharing plan requirements to all grants (not just those over 500K) is an excellent first step. However, it would be better still to couple the data sharing plan with reporting requirements to ensure that plans are followed through. Moreover plans themselves should be more structured, requiring elements such as the creation and reuse of identifiers, licensure of all products (software, ontologies, data, publications), as well as the ontologies or standards that will be used, if any. All data sharing plans should require an archiving or metadata persistence plan.

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as, but not limited to Digital Object Identifier (DOI). We note that in the W3C dataset description community standard[4] and in the FORCE11 data citation manuscript[11], there are numerous persistent identifiers that are recommended. Different communities use different persistent IDs and this diversity MUST be supported. Primary data: Dryad and Figshare (which use DOIs) are well suited to accommodating the long-tail of experimental data. However, a number of considerations should be factored in for other kinds of data (and there are many). These considerations are summarized by McMurry et al [here]. N2t.net identifiers.org and others have recently taken steps toward harmonizing identifier prefixes from databases that do not use DOIs; this work will make persistent identifiers easier to cite.[12,13] The Force 11 Software Citation group has published sensible guidelines for the citation of software.[14] Knowledge bases should follow the best practice for identifier design, provision, and reuse. [3]

b. Inclusion of a link to the data/software resource with the citation in the report

As scientific data becomes ever larger, more interconnected, and more natively web-based, the definition of ‘citation’ should be broadened.[15] We recommend each data provider transparently document how they wish to BE cited; this practice encourages attribution and make it easier to comply.[3] Cultural norms toward FAIR-TLC won't shift overnight: Standards must permit proportional wins at 'good' 'better' 'best'. Making the right thing the easier thing is an essential strategy toward FAIRness. To this end, we recommend the following: Include citation of data and other scholarly products within reference managers. Better adoption of JATS[16] and DATS[17] by journals; better software tooling for authors to produce JATS-compliant data references, preferably with dual PIDs. Citing dual PIDs (one native one from the provider, and corresponding one issued by a resolver) would a) make registries more receptive to the use of 3rd party resolver URIs for citation while also b) providing an inexpensive failsafe against 3rd party resolver failure. Text mining

applications such as Jannotator[18] whereby authors can upload a manuscript and detect possible data elements / entities and prompt the author to include the citation to them. Data citations should have the same structure as other kinds of references, but include the following elements: author(s)/contributor, title, repository, year, version and persistent identifier. For curated knowlegebases, the citations may be structured differently but the provenance should nevertheless be transparent. More examples of formatted data references are found in the JATS specification[17] and the HCLS dataset working group [4].

c. Identification of the authors of the Data/Software products

Attribution is a crucial social component to getting sharing to work. Moreover, the prospect of being attributed serves to increase the quality of the work itself. For this reason, more granular attribution is better, where applicable. Attribution should leverage identifiers for people (such as ResearcherID or ORCID) and organizations (such as Digital Science GRID or OCLC). The Force11 Attribution Working group [19] and the Open Research Information system[20] VIVO-ISF have been working to define contribution roles for all data, software, and other scholarly product contributions. This creates a computational framework for foundational work done by the Credit taxonomy[21] working group. Contribution roles have been implemented in a number of places, such as in Wikidata[22], ontologies, biosketches, and in dataset descriptions. These efforts ensure that data and data-related products are first-class scholarly products and not merely relegated to non computable acknowledgement sections of manuscripts. It should be noted that often contributions to curated knowledgebases, ontologies, and other data artifacts often go unattributed and are therefore difficult to value. Some communities, such as the Open Biomedical Ontologies community have therefore implemented citation policies to track contribution[23].

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

At Monarch, we work in the realm of curated knowlegebases. There are many people who contribute to the overall database in different ways. For instance, the authors of the primary paper, the curators, the database maintainers etc. In many cases, the granular citation is the only one that makes sense. For instance, if I'm publishing a paper about a metastudy of all Fanconi-associated variants in ClinVar, there is currently no way to reference this set explicitly unless the clinvar ID is pasted for all. Nor would it make sense to reference each of the primary papers associated with the variants. What is needed is a mechanism[24,25] of attribution whereby citing a resource transitively attributes all those in the chain of contributions.

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

Identifying repositories *themselves* is a tricky business. Confounding issues include but are not limited to: Nestedness (e.g. is the NCBI Gene database represented together with, separately from, or underneath "NCBI") Consortial efforts (e.g. are the BioSamples database at EBI and the corresponding BioSamples database at NCBI represented as a single resource or as separate ones?) Overlapping remit. Various groups have generated citable identifiers for repositories. Such groups include but are not limited to NAR[26], BioDBCore[27], Re3Data[28], identifiers.org[29], and SciCrunch[30]. The scope of what is identified differs between these groups and there are different approaches to the nestedness and consortial challenges making the mapping of the identifiers between these authorities rather fraught. Further, most of the repositories themselves do not refer to themselves or endorse any of these identifier strategies. The Biocuration society aims to help alleviate this problem. This information should be part of the required metadata for the data/software resource – the publisher property – as described in the Force11 recommendations for data repositories [10].

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

-specific funding that will cover the costs of data sharing, including creating infrastructure and maintaining records. This could be in the form of new grants that are specifically for this cause. It should also be made clear in the grant instructions that NIH funding can be used to cover costs associated with data and software sharing. -incentivize data

sharing by creating specific awards; these awards can also be listed as achievements on RPPRs and grant renewals -data sharing plan should impact the score of the grant application

4. Any other relevant issues respondents recognize as important for NIH to consider

- Require data sharing for all grants - not just grants for $500,000 or greater (which is the current policy) -Genomics data is often singled out, it would be nice to note other data types to also be shared, for example, phenomics, metabolomics, etc. -Include data stewards and biocurators on grant review panels. Work with the Biocuration society to determine expertise. Not all knowledgeable experts are primary grant writers or primary authors of manuscripts- this is a fundamental flaw in evaluating the informatics and data stewardship in the grant review process. - Build in collaborative funding to work together rather than compete. The new NCI ITCR program does this. Conversely, the NCATS OT3 mechanism created a collaborative consortium to build infrastructure based on expertise rather than traditional competitive mechanisms. - Licensure and availability of data reuse information whenever referencing any data, so that it can be reused legally and ethically. Licenses should not require negotiation and licenses themselves should be legally re-distributable without engaging legal counsel. Not all data resources are free to use, derive, and redistribute, even if they are publicly funded and seemingly publicly available. This is true for almost all existing NIH-funded resources. We note that considerations for data are significantly different than those for software and they must be considered separately (see this blog for example[31]). For licensing metrics see also the other RFI response[8].

Additional Comments

ResponsetoRFIonDataSharing2017.pdf (142 KB)

Practical Steps Toward FAIR-TLC Data Management, Citation, and Sharing

Authors: Julie A McMurry, Lilly Winfree, Melissa Haendel on behalf of the Monarch Consortium

NIH Request for Information (RFI)
Strategies for NIH Data Management, Sharing, and Citation, NOT-OD-17-015
URL: https://grants.nih.gov/grants/guide/notice-files/NOT-OD-17-015.html

Repositories You or Your Organization Primarily Utilize

The Monarch Initiative is a large-scale data integration project, and so this is a long list of data sources, annotations, knowledgbases, and ontologies:

Animal QTLDB
APB
BGee
BioGRID
BIOGRID
ChEBI
CHR: Chromosome Ontology
CID
CL: Cell Ontology
ClinVar
CLO: Cell Line Ontology
CMMR
CMO: Clinical Measurements Ontology
Coriell
COSMIC
CTD: Comparative Toxicogenomics Database
dbSNP
dbVar
DC: Dublin Core
DECIPHER
dictyBase
DOID
DrugBank
EC code
ECO: Evidence Code Ontology
EcoGene
EDAM: Data and Methods Ontology
EFO: Experimental Factor Ontology
EMMA Infrafrontier
Encode
ENSEMBL
ENVO: Environment Ontology
EOM: Elements of morphology
ERO: eagle-i resource ontology
FALDO: Feature Annotation Location Description Ontology

Fantom5
FlyBase
GenBank
GENO: Genotype Partonomy Ontology
GO: Gene Ontology
GTEx
HGMD
HGNC
HP: Human Phenotype Ontology
HPRD
Human Phenotype Ontology
IAO: Information Artifact Ontology
IMPC: International Mouse Phenotyping Consortium
Jackson Laboratories
KEGG
LPT: Livestock Phenotypic Trait Ontology
MA: Mouse Anatomy Ontology
MedGen: Medical Genetics
MeSH
MGI
miRBase
MMRRC
MP: Mammalian Phenotype Ontology
MPATH: Mammalian Pathology Ontology
MPD: Mouse Phenome Database
MyGene.info
NBO: NeuroBehavior Ontology
NCBI CCDS
NCBI GeneReviews
NCBI HOMOLOGENE
NCBIAssembly
NCBIGene
NCBIGenome
NCBIProtein
NCBITaxon
NCIMR
OBA: Ontology of Biological Attributes
OBAN: Open Biomedical Annotation Model
OBI: Ontology of Biomedical Investigations
OMIA
OMIM
ORPHANET
Orphanet: rare diseases and orphan drugs
PANTHER Orthology Database
PATO: Phenotypic Quality Ontology
PCO: Population and Community Ontology
PDB
PomBase

PRO: protein ontology
PW: pathway ontology
RBRC
REACTOME
RefSeq
RGD
RO: Relationship Ontology
SEPIO
SGD
SIO: SemanticScience Integrated Ontology
SNOMED
SO: Sequence Ontology
Statistics Ontology
SwissProt
TAIR
TrEMBL
UBERON
UCSC
UMLS
UniProtKB
UO: units of measurements
UPHENO
VIVO
VT: Vertebrate Trait Ontology
Wormbase
XCO: Experimental Conditions Ontology
Xenbase
ZFIN: Zebrafish Information Network

SECTION I. Data Sharing Strategy Development.
The NIH seeks comment on any or all of the following topics to help formulate strategic approaches to prioritizing its data management and sharing activities:

1. The highest-priority types of data to be shared and value in sharing such data

1) **Primary experimental data**, especially that used to generate figures. Value: Improves transparency and the ability to assess whether a repeated experiment is consistent with the original. Important is to include negative experimental data and pre-analytical data (such as clinical trial registration data).
2) **Curated knowledge bases**, whether they are curated from the primary literature or directly from data. Value: Improves findability and accessibility of data that would otherwise require vast review and retention of the literature.
3) **Ontologies, vocabularies, standards, models, vocabularies, APIs, ID schemas, docker containers, etc.** Value: Anything that makes data move more effectively, efficiently, and reproducibly.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

In order to make things useful and available for use long term, they must be made available in formats and access structures that openly utilizable, eg. not in proprietary formats such as MS word. Given this caveat, we recommend the following according to the data types described in the above question:

1) **Primary experimental data:** Depends on size, complexity, and cost as well as to the degree of lossiness relative to more compact forms of the knowledge such as curated knowledge bases. If cost or size are barriers, then permanent metadata records should be considered.
2) **Curated knowledge bases:** Should be persistently sustained forever (subsumed / archived is OK; versioning critical).
3) **Ontologies, vocabularies, standards, models, vocabularies, APIs, ID schemas, docker containers, etc.:** Should be persistently sustained forever (subsumed / archived is OK; versioning critical).

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

1) **Bandwidth**. Dedicating people. Solutions: Embed data stewards or librarians into all research projects. Require collaborations with knowledgebases and experts in data stewardship.
2) **Funding**. Solution: require data sharing costs (especially FTE) be included in grant budgets; fund/sustain centralized resources (such as repositories and curators).
3) **Inconsistent Policies**. Solution: data sharing policies for journals should be standardized [1].
4) **Poorly identified data.** Solution: Require consistent data citation and an identifier policy for all DMPs. Improper management of identifiers is a key problem[2] and the community has made extensive recommendations for data providers[3][4].
5) **Lack of Incentives**. Solution: The most important factor is creating a culture wherein it is cool to share [5]; this is especially hard without making the metrics of sharing public and better tracked. Other potential carrots are summarized in a recent study.[6]
6) **Confusion.** Researcher confusion about how to share/where to share. Solution: Portals or helpdesk function that guide researchers through the process.
7) **Lack of Attribution.** Changing the culture so that data reuse benefits rather than competes with the sharer. Solution: Attribute all contributions.
8) **Lack of examples.** Many researchers don't know what a good data sharing plan looks like; there are three examples on the NIH website[7] but these pertain to clinical data and are therefore less relevant for the open access infrastructure. Solution: Several exemplary DMPs should be made available for different research (eg. biochemical, genomic, clinical case study, basic, translational, clinical trial). DMPs for EVERY funded grant should be publicly accessible within NIH reporter.

4. Any other relevant issues respondents recognize as important for NIH to consider

In response to the RFI on repository metrics [8], we augmented FAIR with Traceability, Licensure, and Connectedness. We applied FAIR-TLC to the Open Science Prize candidates[9]. A number of candidates did not meet criteria for access, licensing, or identifiers - extensively limiting data citation or reuse.

FINDABILITY

F1: Discoverable through various external mechanisms

F2: Contents/components are well documented and searchable

ACCESSIBILITY

A1: Diverse data access mechanisms

A2: Well structured and provisioned APIs

A3: Understandable data and scope

INTEROPERABILITY

I1: Identifiers should be simple and durable
I2: Vocabularies, Ontologies, and exchange standards
I3: Versioning changes are documented, keep prior versions

REUSE

R1: Is the resource being used?
R2: Impact: how is the data being used?
R3: Awareness and responsiveness to user needs
R4: Quality of data content and service

TRACEABILITY

T1: Provenance. Tracing the origin/creators of the data
T2: Accurately attributing contributions

LICENSURE

L1: Documented, clear, standard, minimally restrictive, contactable
L2: Transparent about flowthrough implications

CONNECTEDNESS

C1: Having diverse data in the same warehouse can be a good starting point, but it does not make the data inherently more usable or integrated.

All materials created by defunct organizations, or by deceased people is one of thorniest copyright problems; licenses should explicitly state the authorized scope of use after the copyright holder ceases to exist.[10] Such terms should cover the record and the identifier for the record, so that if a journal or database disappears, the identifier itself (e.g. a DOI) could be rerouted to its new location.

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

The NIH seeks comment on any or all of the following topics:

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

The requirement to roll out data sharing plan requirements to all grants (not just those over 500K) is an excellent first step. However, it would be better still to couple the data sharing plan with reporting requirements to ensure that plans are followed through. Moreover plans themselves should be more structured, requiring elements such as the creation and reuse of identifiers, licensure of all products (software, ontologies, data, publications), as well as the ontologies or standards that will be used, if any. All data sharing plans should require an archiving or metadata persistence plan.

2. Important features of technical guidance for data and software citation in reports to NIH, which may include:

1) Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as, but not limited to Digital Object Identifier (DOI). We note that in the W3C dataset description community standard[4] and in the FORCE11 data citation manuscript[11], there are numerous persistent identifiers that are recommended. Different communities use different persistent IDs and this diversity MUST be supported.

a) Primary data: Dryad and Figshare (which use DOIs) are well suited to accommodating the long-tail of experimental data. However, a number of considerations should be factored in for other kinds of data (and there are many). These considerations are summarized in the S4 Supplement of McMurry et al[3]. N2t.net identifiers.org and others have recently taken steps toward harmonizing identifier prefixes from databases that do not use DOIs; this work will make persistent identifiers easier to cite.[12,13]
b) The Force 11 Software Citation group has published sensible guidelines for the citation of software.[14]
c) Knowledge bases should follow the best practice for identifier design, provision, and reuse. [3]

2) Inclusion of a link to the data/software resource with the citation in the report

As scientific data becomes ever larger, more interconnected, and more natively web-based, the definition of 'citation' should be broadened.[15] We recommend each data provider transparently document how they wish to BE cited; this practice encourages attribution and make it easier to comply.[3]

Cultural norms toward FAIR-TLC won't shift overnight: Standards must permit proportional wins at 'good' 'better' 'best'. Making the right thing the easier thing is an essential strategy toward FAIRness. To this end, we recommend the following:

- Include citation of data and other scholarly products within reference managers.
- Better adoption of JATS[16] and DATS[17] by journals; better software tooling for authors to produce JATS-compliant data references, preferably with dual PIDs. Citing dual PIDs (one native one from the provider, and corresponding one issued by a resolver) would a) make registries more receptive to the use of 3rd party resolver URIs for citation while also b) providing an inexpensive failsafe against 3rd party resolver failure.
- Text mining applications such as Jannotator[18] whereby authors can upload a manuscript and detect possible data elements / entities and prompt the author to include the citation to them.

Data citations should have the same structure as other kinds of references, but include the following elements: author(s)/contributor, title, repository, year, version and persistent identifier. For curated knowlegebases, the citations may be structured differently but the provenance should nevertheless be transparent. More examples of formatted data references are found in the JATS specification[17] and the HCLS dataset working group [4].

3) Identification of the authors and their contributions of the Data/Software products

Attribution is a crucial social component to getting sharing to work. Moreover, the prospect of being attributed serves to increase the quality of the work itself. For this reason, more granular attribution is better, where applicable. Attribution should leverage identifiers for people (such as ResearcherID or ORCID) and organizations (such as Digital Science GRID or OCLC). The Force11 Attribution Working group [19] and the Open Research Information system[20] VIVO-ISF have been working to define contribution roles for all data, software, and other scholarly product contributions. This creates a computational framework for foundational work done by the Credit taxonomy[21] working group. Contribution roles have been implemented in a number of places, such as in Wikidata[22], ontologies, biosketches, and in dataset descriptions. These efforts ensure that data and data-related products are first-class scholarly products and not merely relegated to non computable acknowledgement sections of

manuscripts. It should be noted that often contributions to curated knowledgebases, ontologies, and other data artifacts often go unattributed and are therefore difficult to value. Some communities, such as the Open Biomedical Ontologies community have therefore implemented citation policies to track contribution[23].

.

4) **Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately**

At Monarch, we work in the realm of curated knowlegebases. There are many people who contribute to the overall database in different ways. For instance, the authors of the primary paper, the curators, the database maintainers etc. In many cases, the granular citation is the only one that makes sense. For instance, if I'm publishing a paper about a metastudy of all Fanconi-associated variants in ClinVar, there is currently no way to reference this set explicitly unless the clinvar ID is pasted for all. Nor would it make sense to reference each of the primary papers associated with the variants. What is needed is a mechanism[24,25] of attribution whereby citing a resource transitively attributes all those in the chain of contributions.

5) Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

Identifying repositories *themselves* is a tricky business. Confounding issues include but are not limited to:

- Nestedness (e.g. is the NCBI Gene database represented together with, separately from, or underneath "NCBI")
- Consortial efforts (e.g. are the BioSamples database at EBI and the corresponding BioSamples database at NCBI represented as a single resource or as separate ones?)
- Overlapping remit. Various groups have generated citable identifiers for repositories. Such groups include but are not limited to NAR[26], BioDBCore[27], Re3Data[28], identifiers.org[29], and SciCrunch[30]. The scope of what is identified differs between these groups and there are different approaches to the nestedness and consortial challenges making the mapping of the identifiers between these authorities rather fraught. Further, most of the repositories themselves do not refer to themselves or endorse any of these identifier strategies. The Biocuration society aims to help alleviate this problem.

This information should be part of the required metadata for the data/software resource – the publisher property – as described in the Force11 recommendations for data repositories [10].

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

-Specific funding that will cover the costs of data sharing, including creating infrastructure and maintaining records. This could be in the form of new grants that are specifically for this cause. It should also be made clear in the grant instructions that NIH funding can be used to cover costs associated with data and software sharing.

-Incentivize data sharing by creating specific awards; these awards can also be listed as achievements on RPPRs and grant renewals

-Data sharing plan should impact the score of the grant application

4. Any other relevant issues respondents recognize as important for NIH to consider

- Require data sharing for all grants - not just grants for $500,000 or greater (which is the current policy)

-Genomics data is often singled out, it would be nice to note other data types to also be shared, for example, phenomics, metabolomics, etc.

-Include data stewards and biocurators on grant review panels. Work with the Biocuration society to determine expertise. Not all knowledgeable experts are primary grant writers or primary authors of manuscripts- this is a fundamental flaw in evaluating the informatics and data stewardship in the grant review process.

- Build in collaborative funding to work together rather than compete. The new NCI ITCR program does this. Conversely, the NCATS OT3 mechanism created a collaborative consortium to build infrastructure based on expertise rather than traditional competitive mechanisms.

- Licensure and availability of data reuse information whenever referencing any data, so that it can be reused legally and ethically. Licenses should not require negotiation and licenses themselves should be legally re-distributable without engaging legal counsel. Not all data resources are free to use, derive, and redistribute, even if they are publicly funded and seemingly publicly available. This is true for almost all existing NIH-funded resources. We note that considerations for data are significantly different than those for software and they must be considered separately (see this blog for example[31]). For licensing metrics see also the other RFI response[8].

References

1. Vasilevsky NA, Minnier J, Haendel MA, Champieux RE. Reproducible and reusable research: Are journal data sharing policies meeting the mark? [Internet]. PeerJ Preprints; 2016 Nov. Report No.: e2588v1. doi:10.7287/peerj.preprints.2588v1

2. Vasilevsky NA, Brush MH, Paddock H, Ponting L, Tripathy SJ, Larocca GM, et al. On the reproducibility of science: unique identification of research resources in the biomedical literature. PeerJ. 2013;1: e148.

3. McMurry J, 40 Additional Authors. Identifiers For The 21St Century: How To Design, Provision, And Reuse Identifiers To Maximize Data Utility And Impact. Zenodo; 2016; doi:10.5281/zenodo.163459

4. Gray AJG, Baran J, Marshall MS, Dumontier M. Dataset descriptions: HCLS community profile. Interest group note, W3C (May 2015) http://www w3 org/TR/hcls-dataset. 2015; Available: https://www.w3.org/TR/hcls-dataset/

5. The Tax Man Nudgeth: Full Transcript - Freakonomics. In: Freakonomics [Internet]. 3 Apr 2013 [cited 19 Jan 2017]. Available: http://freakonomics.com/2013/04/03/the-tax-man-nudgeth-full-transcript/

6. van Den Eynden V, Knight G, Vlad A, Radler B, Tenopir C, Leon D, et al. Survey of Wellcome

researchers and their attitudes to open research. figshare; 2016; 66.

7. NIH Data Sharing Policy and Implementation Guidance [Internet]. [cited 19 Jan 2017]. Available: https://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm

8. Haendel M, Su A, McMurry J. Metrics To Assess Value Of Biomedical Digital Repositories: Response To Rfi Not-Od-16-133 [Internet]. Zenodo; 2016. doi:10.5281/zenodo.203295

9. Musings about the Open Science Prize. In: FORCE11 [Internet]. 31 Dec 2016 [cited 19 Jan 2017]. Available: https://www.force11.org/blog/musings-about-open-science-prize

10. Nancy Sims on Twitter. In: Twitter [Internet]. [cited 19 Jan 2017]. Available: https://twitter.com/CopyrightLibn/status/752734872578338819

11. Starr J, Castro E, Crosas M, Dumontier M, Downs RR, Duerr R, et al. Achieving human and machine accessibility of cited data in scholarly publications. PeerJ Comput Sci. 2015;1. doi:10.7717/peerj-cs.1

12. Wimalaratne S, Juty N, Kunze J, Janée G, McMurry JA, Beard N, et al. Uniform Resolution of Compact Identifiers for Biomedical Data [Internet]. bioRxiv. 2017. p. 101279. doi:10.1101/101279

13. prefixcommons.org [Internet]. [cited 19 Jan 2017]. Available: http://prefixcommons.org

14. Smith AM, Katz DS, Niemeyer KE, FORCE11 Software Citation Working Group. Software citation principles [Internet]. 2016. doi:10.7717/peerj-cs.86

15. dkernohan. What constitutes “research data”? What is “citation”? | Jisc Research Data Metrics [Internet]. [cited 19 Jan 2017]. Available: https://rdmetrics.jiscinvolve.org/wp/2016/04/12/what-constitutes-research-data-what-is-citation/

16. Journal Article Tag Suite [Internet]. [cited 19 Jan 2017]. Available: https://jats.nlm.nih.gov/

17. Mietchen D, McEntyre J, Beck J, Maloney C, Force11 Data Citation Implementation Group. Adapting JATS to support data citation. National Center for Biotechnology Information (US); 2015.

18. Tudor Groza CM. Jannotator [Internet]. 1 Jan 2017 [cited 19 Jan 2017]. Available: http://jannotator.monarchinitiative.org/#/

19. Attribution Working Group. In: FORCE11 [Internet]. 11 Jan 2015 [cited 19 Jan 2017]. Available: https://www.force11.org/group/attributionwg

20. OpenRIF. In: GitHub [Internet]. [cited 19 Jan 2017]. Available: https://github.com/openrif

21. Allen L, Scott J, Brand A, Hlava M, Altman M. Publishing: Credit where credit is due. Nature. 2014;508: 312–313.

22. Vrandečić D. Wikidata: A New Platform for Collaborative Data Collection. Proceedings of the 21st International Conference on World Wide Web. New York, NY, USA: ACM; 2012. pp. 1063–1064.

23. Wg OT. Citation [Internet]. [cited 19 Jan 2017]. Available: http://www.obofoundry.org/docs/Citation.html

24. Transitive Credit as a Means to Address Social and Technological Concerns Stemming from Citation and Attribution of Digital Products. Journal of Open Research Software. 2014;2: e20.

25. mhaendel. Envisioning a world where everyone helps solve disease [Internet]. [cited 19 Jan 2017]. Available: http://www.slideshare.net/mhaendel/envisioning-a-world-where-everyone-helps-solve-disease

26. Nucleic Acids Research | Oxford Academic [Internet]. [cited 19 Jan 2017]. Available: https://academic.oup.com/nar

27. Gaudet P, Bairoch A, Field D, Sansone S-A, Taylor C, Attwood TK, et al. Towards BioDBcore: a community-defined information specification for biological databases. Database. Oxford University Press; 2011;2011. doi:10.1093/database/baq027

28. Pampel H, Vierkant P, Scholze F, Bertelmann R, Kindling M, Klump J, et al. Making research data repositories visible: the re3data.org Registry. PLoS One. 2013;8: e78080.

29. Juty N, Le Novère N, Laibe C. Identifiers.org and MIRIAM Registry: community resources to provide persistent identification. Nucleic Acids Res. 2012;40: D580–6.

30. RRID | Searching in Tools [Internet]. [cited 19 Jan 2017]. Available: https://scicrunch.org/resources/Tools/search

31. Copyleft and data: databases as poor subject [Internet]. [cited 19 Jan 2017]. Available: http://lu.is/blog/2016/09/14/copyleft-and-data-databases-as-poor-subject/

Submission Date

01/19/2017

Submitter Name

Mark Gerstein

Name of Organization

Yale University

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

Genomics

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

4. Any other relevant issues respondents recognize as important for NIH to consider

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

b. Inclusion of a link to the data/software resource with the citation in the report

c. Identification of the authors of the Data/Software products

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

4. Any other relevant issues respondents recognize as important for NIH to consider

Additional Comments

response-to-data-sharing-rfi.pdf (28 KB)

I am very interested in data sharing in this RFI (http://osp.od.nih.gov/content/nih-request-information-strategies-nih-data-management-sharing-and-citation). In the past I have written a number of pieces about the subject and below I summarize some of the main points and list relevant references.

My Response

(1) The dynamic between databases and journals and between traditional reading and other forms of access should be considered (Reference Collection #1).
(2) There is a substantial cost in maintaining large data sets, both in terms of keeping up internet infrastructure (ie security) and the exponential scaling of data size and compute needs (ref. #2).
(3) The current journal publishing system should be updated to allow for computer parsing of papers and machine readable standards and to make the journal article more like a "mineable dataset" (ref #3).
(4))) Sharing private, patient data is problematic; solutions may lie in the framework of a central NIH sponsored resource and in specialized data standards (ref #4).

Reference Collection #1

E-publishing on the Web: promises, pitfalls, and payoffs for bioinformatics.
M Gerstein (1999). Bioinformatics 15: 429-31.

Annotation of the human genome.
M Gerstein (2000). Science 288: 1590.

Blurring the boundaries between scientific 'papers' and biological databases
M Gerstein, J Junker (2002). Nature Yearbook of Science and Technology 210-212 (ed. D Butler, Palgrave Macmillan Publishers)

An analysis of the present system of scientific publishing: what's wrong and where to go from here
D Greenbaum, J Lim, M Gerstein (2003). Interdiscip Sci Rev 28:293-302

The Death of the Scientific Paper
Seringhaus M, Gerstein M (2006). The Scientist. 20(9): 25

Open access: taking full advantage of the content.
PE Bourne, JL Fink, M Gerstein (2008). PLoS Comput Biol 4: e1000037.

Reproducible Research: Addressing the need for data and code sharing in computational science

Yale Law School Roundtable on Data and Code Sharing (2010). Computing in Science & Engineering 12(5): 8-13 (Sept/Oct).

Reference Collection #2

Computer security in academia-a potential roadblock to distributed annotation of the human genome.
D Greenbaum, SM Douglas, A Smith, J Lim, M Fischer, M Schultz, M Gerstein (2004). Nat Biotechnol 22: 771-2.

Impediments to database interoperation: legal issues and security concerns.
D Greenbaum, A Smith, M Gerstein (2005). Nucleic Acids Res 33: D3-4.

Network security and data integrity in academia: an assessment and a proposal for large-scale archiving.
A Smith, D Greenbaum, SM Douglas, M Long, M Gerstein (2005). Genome Biol 6: 119.

The real cost of sequencing: scaling computation to keep pace with data generation.
P Muir, S Li, S Lou, D Wang, DJ Spakowicz, L Salichos, J Zhang, GM Weinstock, F Isaacs, J Rozowsky, M Gerstein (2016). Genome Biol 17: 53.
(http://papers.gersteinlab.org/papers/costseq2)

Reference Collection #3

Structured digital abstract makes text mining easy.
M Gerstein, M Seringhaus, S Fields (2007). Nature 447: 142.

Structured digital tables on the Semantic Web: toward a structured digital literature.
KH Cheung, M Samwald, RK Auerbach, MB Gerstein (2010). Mol Syst Biol 6: 403.

Manually structured digital abstracts: a scaffold for automatic text mining.
M Seringhaus, M Gerstein (2008). FEBS Lett 582: 1170.

Seeking a new biology through text mining.
A Rzhetsky, M Seringhaus, M Gerstein (2008). Cell 134: 9-13.

Getting started in text mining: part two.
A Rzhetsky, M Seringhaus, MB Gerstein (2009). PLoS Comput Biol 5: e1000411.

Reference Collection #4

Genomics and Privacy: Implications of the New Reality of Closed Data for the Field
D Greenbaum, A Sboner, X J Mu, M Gerstein (2011). PLoS Comput Biol 7: e1002278

The role of cloud computing in managing the deluge of potentially private genetic data.
D Greenbaum, M Gerstein (2011). Am J Bioeth 11: 39-41.

Proceed with Caution
D Greenbaum, M Gerstein (2013). The Scientist 27:26 (1 Oct.)

Submission Date

01/19/2017

Submitter Name

ASCO

Name of Organization

American Society of Clinical Oncology

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

Oncology

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

4. Any other relevant issues respondents recognize as important for NIH to consider

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

b. Inclusion of a link to the data/software resource with the citation in the report

c. Identification of the authors of the Data/Software products

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

4. Any other relevant issues respondents recognize as important for NIH to consider

Additional Comments

FINAL ASCO NIHRFI LtrDataSharing 1-19-17.pdf (520 KB)

ASCO®
American Society of Clinical Oncology

PRESIDENT
Daniel F. Hayes, MD, FACP, FASCO

IMMEDIATE PAST PRESIDENT
Julie M. Vose, MD, MBA, FASCO

PRESIDENT-ELECT
Bruce E. Johnson, MD, FASCO

TREASURER
Craig R. Nichols, MD, FACP

DIRECTORS
Peter C. Adamson, MD
Charles D. Blanke, MD, FACP, FASCO
Linda D. Bosserman, MD, FACP
Walter J. Curran, Jr., MD, FACR
Stephen B. Edge, MD, FACS, FASCO
Paulo M. G. Hoff, MD, PhD, FACP
Arti Hurria, MD
Maha H. A. Hussain, MD, FACP, FASCO
David Khayat, MD, PhD, FASCO
Neal J. Meropol, MD, FASCO
Therese M. Mulvey, MD, FASCO
J. Chris Nunnink, MD, FASCO
Jaap Verweij, MD, PhD
Jedd D. Wolchok, MD, PhD

EX-OFFICIO MEMBERS
Clifford A. Hudis, MD, FACP, FASCO
ASCO Chief Executive Officer
Thomas G. Roberts, Jr., MD
Chair, Conquer Cancer Foundation Board of Directors

Via Electronic Submission

January 19, 2017

Francis S. Collins, MD, PhD
Director
National Institutes of Health
9000 Rockville Pike
Bethesda, Maryland 20892

Subject: NIH Request for Information (RFI): Strategies for NIH Data Management, Sharing, and Citation (Notice Number: NOT-OD-17-015)

Dear Dr. Collins:

The American Society of Clinical Oncology (ASCO) appreciates the opportunity to provide comments and information to help inform the National Institutes of Health's (NIH's) plans for data management and data sharing strategies. ASCO represents more than 41,000 physicians and other professionals who care for people with cancer and conduct research to improve cancer treatment.

General Comments

ASCO strongly supports the NIH efforts to develop a framework and strategies for data sharing and data management, which are essential for expediting the translation of research results and the dissemination of best practices to improve public health. ASCO agrees with NIH's principle that "the results of federally-funded scientific research are made available to and are useful for the general public, industry, and the community".[1] Further, effective data sharing relies upon appropriate identification, adoption, and crediting of good data management and sharing practices, and the principles to make data "FAIR" (Findable, Accessible, Interoperable, and Reusable). [2]

In addition to these principles, we note two additional factors we strongly urge the NIH to consider in the development of any data management and sharing strategies: 1) additional funding is critical for effective accessibility and utilization of data to occur and

[1] https://grants.nih.gov/policy/sharing.htm
[2] http://www.nature.com/articles/sdata201618

2318 Mill Road, Suite 800
Alexandria, VA 22314
T: 571-483-1300
F: 571-366-9530
www.asco.org

2) NIH should reconcile various data sharing policies that have emerged from several offices within the agency, and harmonize its policy (to the greatest extent possible) with other requirements researchers face. These requirements include results reporting on ClinicalTrials.gov and journal requirements for data submission as well as emerging journal requirements for data sharing. These points will also be detailed below in our response to the specifics within this Request for Information (RFI).
Our comments below are provided in the context of goals and strategies we have considered in the development of two ASCO data sharing platforms, ASCO's rapid learning system, CancerLinQ, LLC (CLQ), and the Targeted Agent and Profiling Utilization Registry (TAPUR), a prospective clinical trial. Both initiatives have highlighted the importance of common data elements and structured data collection to enable searching for and sharing information and integrating data across datasets.

The development of our rapid learning health care system, CLQ, will allow clinicians to analyze aggregated, real-world cancer clinical data from electronic health records (EHR). Clinical trials like ASCO's TAPUR Study include multiple therapeutic options and operate across all cancer types. The TAPUR Study leverages rapid advances in tumor genomic sequencing and broadening availability of such testing to facilitate participation by cancer patients with all cancer types. The goals of these initiatives are to drive better cancer diagnosis and treatment.

Information Requested

Section 1. Data Sharing Strategy Development

In this section the NIH notes that many factors must be considered when determining what, when, and how data should be managed and shared. These factors include, for example, the purpose for sharing, supporting data re-use and reproducibility, maturity of the science, the uniqueness of the data, and ethical considerations.

The CLQ and TAPUR platforms are examples of the valuable types of data to be shared because they represent the experiences of the diverse patient populations treated in real world clinical practice. The value in sharing such data is the ability to learn from every patient, with the expectation that this will accelerate progress against cancer and provide patients and physicians more comprehensive information to guide decisions about cancer prognosis and treatment. The sharing of data in a timely manner, once completed studies are reviewed and accepted for publication, provides researchers and clinicians the ability to deliver better treatments to the right patients at the right time.

General policies and aspects of ASCO's data sharing platforms include:

- Providing diagnostic and clinical data (i.e. patient demographics, diagnostic tests, treatments administered, clinical outcomes, adverse events, and patient-reported outcomes) to enable translation of basic science findings into clinical outcomes. This is particularly important in the era of precision medicine.

- Standardizing the data elements to ensure they can be used in a broad range of applications.

- Offering thorough documentation and guidance to ensure that investigators not only understand the data but also the initial purpose for its collection, data sources, validation processes, methods and tools that were used for collection and analysis, and recommended analytic methods to employ.

- Fostering a process that encourages and enables subsequent researchers who use the data to share with the original researcher information about data curation, data mapping, and methods that were used to work with the data to ensure that others can learn from any new methods or enhancements.

- Requiring good data stewardship, including a commitment to privacy, strong data security and transparency about the uses of the data.

In addition to these principles, there are a few complex aspects that can cause barriers to data sharing including the need for standardization of data. Below, ASCO has previously shared a few of these key provisions with the agency in 2015, in response to its *NIH Request for Public Comments on the Draft NIH Policy on Dissemination of NIH-Funded Clinical Trial Information* (NOT-OD-15-019)[3]:

1. Strongly recommend NIH to take the lead and work with stakeholders in the community to develop standardized common data elements and encourage researchers to share data in the agreed-upon format. ASCO is particularly supportive of requiring more standardized common data elements in electronic health records and other databases because this will make it easier for users of the data to search for information and integrate it with other datasets. Databases are only valuable to the extent that they are high-quality, well organized, and are standardized across entries.

2. In clinical trials, strongly recommend NIH follow the FDA standard requiring information about the attribution of adverse events to promote consistency in reporting for IND studies, and also use this standard for non-IND studies. The attribution of adverse events is an example of the broad category of data annotation which is essential for effective data sharing. Cancer is a complex and deadly disease with disease-related adverse events. In addition, trials often combine the investigational agent with standard of care treatments. Both of these facts make it difficult, at times, to determine whether an adverse event is caused by the investigational agent, the standard of care agent(s), or the underlying cancer. Thus, it is important that the information on adverse events accurately reflect what information the researchers are reporting (i.e., whether the record includes all adverse events or only adverse events that the researchers believe are attributable to the study drug or intervention).

3. Strongly recommend NIH to consider in the development of its strategies that it is impractical to ask researchers to comply with different data sharing rules, thus guidance is needed on how to coordinate the various approaches regardless of sponsor.

---

[3] https://grants.nih.gov/grants/guide/notice-files/NOT-OD-15-019.html

4. Strongly recommend NIH to allow researchers to request funding as an allowable budget item in NIH grants to support resources required for ClinicalTrials.gov reporting requirements; maintaining, organizing, and cleaning of data sets and to provide access to researchers who request the data; and the acquiring and analysis of data sets needed for research. This would ease the implementation of the NIH policy by providing researchers with the resources to meet these requirements. It would also align the new policy with the 2015 Institute of Medicine report, *Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risks*, which recommended that sponsors and funders "provide funding to investigators for sharing of clinical trial data as a line item in grants and contracts." [4]

Section 2. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports (RPPR) and Grant Applications

NIH grantees are required to report "other products of the research" in their Research Progress Reports (RPPR) and grant applications including data, databases, and software, in section C5a of their annual RPPR submission. ASCO agrees that more thorough reporting of data and software products in the RPPR and in Competitive Grant Renewal applications may strengthen documentation of productivity and may also identify projects and investigators who most effectively share data and software.

The NIH also requests in this section additional routes by which NIH might strengthen and incentivize data and software sharing beyond the benefits that result from reporting in RPPRs and Competitive Grant Renewals applications. As noted above, incentives should include additional resources to enable effective data and software sharing. This includes not only the preparation of data to be shared, but also the hosting of data and the associated long-term resource implications. We strongly urge the NIH to identify strategies to improve researchers' compliance with other existing and developing rules governing data sharing. Guidance is needed on how to coordinate various approaches and harmonize policies on data sharing that would otherwise overlap for stakeholders who conduct research projects that must adhere to requirements for ClinicalTrials.gov, various institutes and centers within the NIH (i.e. NIH Data Sharing Policy, NIH Public Access Policy, and the NIH Genomic Data Sharing Policy), trial sponsors, the U.S. Food and Drug Administration, and journal editors.

As another example of overlapping developing policies within the area of data sharing, the recent NIH RFI *Including Preprints and Interim Research Products in NIH Applications and Reports* (NOT-OD-17-006)[5] sought input on whether preprints and other interim research products should be included in NIH applications and reports, and how investigators could report them. In particular, we believe that studies reporting clinical data that have a potential impact on the lives of patients should continue to be shared according to the current standards used within research. Currently, authors are allowed to submit and present abstracts (i.e. oral or poster presentations) of their research in open, scientific meetings. These are considered preliminary data that are not generally viewed by physicians as sufficiently mature to

---

4 https://www.nap.edu/catalog/18998/sharing-clinical-trial-data-maximizing-benefits-minimizing-risk

5 https://grants.nih.gov/grants/guide/notice-files/NOT-OD-17-006.html

justify a change in clinical practice, but extremely important for continuing education and knowledge on the advancements in science. Further, the current standard is a proven approach for the dissemination of preliminary results, for establishing research priorities, to provide evidence of independence and productivity, to gain feedback from peers, correct errors, and ensure the protection of patient privacy.

In conclusion, ASCO urges the NIH to develop a comprehensive data management and sharing framework that reflects the principles of responsible sharing of data in a timely way and that maximizes the benefits to research, care, and the general public. A fundamental principle of all NIH-funded research is that the results must be disseminated in order to contribute to the general body of scientific knowledge and, ultimately, to the public health. We agree and believe NIH awardees should continue to be expected to make their data and accomplishments of their activities available to the research community and to the public at large for the greater good. We encourage the NIH to continue working towards a broad and coordinated framework that will enable researchers and clinicians to learn from and extend the work of others, to readily share those insights, and to translate research results into clinical advances for patients.

We look forward to working with the NIH toward these important goals.

Thank you for the opportunity to provide comments on this RFI on Strategies for NIH Data Management, Sharing, and Citation. Please contact Shimere Williams Sherwood (Shimere.Sherwood@asco.org) at ASCO with any further questions.

Sincerely,

Daniel F. Hayes, MD, FASCO, FACP
President, American Society of Clinical Oncology

Submission Date

01/19/2017

Submitter Name

Veronique Kiermer

Name of Organization

PLOS

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

Publishing research across the spectrum of sciences with a focus on biomedical research

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

4. Any other relevant issues respondents recognize as important for NIH to consider

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

b. Inclusion of a link to the data/software resource with the citation in the report

c. Identification of the authors of the Data/Software products

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

4. Any other relevant issues respondents recognize as important for NIH to consider

Additional Comments

NIHRFI-DataManagementandCitation_PLOS.pdf (91 KB)

Response to NIH RFI -- Strategies for NIH Data Management, Sharing, and Citation

We write to express the views of the Public Library of Science, a fully Open Access Publisher of seven Research Journals, in response to your RFI on Data Sharing, Management, and Citation. Open access to Research Articles is just the first step in what we consider should be the end state for all publicly funded research, and we support broader efforts towards open science. We are developing our own policies to help establish a new norm in which upon publication of a journal article, if not before, all of the underlying data (where ethically appropriate) is openly available to access and reuse without restriction according to the FAIR principles for data management to make data Findable, Accessible, Interoperable and Re--usable[1].

Since our first journal, *PLOS Biology*, launched in 2003 all PLOS journals had a data policy that expected researchers to make available their data upon reasonable request. Even though many publishers have a very similar policy, there is strong evidence that data availability declines over time once a study has been published[2]. Such a 'good faith' style policy was not sufficient to ensure data availability and therefore, in March 2014, PLOS updated its data policy[3]. Under the new policy, every article now includes an explicit Data Availability Statement that details where the underlying data are located or provides information about any exceptions (where sharing is ethically not possible, or if the data are owned by a third party). Since the inception of this policy, more than 66,000 articles have been published in PLOS journals with an associated data availability statement. We have seen clear increases in data availability, and we have evidence of greater engagement from our communities of academic editors and reviewers in assessing compliance with the policy during peer review. In 2016 alone, about 4,000 datasets associated with PLOS articles were deposited in open repositories. To continually improve on our implementation of this policy, PLOS has assembled a Data Advisory Group of practicing researchers across many disciplines who we consult regularly.

PLOS' experience around the implementation of our 2014 open data policy has also highlighted some barriers that exist to data sharing, even when researchers are willing to comply. The barriers are multi--faceted, some are by their nature technological, some are policy--based, and others are cultural. It is clear that ongoing efforts are required to overcome these barriers and move the scientific community towards a standard practice of data sharing as the norm.

To this end, PLOS is already involved in several initiatives that aim to bring together relevant stakeholders and define standards and standard practice around data sharing, management,

---

[1] http://www.nature.com/articles/sdata201618

[2] Timothy H. Vines et al., (2014) The Availability of Research Data Declines Rapidly with Article Age, *Current Biology* 24 (1). 10.1016/j.cub.2013.11.014

[3] http://journals.plos.org/plosone/s/data--availability

and citation e.g. FORCE11[4], THOR[5]. However, our experience with implementation of our data policy has also made us aware that some of the current barriers do not have obvious solutions, especially for clinical and translational research, areas where data transparency is of the utmost importance to ensure integrity, reproducibility and potentially speed of research. All research disciplines benefit from data sharing, and where obstacles persist, we stand ready to work with the community of researchers and their institutions to help bring down these obstacles. We recognize and appreciate the critical leverage of the NIH to facilitate further progress.

Our experience has shown that data sharing on a large scale is possible, while still a work in progress. We hope that sharing some of our experience will usefully inform the development of the NIH policy. We believe the NIH has a unique opportunity to influence a cultural change in data sharing to benefit the research community and society at large.

Because of our experience with the PLOS data policy, we focus our comments on data sharing and citation, however, most considerations will also directly apply to code and software.
We will discuss the relevant technology and policy issues and then the cultural barriers that we have encountered.

Our key comments for the NIH to consider are:

1. A major impediment to data sharing remains cultural: the lack of academic credit for sharing data. Therefore, we applaud the efforts of the NIH to encourage thorough data citation in progress reports and grant applications, because of the tremendous impact this new norm could have on research assessment and academic credit. We urge that these encouragements be accompanied by operational mechanisms to ensure that this information is *actively* considered by reviewer panels during the evaluation process.
2. Important features of technical guidance for data citations, to facilitate academic credit, include:
    a. Persistent unique identifiers, such as DOIs, for datasets;;
    b. Usage, and identification in the citation, of repositories that support the FAIR principles for data management;;
    c. Data citation formats that identify authors and are consistent with publishing formats (see the work of the FORCE 11 DCIP[6] group);;
    d. In addition, allowing the citation of preprints can help provide additional context to cited data.
3. Achieving such quality and prominence of data citation in progress reports demands a data management and sharing strategy emphasizing the use of repositories that issue persistent identifiers and facilitate the attribution of credit to data providers.
4. Agreements around elements that constitute pillars of effective data management and sharing strategy are still needed:

[4] https://www.force11.org
[5] https://project--thor.eu/
[6] https://www.force11.org/group/dcip

   a. Community--endorsed metadata standards that are discipline-- and data type--specific, to maximize the capacity for data reuse and minimize risks of misinterpretation;; support for researchers willing to invest time in these standards is critical.
   b. Guidelines about the length of time that data managed outside of repositories needs to be maintained;; a rule of thumb agreed between funders, journals and research institutions will help planning of management efforts.
   c. Guidance and best practices for mechanisms for effective, safe and responsible data sharing of clinical data and other types of sensitive data;; structures like independent data access committees are required and while their centralization is impractical and costly, guidance for their consistent deployment at the institutional and departmental level would provide a way forward.
5. We caution against the use of embargoes for data sharing intended to allow the pursuit of further academic credit. Instead, we urge more efforts to provide academic credit for data and data--related contributions, to ease the pressures on researchers that may otherwise undermine data sharing efforts.
6. When prioritization is needed, we reason that whilst the expected volume of reuse may influence the amount of data curation required and length of time a dataset is maintained, a single reuse, such as a validation by peers or inclusion in a meta--analysis, can have enormous benefits for the progress of research. Therefore, minimal standards for sharing all datasets are important.

I. Technology and policy questions pertaining to data sharing and strategy development.

There are costs and efforts needed to move to a state where data sharing and citation are the norm. They relate to infrastructure, knowledge and incentives.

1. Prioritization

Where prioritization is a factor for the NIH, we recommend that the inherent purpose of data sharing be actively considered. If the purpose is to maximize reuse of data[7], the highest standards of data curation and description, as well as appropriate licensing (we recommend CC0) must be applied. These high standards require more regulation, and more efforts for data providers, hosts and curators.

In some cases, the purpose of data sharing is focused on transparency and reproducibility[8], and other reuse is unlikely. Data can then, in theory, be shared with less regulation and cost, but a minimal level of metadata is still required (see below).

[7] http://bjoern.brembs.net/2015/11/dont--be--afraid--of--open--data/

[8] http://www.nature.com/news/1--500--scientists--lift--the--lid--on--reproducibility--1.19970

In both use cases, data persistence and having a means to unambiguously identify a dataset (via a persistent identifier) remain critical features, best achieved via data repositories.

2. Metadata

In our experience, 'Data' means different things to people both within and between fields. Correct interpretation of data requires that high quality metadata is also made available. Anecdotally, the risk of data being misinterpreted by others is a concern cited by researchers when discussing the PLOS data requirements.

a) Metadata Standards;; In our experience, metadata standards are better established by the community----the MIAME standards are the quintessential example of such a grass--root effort[9].

In the absence of such standards at the dataset level, published articles serve as a complementary description that provides context. Therefore, reporting standards for publication, such as those maintained by the EQUATOR network[10] play an equally important role to metadata standards.

Robust linking between uniquely identified publications and datasets is therefore crucial, and efforts like those of the European commission funded THOR[11] and OpenAire[12] projects are aimed at establishing an infrastructure for assigning persistent identifiers and cross--linking them.

Finally, given the plethora of data types and discipline--specific needs, we have also experienced the usefulness of community discussions about what data must be shared to effectively support a publication ---- for example, a group of academic volunteers coordinated by the PLOS Genetics Editors have established useful guidelines for genetics[13].

The lack of agreed--upon metadata standards remains a barrier to data sharing. Funding agencies such as the NIH are well placed to support the coordination and financing of community efforts towards metadata standards, as well as rewarding participants with academic credit.

b) Data Repositories;; Data repositories can play a critical role in collecting and curating metadata. Structured specialized repositories such as those maintained by NCBI and EBI, provide a critical service in supporting reuse of data. Umbrella organizations can also help

[9] https://www.ncbi.nlm.nih.gov/pubmed/11726920
[10] http://www.equator--network.org/
[11] https://project--thor.eu/
[12] https://www.openaire.eu/
[13] Gregory S. Barsh et al., (2015) *PLOS Genetics* Data Sharing Policy: In Pursuit of Functional Utility. PLOS Genetics 11(12): e1005716. doi: 10.1371/journal.pgen.1005716

smaller repositories with metadata standards e.g. the NSF--funded DataONE[14] project. In disciplines where structured repositories do not yet exist, unstructured repositories (Dryad[15], Zenodo[16], Open Science Framework[17]), do not enforce metadata standards, but they are useful to allow data deposition and persistence.

At a minimum, it is important for all repositories to facilitate the reference to and citation of datasets in order to facilitate credit for data sharing (see section on cultural issues). For example, many repositories do not require author/generator information, which is essential for data citation.

PLOS does not dictate repository selection;; authors are encouraged to select the repository most appropriate for their research. However, we recommend a list of vetted repositories[18]. Minimum requirements are necessary for the repository itself with respect to licensing (it should not be more restrictive than CC--BY), openness (for access and deposition), longevity, use of persistent identifiers, and costs for deposition).

3. Persistence

There is no consensus on the length of time that data should be available for. Repositories that issue DOIs (or equivalent persistent identifiers) are the most effective way to ensure long--term persistence. Author stewardship of data is not sufficient[19]. However, as deposition in repositories is not yet the norm, guidelines are needed to determine how long a researcher should provide access to, and/or 'user support' for his/her data.

In different contexts, periods of 3 to 10 years have been floated as reasonable timeframes. We receive requests for data (related to papers that we published before our updated Data Availability policy was in place), several years after publication. We suggest that 3 years is probably too short for the purposes of assessing replication and re--analysis, while a 10--year commitment could raise issues of data formats and readability, which may impose a further burden on data producers or repositories. Whilst discipline--specificity, data types and intended purposes will call for different lengths of time, we would welcome an agreement for funders, repositories, and publishers, to provide consistent direction to researchers.

4. Sensitive Data

We wish to raise the difficult challenges observed with sharing of several kinds of sensitive data that are unsuitable for public release. A means to provide effective, safe and responsible data sharing of such sensitive data will require investment in infrastructures, education and resources. Such data includes, but may not be limited to:

---

[14] https://www.dataone.org/
[15] http://datadryad.org/
[16] https://zenodo.org/
[17] https://osf.io/
[18] http://journals.plos.org/plosone/s/data--availability#loc--recommended--repositories
[19] Timothy H. Vines et al., (2014) The Availability of Research Data Declines Rapidly with Article Age, Current Biology 24 (1). 10.1016/j.cub.2013.11.014

a) **Ethical or legally restricted data**;; in particular, data from clinical trials and other studies involving patients and research subjects.
b) **Data deposition might present some threat**;; for example, locations of fossil deposits and endangered species.

We see a need for best practices and guidance for the establishment and operation of data access committees to oversee and mediate requests for access to the data. Independent data access committees can be effective[20] and could be deployed for example at the institutional level. We also note a need for clarification and best practices for informed consent to avoid limiting future sharing of data. Although there are existing guidelines about how to anonymize clinical data for sharing, this is an area that would also benefit from a broadly disseminated directive.

II. Important features of technical guidance for data citations in NIH progress reports and grant applications.

Data citation in grant applications and progress reports is not only critical to ensure compliance with a data management and sharing strategy, it also provides an essential means of assigning proper academic credit for sharing data. Taking credit and the above considerations into account, we suggest that important features of technical guidance include:

b) Use of persistent unique identifiers, such as DOIs, for datasets;;
c) Encourage the usage, and identification in the citation, of repositories that support the FAIR principles for data management;;
d) Use data citation formats that identify authors and, for consistency, are consistent with formats considered by the publishing industry (see the work of the FORCE 11 DCIP group);;
e) In addition, allowing the citation of preprints can help provide additional context to data cited in this way.

III. Cultural questions and the inclusion of data citation in NIH performance progress reports and grant applications.

As discussed above, there are emerging services, standards and best practices that in time can overcome the technical and infrastructure barriers. The lack of academic credit for data sharing, however, remains the major cultural impediment. The larger challenge, therefore, is to transform the existing culture into a culture in which data sharing, transparent reporting and good data stewardship are given at least as much prominence and status as journal publications.

By encouraging data citations in grant applications and progress reports, the NIH is deploying a

[20] Krumholz et al, N Engl J Med 2016;; 375:403--405, 2016, http://dx.doi.org/10.1056/NEJMp1607342

mechanism for data--focused academic credit, which we applaud. However, without tangible benefits to individual researchers, through funding and career opportunities, there will remain little incentive to make use of this mechanism.

Therefore, we suggest consideration of four key points:

1. **Operationally increase the value of data in research assessment**
   Encouraging data citations in progress reports is a clear signal that the value placed on datasets *per se* is the same as that for journal publications. However, for this policy to take root, it must penetrate deep into the workings of reviewer panels. Reviewers should pay as much attention to the data outputs of researchers as they do to their publication record. We therefore urge that the encouragement of data citations be accompanied by mechanisms that ensure that this information is *actively* considered by reviewers. Overcoming cultural barriers will also require the community to move beyond old habits and the reliance on journal titles to evaluate research outputs.

   The same habits need to be overhauled in tenure and promotion committees within universities and research institutions. Funding agencies like the NIH have critical leverage in this cultural change, leading by example, establishing new incentives and changing individual behavior in all contexts of research assessment.

2. **Establish expectations about data sharing in grant applications**
   For data to be shared effectively, it must be collected, analyzed, transformed and stored with the intention to share. Data sharing must be an integral part of the research cycle, not an afterthought at publication or in progress reports. NIH's implementation of data citation in progress reports should intimately connected to the provision of data management plans in grant applications.

3. **Embargoes and credit**
   In different contexts, post--publication embargoes have been suggested to allow data producers more time to pursue further academic credit (in the form of additional publications). The PLOS data policy requires the data underpinning the conclusions of an article to be released at the time of publication----based on consideration of reproducibility, public trust, and rapid research progress. The vast majority of PLOS authors voluntarily comply, but in some communities, concerns have been expressed about the potential impact on future publications. As long as the only form of academic credit is a first--author article publication, the pressure to publish will continue to undermine efforts to facilitate data sharing. We argue that providing academic credit for data producers and curators is the best way to counter such pressure on researchers.

   Releasing data *before* publication can of course be beneficial also. In fact, in public health emergencies, it is essential and can help save lives[21]. In 2015, several leading journals

[21] Nathan L. Yozwiak, Stephen F. Schaffner & Pardis C. Sabeti (2015) Data sharing: Make outbreak research open access. *Nature* 518, 477--479 doi:10.1038/518477a

issued a statement to emphasize that they encourage and early release of data in public health emergencies[22]. In such circumstances, context for the data is also particularly important, and a preprint of a planned research publication may serve this purpose. To this end, since the beginning of the zika outbreak, PLOS encouraged all authors of relevant manuscripts to deposit their manuscripts in bioRxiv at the time of submission, while peer review was ongoing. Similar mechanisms would apply to the citation of data and preprints in grant applications and progress reports.

4. Policies and best practices alignment

Compatibility and consistency of policies and technical guidance between funders, publishers and repositories can help alleviate the burden on authors of data sharing and citation and facilitate adoption. Here we mention a few initiatives in development among publishers and repositories that are relevant to the NIH consideration:

Publishing initiatives and responsibilities

a) **Data policies:** The PLOS data policy implementation has demonstrated that publishers can influence the amount of data being made available to other researchers as part of the publication process. Others have effectively implemented their own guidance (e.g., the Nature data policy[23]). Publishers therefore can reinforce standards set by funding agencies.

b) **Data citations:** By facilitating citations to data in consistent formats, journals can reinforce the equal status of publications and datasets, and make the impact of reuse more transparent. Efforts are ongoing in the publishing community to standardize data citations. We therefore encourage the NIH to take advantage of existing efforts such as FORCE11 Joint Declaration of Data Citation Principles[24] and the ongoing DCIP group[25].

c) **Granularity of author contributions**: Current authorship conventions for journal publications poorly capture the complexities of academic efforts and achievements. In addition to raising the profile of datasets as research outcomes *per se*, data--related contributions to research articles can also be highlighted. With initiatives like ORCID[26] and the CRediT taxonomy[27], publishers can now allow authors to express their data--related (and other) contributions in terms that are both human-- and machine--readable, making such contributions portable to indexers, CVs and professional profiles (see the PLOS implementation[28]). We recommend that NIH

[22] http://blogs.plos.org/plos/2016/02/statement--on--data--sharing--in--public--health--emergencies/

[23] http://www.nature.com/news/announcement--where--are--the--data--1.20541

[24] Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11;; 2014 [https://www.force11.org/group/joint--declaration--data--citation--principles--final].

[25] https://www.force11.org/group/dcip

[26] https://orcid.org/

[27] http://docs.casrai.org/CRediT

[28] http://blogs.plos.org/plos/2016/12/author--credit--2016--roundup/

develop mechanisms to make use of this information as part of their research assessment efforts.

Repositories initiatives and responsibilities

In addition to their roles in collecting and curating metadata, and supporting compliance with metadata standards, repositories have additional capabilities to support the ecosystem and facilitate a cultural shift in credit attribution ---- in particular, the provision of dataset--level certificates and metrics that demonstrate data quality or reuse. Many repositories, and umbrella organizations like DataCite[29] and RDA[30] are engaged in such initiatives. We note, however, that the development of data--level metrics is in its infancy and there is no consensus yet on the relevant metrics. For example, a single instance of reuse could have a powerful effect on the research community, and volume of reuse is not necessarily a good indicator.

In closing,

We want to thank the NIH for allowing our and other's comments to inform the development of their policy and guidance on the critical topics of data sharing and credit for data as a product of research. A strong lead from such a major funding agency has the potential for a global impact on how research outputs are communicated and evaluated.

Sincerely,

Emma Ganley, PLOS Data Program Lead and Chief Editor, PLOS Biology
Catriona McCallum, Advocacy Director, PLOS
Veronique Kiermer, Executive Editor, PLOS

[29] https://www.datacite.org/
[30] https://www.rd--alliance.org/

Submission Date

01/19/2017

Submitter Name

Margaret Levenstein

Name of Organization

Inter-university Consortium for Political and Social Research

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

social, political, economic, and environmental determinants of health

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

Please see attached submission.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

In perpetuity, as indicated in attached submission.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

Please see attached submission.

4. Any other relevant issues respondents recognize as important for NIH to consider

Please see attached submission.

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

Please see attached submission.

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

DOIs are extremely valuable for creating self-reinforcing incentives for data production, sharing, and citation. Please see attached submission.

b. Inclusion of a link to the data/software resource with the citation in the report

Please see attached submission.

c. Identification of the authors of the Data/Software products

Please see attached submission.

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

Please see attached submission.

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

Please see attached submission.

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

Please see attached submission.

4. Any other relevant issues respondents recognize as important for NIH to consider

Please see attached submission.

Additional Comments

ICPSR submission NIH Data Sharing RFI January 2017.docx (16 KB)

NIH Request for Information (RFI): Strategies for NIH Data Management, Sharing, and Citation

SECTION I: Data Sharing Strategy Development

While all data have potential value, and data sharing is fundamental to the principles of scientific replication and cumulative progress, so that publication-related should data be made available to allow for replication (not necessarily requiring significant investment in curation), some data have particular value that add significantly to the scientific infrastructure. Particularly important are large, complex data collection efforts that are expensive to create. These investments generate the greatest scientific return when these data are shared; other researchers can leverage prior investments in data collection rather than request funds for additional data collection effort. Datasets will have lasting impact on the scientific community, and should be prioritized for sharing in both the short- and long-term, when they were:

1) collected using a probability sampling frame (especially nationally representative samples)

2) from or about an understudied demographic group

3) collected at multiple time points (longitudinal or repeated cross-sections)

4) represent methodological innovations in either collection strategies or content

5) underly highly cited publications or researchers

6) represent key theoretical frameworks or are otherwise critical to a field's collective knowledge (Hoelter, Pienta, and Lyle, 2015).

Sharing these high-value data allows researchers to answer questions that would otherwise not be possible, to generalize findings to a broader population, and test causal models.

In such cases it is important to preserve access to the data over time and to assure that they are shared and curated to keep them viable despite technological change. This is best done by depositing data into an established repository. Sharing data with more ad hoc solutions, such as via one's personal Website may appear satisfactory in the short-term; in the longer run it can undermine data access if the original researcher changes institutions, the host of the site goes out of business, etc. Software updates and changing popularity of software packages – including differences in disciplinary preferences –mean that files created in one version of a package common today may not be readable by the software of choice in just a few years. To prevent this, data must be released in a variety of formats, and data must be

preserved in an accepted archival format such as SPSS portable files or comma-separated variables (CSV), and XML or PDF/A for documentation, allowing for new versions to be created when needed.

Lastly, saving and sharing data are not sufficient. For data to continue to have scientific value, they must be well documented and discoverable. Both of these features require curation. That is, metadata – data about the data – is critical to the successful reuse of any dataset. A researcher must be able to evaluate the quality of the data in terms of sample design and size, question order and wording, data completeness, data collection at multiple points in time if causality is to be inferred, and other characteristics; this is only possible if documentation have been created that fully describe these critical elements. Creating metadata for both the study and the variables within the study provides information for search engines and online catalogs. Great data that cannot be found by anyone outside the original research team will not and cannot contribute to future science.

Commonly cited barriers to data stewardship include cost (in both time and dollars) to prepare the data for sharing, fear of having one's project "scooped" by another researcher, privacy and confidentiality, and proprietary and commercial value. There are well-established methods for addressing each of these issues. In terms of preparation necessary to share data, software packages and platforms now exist that allow researchers to document this step in the research process as it occurs, so that once the data collection is complete, all relevant metadata are automatically created. For NIH policies on data sharing to be successful, researchers must be trained to use such software, so that the burden on PIs of engaging is data sharing is minimized. Some data collection software (e.g., BLAISE) can export variable-level metadata – including features such skip patterns – with the data file itself. Documenting decisions throughout the research process, rather than after publishing results, significantly reduces the burden and results in higher quality documentation. This requires education and guidance for researchers at the beginning of new studies. Pre-analysis plans and research registries allow investigators to "lay claim" to their ideas, establishing intellectual property rights to ideas, while reducing the scope for researchers to engage in specification search. Most importantly, the amount and range of data in most data collections is such that the original investigators will almost never be able to analyze fully the data collected, and researchers from different disciplinary perspectives may ask questions never envisioned by the original investigators. Finally, working with established archives can reduce concerns about protecting privacy and confidentiality. For example, the Inter-university Consortium for Political and Social Research (ICPSR) has a standard approach to reviewing data for potential disclosure issues and to address such issues when they arise. These solutions include aggregation or suppression of observations, "swapping" values, suppressing variables that alone or in combination increase the risk of re-identification, synthesizing data, or restricting access to approved researchers using a secure computing environment.

SECTION II: Inclusion of Data and Software Citation in NIH Research Performance Progress Reports (RPPR) and Grant Applications

Data citation in research output is critical to creating a scientific ecology in which researchers have an incentive to produce and share data. NIH's requiring such citations in RPPRs and grant applications reinforces the practice of data citation in research. But researchers need to be provided with the skills and resources necessary to share data effectively or these policies will become an additional burden, a checkoff on the "to do" list, rather than a valued part of the research process. Citation is important not only for showing impact on a researcher's career or on science as a whole; citing data makes it easier for linkages between data and published papers or articles to be made, and therefore for researchers to receive credit for their investment in data creation and their willingness to share that data. Depositing to established repositories that issue DOIs facilitates data citation. The linking of data and publications, already being done by ICPSR, Web of Science, and others, makes it possible to measure the breadth of work based on a single data source and to identify gaps that remain to be filled. Those new to research –

particularly undergraduate and graduate students – are aided greatly when data and publications are linked. They can use publication searches to point them toward important data in the areas of interest. When this information must be gathered manually (as is the case when data are not cited), there is a greater chance of a key data source going "unfound."

Data citation is critical for replication and transparency, especially when an author uses data that are available from multiple sources. Data repositories and other data disseminators should version their data and include that in the information they make available to users for citation purposes. The essential elements of a data citation include: 1) the name(s) of each individual or organization responsible for the creation of the datasets (author); 2) the year the dataset was disseminated (date of publication); 3) the complete title of the dataset, including edition or version number if applicable; 4) the organization responsible for archiving, producing, publishing, and/or distributing the dataset (publisher and/or distributor); and 5) a Web address or other unique identifier, preferably a persistent global identifier such as a DOI (electronic location or identifier) (e.g., IASSIST, 2012; Economic and Social Research Council, nd). Data providers can make citation easier by clearly providing these elements or the preferred citation. ICPSR, for example, provides a citation on every study homepage as well as in documentation files and data download packages. Policies such as requiring data citations in RPPR or grant applications are useful to inculcate researchers with the habit of citing data and the establishment of a culture and self-reinforcing ecology of data citation. They will be most effective when accompanied by training in how to collect and prepare data for sharing in an established data repository.

References:

Economic and Social Research Council (ESRC). ND. *Data Citation: What you need to know.*

Hoelter, Lynette F., Amy Pienta, and Jared Lyle. 2016. Data Preservation, Secondary Analysis, and Replication: Learning from Existing Data. Ch. 40 in *The SAGE Handbook of Survey Methodology*, Christof Wolf, Dominique Joye, Thomas W. Smith, and Yang-chih Fu (Eds.).

International Association for Social Science Information Services and Technology (IASSIST). 2012. *Quick Guide to Data Citation: Identify, retrieve, attribute*.

Submission Date

01/19/2017

Submitter Name

GQ Zhang

Name of Organization

University of Kentucky

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

Cancer, Neuroscience, Cardiovascular disease/stroke, Diabetes/obesity and Substance abuse

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

Physiological signals provide a fundamental source of phenotypic information related to major human organ systems such as the nervous, the cardiovascular, and the muscular-skeletal systems. The use of such signal data permeates almost all medical specialties, and is especially evident in sleep medicine and neurological diseases where biophysiological signals are routinely collected as part of diagnostic and therapeutic initiatives. Physiological signals comprise a unique data modality since they are acquired non-invasively or minimally invasively over time and provide temporally informative data. Physiological signals provide key diagnostic and prognostic information for a wide range of disorders - from myocardial infarction to sleep apnea. Furthermore, these data types are increasingly captured by smart sensors and mobile devices to provide day-to-day (if not minute-by-minute) monitoring data. Therefore, fully utilizing this type of rich data source represents a major opportunity for the scientific and healthcare community.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

In a large collaborative center setting, the length of time the clinical data should be made available for secondary research purposes should be no more than a year. The appropriate means for maintaining such data should leverage the latest data science and informatics advances. For example, for the center for SUDEP research project (csr.case.edu), the entire data capture to data warehousing and data sharing pipeline is created and maintained by a relatively small group of researhers, and data is available for secondary research purposes within a couple of weeks from the initial collection. We must continue innovate tools and technology to achieve highly effective means for maintaining and sustaining such data. The funding and compitational reseources needed for sustaining the data would be greatly overweighed by the research value they bring. Reusing data is less expensive an avenue than regenerating and recollecting data - the latter continues to be the de fecto mode of funded research.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

To understand the burden to the research community under the existing paradigm, consider a typical scenario of a sleep researcher who is interested in extracting sleep EEG spindles as an inherited phenotype predictive of cognitive performance. To identify sleep spindles, characterized as bursts of oscillatory brain activity visible on an EEG in non-REM (NREM) sleep, one must download the entire collection of PSG EDF files, extract the EEG channels, and then extract the specific NREM stage sleep signal fragments for further analysis. Even though stage NREM sleep data may only be a portion (2%) of the overall PSG signal data (e.g., over 500MB per sleep study), the entire PSG must be downloaded. If interest was in stage 2 sleep (when spindle activity is highest), there are even greater needs to improve the approach to extraction. When thousands of researchers are interested in analyzing sleep spindle and similar channel-specific, event-specific physiological signal phenotypes, each researcher must go through the same duplicated process of EDF file downloading, channel extraction, and event-specific signal segmentation. A similar inefficient paradigm exists for TCGA

and other genomic data. The main barrier is that a new paradigm overcoming these redundancies and inefficiencies requires NIH panels and program directors to appreciate and embrace the "disrruptive" data sharing and management paradigms that are being proposed by computer and data scientists. Continued innovation in data science and technology is the key.

4. Any other relevant issues respondents recognize as important for NIH to consider

Priority should be given to sharing data to encourage new discovery--ie, many studies come to completion without fully exploring the potential number of questions or apply newly developed or emerging techniques. This relates to identifying new prediction models, signatures of disease, etc. Data sharing also provides the ability to aggregate data to get greater power, diversity, and to conduct subgroup analyses for "precision medicine" and explore issues such as sex-specific effects. Data sharing also enhances transparency and supports training -attracting new people to the field.

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

This would be encouraged, but should not be the only measure to judge the progress and success of a project. Depending on specific domains, some data resource may have a longer life cycle, others may be shorter.

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

This is a secondary issue. The primary issue is the availability and usability of the data that are shared.

b. Inclusion of a link to the data/software resource with the citation in the report

This is a first step. The question of how long the link will need to be active and accessible, and who is going to pay to make the link staying active and with fully usable content, is another matter. We often find broken and password-protected links.

c. Identification of the authors of the Data/Software products

There should be publication venues dedicated for describing them. User manuals and technical manuals should also be made available, such as those found at the sleepdata.org site.

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

Depends on the types of data.

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

Not sure what this question means.

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

R24, R01, and U01 are all appropriate mechanisms.

4. Any other relevant issues respondents recognize as important for NIH to consider

Priority should be given to sharing data to encourage new discovery--ie, many studies come to completion without fully exploring the potential number of questions or apply newly developed or emerging techniques. This relates to identifying new prediction models, signatures of disease, etc. Data sharing also provides the ability to aggregate data to get greater power, diversity, and to conduct subgroup analyses for "precision medicine" and explore issues such as sex-

specific effects. Data sharing also enhances transparency and supports training -attracting new people to the field.

Additional Comments

Submission Date

01/19/2017

Submitter Name

Heidi Imker

Name of Organization

University of Illinois at Urbana-Champaign

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

Multidisciplinary

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data

At an agency level, thinking of data has having "localized" or "transformative" value may help prioritization. Localized referring to data underling publications in which inspection allows readers to judge the work's validity beyond what's available in a low-res article figure. e.g., inspecting chemical spectra for peak width, splitting, etc. Beyond inspection, sharing data may also allow for re-use. e.g., re-using a set of genomic sequences from a paper that included phylogenetic classification would be more efficient (no need to re-collect) and more reproducible (exactly same set as previously published). In these cases though, value is tied to original research and subsequent inspection/re-use would be closely related (generally), so impact is likely localized. Because of the pace and nature of science, data in these cases, when shared in an isolated way (e.g. not part of collection of like-data but rather in a supplementary fashion) are likely to have short-term value but *inconsistent* long-term value. "Transformative" refers to when data availability has potential beyond the initial research thread and in particular when sharing (*with QA/QC, standardization*) may enable new questions, analyses, and subsequent scientific knowledge. Genomic data is the poster child - had sequences always been shared as a supplement to a paper or even as a collection but without standardized metadata, formats, and tools then genomics wouldn't be where it is today. While both have value, pragmatically need to prioritize data with potential transformative value over localized for long-term commitment and resource-intensive applications.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

Time: Difficult to anticipate, so set a minimum and a caveat that data that continues to be of use past the minimum should be retained; particularly useful for "localized" value. Our campus data repository uses a 5 year minimum retention, with provisions for preservation review after 5. Aligns with HHS's ORI documentation (http://ori.hhs.gov/education/products/rcradmin/topics/data/tutorial_11.shtml) although would help if an authority body would clarify if data = record. We're taking this approach because allows us to positively manage our long-term data commitments (https://databank.illinois.edu/policies#preservation_review). Means: Scoping is a means. Need centralized, government-supported repos for "transformative" but can probably rely on universities or 3rd party repos for "localized." Unbounded growth isn't sustainable so a "do it all/keep it all" stance is nonsensical and is why focusing on maintaining/sustaining transformative necessary. Define criteria and expectations that can be rolled into processes and protocols, and then define and implement metrics so when arguments need to be made to invest more/less in certain resources then metrics support decisions. Implications: Long term preservation isn't free. Moreover, people just don't want preserved data, they want resources to use that data. Money (for both hardware and staffing) will have to come from somewhere. Much concern right now about lost data, but we would do better to embrace data triage as a core part of research data management, but that triage should be intentional and controlled, not begin neglect.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

The knee-jerk concern is that with monies scarce, the cost of storage is a burden – and that is a legitimate burden at scale. But ultimately the biggest barriers are in 1) scoping what data has or could have transformative value (e.g. where to focus most effort) and 2) articulating and implementing standards to maximize that data's potential and the 3) getting the community to support and adopt those standards.

4. Any other relevant issues respondents recognize as important for NIH to consider

Academic institutions are first and foremost learning centers. Individuals come to gain knowledge and experience - mistakes, technique-building, risk-taking, and *iterative understanding* of the science are all core to a student's experience -- are core to science itself since all researchers are students their entire career one way or another. So it should be emphasized that the goal of sharing is not to police research and/or ensure that others can verify that a given set of results are unequivocally "correct." *The latitude to make mistakes and revise hypotheses must be preserved.* To take it a step further, these are fairly new levels of transparency and to set expectations, NIH could consider issuing an accompanying Code of Conduct for review and critique of shared data/software. A "sharing must not be used as an opportunity to gain unfair advantages, bash alternate methodology, act on personal agendas, etc." sort of document. Such social contracts shouldn't be necessary, yet it wouldn't hurt to see this in writing and endorsed by a major agency.

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

Impact is that it will help to establish a "new normal" for the researchers and adds dimensionality to the work, may get researchers to think about designing research with creation of valuable datasets in mind from the onset as opposed to as a by-product. Will be complaints about the additional reporting, and reporting for the sake of reporting is, in fact, bureaucratic. So best to convey all new reporting requirements in terms of how reporting of dataset/software products will be used and the benefits of reporting. For example, will NIH track reported (and verified?) shared datasets/software in an author/grantee database? If so, how will that be used to the discipline's and authors' benefit?

a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI)

This is ideal, but not always available, e.g. accession numbers in and of themselves do not resolve but would be evidence of data sharing if deposited in an NCBI resource. Considering some of the current ambivalence about data sharing, creating a frustrating experience where researchers are asked to do something (in this case provide a resolvable PID) that they may not have control over is not likely to endear the process to them. Instead, encourage inclusion of resolvable identifiers when possible (by way of multiple example citations with and without resolvable PIDs) and expect for reporting to evolve as systems evolve. But the first step seems to be to just acknowledge the data was shared publicly at all and then improve iteratively from there.

b. Inclusion of a link to the data/software resource with the citation in the report

Same as resolvable ID in a) – a DOI, handle, ARK, will do that so if they are available, that will work. But is anyone going to actually click on them in a report? Again, thinking of accession numbers, researchers could provide an accession as a link but could require tedious reformatting which would be senseless unless the link is actually used. Is the expectation that reports will be spit through a program and links will be machine read? And a program officer will get a report that says what resolves or doesn't resolve? And then what? A box gets a check? A judgement about what should have been shared but wasn't? The NOA is held up? Better to ask for "anticipated dataset/software products that will be made publicly available" to be included in a DMP so the reporting is tangible. See answer to 3 below.

c. Identification of the authors of the Data/Software products

Since the question is specifically about grant reports, which would largely be tied to discrete projects, I'm guessing that definition of authors is possible and requiring a formal citation in the report would provide the authors in that case.

Could define/adopt NIH-endorsed authorship guidelines for data and software if too much confusion or there are specific edge cases that are of concern.

d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

For the sake of reporting the impact of the grant, expect best for them all to be reported. Worried about pages and pages of citations? (Or asking grantees to prepare pages and pages of data citations?) If that is the concern, can ask to report of total number of datasets published and then ask them to list "selected datasets" that best exemplify the nature and quality of the work performed under the grant (max 10). Or something similar.

e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed

Impact here may be greater for the repository's metrics than for the individual dataset or software, but a "publisher" (in this case the repository) is a natural inclusion in a citation so along with authorship, asking for formal citations would be most helpful.

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications

One of the most concrete ways to raise awareness and also make subsequent reporting tangible would be to encourage articulation of "expected data/software products that will be made publicly available" in the grant application – probably in a DMP. Guidance will need to be provided to review panel members so they have a way to evaluate what kinds of dataset could/should be expected for a given FOA. But by articulating products up front at the time of application, it brings data to the forefront of the project so researchers will *know what's expected of them* and hopefully be in a better position to publish better quality datasets because of that expectation as well. By defining up front, this would also allow follow-up in PMCID-required sort of way (e.g. you said you'd publish data with each one of your papers, you reported 5 papers, please provide citations for the corresponding published datasets). It could be that not all products will be anticipated, and in the end more will be reported, but that's great. It might be that products were anticipated but failed to materialize, that's not unexpected either, and can provide provisions to submit a short explanation in the RPPR – or suggest publication of negative results, depending on the reason for failure.

4. Any other relevant issues respondents recognize as important for NIH to consider

Academic institutions are first and foremost learning centers. Individuals come to gain knowledge and experience - mistakes, technique-building, risk-taking, and *iterative understanding* of the science are all core to a student's experience -- are core to science itself since all researchers are students their entire career one way or another. So it should be emphasized that the goal of sharing is not to police research and/or ensure that others can verify that a given set of results are unequivocally "correct." *The latitude to make mistakes and revise hypotheses must be preserved.* To take it a step further, these are fairly new levels of transparency and to set expectations, NIH could consider issuing an accompanying Code of Conduct for review and critique of shared data/software. A "sharing must not be used as an opportunity to gain unfair advantages, bash alternate methodology, act on personal agendas, etc." sort of document. Such social contracts shouldn't be necessary, yet it wouldn't hurt to see this in writing and endorsed by a major agency.

Additional Comments