

Compiled Public Comments on
the Request for Information on
the NIH Plan to Increase
Findability and Transparency of
Research Results
Through the Use of Metadata
and Persistent Identifiers

December 17, 2024 – February 21, 2025

Table of Contents

1. N/A
2. Christopher Gregg PhD
3. N/A
4. Christos Katsanos
5. Children's Hospital of Philadelphia
6. ICPSR/University of Michigan
7. Federation of American Societies for Experimental Biology (FASEB)
8. Copyright Clearance Center
9. Elsevier
10. ORCID
11. APLU, ARL, AAU, and COGR
12. Association of American Medical Colleges
13. Association of American Publishers
14. AAAS

Submit date: 12/18/2024

I am responding to this RFI: Behalf of Myself

Name:

Name of Organization:

Type of Organization: University

Type of Organization-Other:

Role: Scientific Researcher

Comments: Every new requirement places additional burdens on already overburdened scientists. We are spending less and less effort on actual science and more and more effort complying with regulations. Grant budgets have not increased with inflation so each grant already buys less science than 10 years ago. Couple that with increasing compliance burdens and requirements to give data to the public. Suddenly, the incentive to collect data is very low.

Submit date: 12/19/2024

I am responding to this RFI: Behalf of Myself

Name: Christopher Gregg PhD

Name of Organization: university of utah school of medicine

Type of Organization: University

Type of Organization-Other:

Role: Scientific Researcher

Comments: One of my great frustrations is that most important clinical research papers are behind paywalls (NEJM, JAMA, etc.), yet patients and doctors (and now AI tools) must be able to access this work. I don't think the new revisions go far enough. We need to enforce retroactive open access and sharing of previously published papers by scientists and clinicians at "any institute that has received or is actively receiving federal research funds". All knowledge relevant to health, disease, and patient care must be public and open-source going back at least 30 years. The mandate for compliance should fall to the funded institutions (universities, health systems etc) to ensure that pdfs are obtained from the paywall journal and deposited in pubmed. Ideally data would also be shared, but that may be too much to ask. Every individual researcher can easily dump a pdf of their previous papers into pubmed.

Submit date: 12/19/2024

I am responding to this RFI: Behalf of Myself

Name:

Name of Organization:

Type of Organization: University

Type of Organization-Other:

Role: Scientific Researcher

Comments: Please force the publishers to make the research public, rather than adding yet another requirement for researchers. Already overtaxed researchers receive more and more work and less support from government, institutions, and for-profit companies. You keep making new regulations and then for-profit publishers and nonprofit universities make our lives difficult and put most of the burden of complying on us. The people who are supposed to be doing the real work. But we are jumping through extra hoops when we should be doing science.

Submit date: 12/20/2024

I am responding to this RFI: Behalf of Myself

Name: Christos Katsanos

Name of Organization:

Type of Organization: University

Type of Organization-Other:

Role: Scientific Researcher

Comments: This plan sounds fine, but it is a small and slow step on how fast everything moves.

NIH is large enough to lead its own publication enterprise for its funded research and in an independent manner as currently does with the grant application review processes.

Journals will be of low cost, or possibly no cost (if NIH funded research) to publish, no impact factors, ensure reputation of the journal/publication and speed of review process. Can directly work an efficient approach in regards to open access vs subscription.

Thanks,

Christos

Christos S. Katsanos, PhD, FACSM, FAPS

Director, Human Obesity Metabolism Laboratory

Director, Biology PhD Program, School of Life Sciences, Arizona State University

Associate Professor, School of Life Sciences, Arizona State University

Adjunct, Department of Physiology and Biomedical Engineering, Mayo Clinic

Health Futures Center, Room 331C

6161 E. Mayo Blvd

Phoenix, AZ 85054

Christos S. Katsanos, PhD, FACSM, FAPS

Director, Human Obesity Metabolism Laboratory

Director, Biology PhD Program, School of Life Sciences, Arizona State University

Associate Professor, School of Life Sciences, Arizona State University

Adjunct, Department of Physiology and Biomedical Engineering, Mayo Clinic

Health Futures Center, Room 331C

6161 E. Mayo Blvd

Phoenix, AZ 85054

Phone: (602) 543-4254

@HOMeScienceLab

Submit date: 12/20/2024

I am responding to this RFI: Behalf of an Organization

Name: Annabel Pinkney (Metadata Librarian), Christopher Forrest (Director, Applied Clinical Research Center)

Name of Organization: Children's Hospital of Philadelphia

Type of Organization: Nonprofit Research Organization

Type of Organization-Other:

Role: Institutional Official

Comments: This policy adequately underscores the importance of consistently and accurately applied metadata. Section IIB could be strengthened by providing specific guidance on the minimum metadata required by NIH-supported repositories. While the section currently notes that any PIDs are generally accepted for names, affiliations, and associated publications, it would be helpful to include a list of accepted PID types/sources for each field (ORCID, ROR, ISNI, DOI, etc.). This would reduce ambiguity and clarify the expectations for content. Additionally, guidance on how to handle cases where a name or entity does not have an existing PID would assist both depositors and repository maintainers. We encourage the NLM to socialize these guidelines with other federal and industrial agencies to provide a national standard.

We also recommend that the plan include a finalized resource for NIH award and project PIDs (Section III). The current draft references the upcoming formalization of an identification system for NIH awards and projects, along with the intention to require the use of these identifiers. The plan will be more effective if this work is completed prior to publication, enabling the inclusion of concrete guidance on the use of these PIDs.

Lastly, while the plan encourages researchers to update their ORCID iD records to reflect their individual research corpus, it could benefit from placing greater emphasis on the overarching benefits of retroactively improving existing metadata. Specifically, guidance on how and what to update would ensure that legacy resources are brought into alignment with updated standards, improving accessibility and consistency across the board.

Submit date: 1/28/2025

I am responding to this RFI: Behalf of an Organization

Name: Margaret Levenstein

Name of Organization: ICPSR/University of Michigan

Type of Organization: University

Type of Organization-Other:

Role: Institutional Official

Comments: Please find attached ICPSR's response and recommendations for NIH Plan to Increase Findability and Transparency of Research Results Through the Use of Metadata and Persistent Identifiers. ICPSR is a data archive that hosts multiple NIH data collections (NACDA (NIA), DSDR (NICHD), and NAHDAP (NIDA)) that preserve and disseminate thousands of datasets whose production was funded by NIH. These data are actively used by thousands of researchers and students.

Uploaded File: 20250114-ICPSR-Comments-on-NIH-plan-to-increase-transparent-of-research-results-via-use-of-metadata-and-PIDs.pdf

Description: ICPSR comments on NIH Plan to Increase Findability and Transparency of Research Results Through the Use of Metadata and Persistent Identifiers i

January 14, 2025

Subject: ICPSR Comments on the [NIH Plan to Increase Findability and Transparency of Research Results Through the Use of Metadata and Persistent Identifiers](#)

Dear NIH Office of Science Policy,

I am submitting these comments in my capacity as Director of the Inter-university Consortium for Political and Social Research ([ICPSR](#)), a unit within the Institute of Social Research at the University of Michigan. ICPSR, a CoreTrustSeal-certified repository that meets the NIH's "Desirable Characteristics of Data Repositories for Federally Funded Research," curates, preserves, and disseminates over 20,000 social and behavioral science data collections, including hundreds originally funded by the National Institutes of Health (NIH). ICPSR also hosts specialized collections supported directly by the NIH Eunice Kennedy Shriver National Institute of Child Health and Human Development, the NIH National Institute on Aging, and the NIH National Institute on Drug Abuse.

I applaud the NIH's plans to increase standardized, publicly available metadata for research outputs, especially to:

- Instruct federally-funded researchers to obtain a digital persistent identifier (PID).
- Assign unique persistent identifiers to all scientific research awards.
- Gather metadata from data repositories.

These steps should dramatically increase the uptake and use of established persistent identifiers in the United States, including ORCiDs and RORs, and increase the interoperability and re-use of NIH-funded research outputs. These recommendations will be more effective if they do not simply impose additional requirements on individual researchers or research projects. The ability of individual researchers to follow NIH guidance in this area depends on the capacity of both researchers and research data institutions, including data archives and organizations providing PIDs, to serve the research community. We encourage NIH to invest in these institutions supporting the research community so that individual PIs can efficiently comply with NIH guidelines and provide research data to the community at the lowest possible cost.

I outline below specific aspects of the plan that warrant further consideration and refinement, drawing on ICPSR's 60 years of experience in data stewardship as a trusted, long-lived repository of NIH research output.

I.C. Submitting Metadata and PIDs when Depositing Scientific Data in Repositories

The proposed plan says it will expect submissions of scientific data to a data repository to include the following metadata:

- ORCID iDs/PIDs and names for contributing senior and key personnel,

ORCiDS are created and managed by individual researchers themselves, which precludes their creation on behalf of deceased or other uncontactable personnel. During the initial takeup of this policy, NIH should provide guidance on how to assign alternative identifiers for deceased or otherwise unreachable senior and key personnel who do not already have a PID.

- affiliations (or other PIDs for affiliations) for contributing senior and key personnel, [a footnote specifies: “Because affiliation information can become ambiguous over time if inconsistently reported, award recipient organizations may optionally work with PID providers such as the Research Organization Registry (ROR) or International Standard Name Identifier (ISNI). Alternatively, award recipient organizations may encourage any NIH-supported research contributors within their organization to use a defined, consistent affiliation text and a specific PID, as available, when reporting research outputs.”]

The NIH should provide a stronger recommendation by requiring (not recommending) the use of PIDs for affiliations and not simply allowing standardized free text if PIDs aren't used. Such a mandate would increase the adoption of PIDs by organizations whose researchers benefit from NIH funding, broadening and deepening the research community's commitment to PIDs. Continued allowance of free text undermines the precision, interoperability, and discoverability of metadata.

- For all other contributors, encourage institutions to consider consistently providing these metadata and PIDs to the repository, as able.

While it is good the NIH recommends PIDs for “other contributors,” the NIH should provide more specific guidance about who they recommend to be considered as the “other contributors.” This is an opportunity for the NIH to augment what information repositories collect and share. Many repositories still cite just the senior personnel. The Force11 Joint Declaration of Data Citation Principles notes, “Data citations should facilitate giving scholarly credit and normative and legal attribution to all contributors to the data, recognizing that a single style or mechanism of attribution may not be

applicable to all data.”¹ The Contributor Role Taxonomy (CRediT)² provides one option for better defining who should be considered as a contributor to a research output.

I.E. Citing and Cross-Linking Metadata and PIDs

The proposed plan encourages researchers to follow community best practices around data citation.

Although data citation standards exist and adoption continues to improve, inconsistencies still exist in how data citation information is recorded in publications.³⁴ This is an area where more NIH guidance could significantly improve adoption. NIH should provide specific citation guidelines on where and how to cite data rather than simply alluding to and relying on “community standards.” NIH’s guidelines and requirements for sharing data will be more effective, and implemented more efficiently, if researchers have positive incentives to share data. Data citations provide such an incentive.

II.B. Collecting Publicly Available Metadata and PIDs for Scientific Data in NIH-Supported Repositories

The proposed plan is expected to require data repositories to collect and make publicly available metadata, including about the submission date of the scientific data.

The NIH should specify acceptable formats, schemas, and protocols for making metadata publicly available. Several options, including APIs and the Open Archives Initiative Protocol for Metadata Harvesting, are widely adopted by data repositories and would effectively allow the NIH to harvest the relevant metadata.⁵

Repositories may collect multiple data submissions connected to one funded research project. The NIH should clarify which of the potentially multiple submission dates are expected to be made available.

¹ <https://doi.org/10.25490/a97f-egyk>

² <https://credit.niso.org/>

³ Citation guidance includes:

- 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. <https://doi.org/10.2481/dsj.OSOM13-043>
- Cousijn, H., Kenall, A., Ganley, E. *et al.* A data citation roadmap for scientific publishers. *Sci Data* 5, 180259 (2018). <https://doi.org/10.1038/sdata.2018.259>

⁴ Donovan, G C and Langseth, M L 2024 Are Researchers Citing Their Data? A Case Study from The U.S. Geological Survey. *Data Science Journal*, 23: 24, pp. 1–14. DOI: <https://doi.org/10.5334/dsj-2024-024>

⁵ <https://www.openarchives.org/pmh/>

ICPSR is excited to see the NIH making significant advancements in promoting the adoption and use of metadata and persistent identifiers, which play a pivotal role in driving data re-use and enhancing interoperability. Thank you for this opportunity to provide comments.

Sincerely,

Margaret C. Levenstein
Director, ICPSR
Professor, School of Information
Research Professor, Institute for Social Research
Adjunct Professor of Business Economics, Ross School of Business
University of Michigan
(734) 615-8400
maggiel@umich.edu

Submit date: 2/5/2025

I am responding to this RFI: Behalf of an Organization

Name: Beth A. Garvy, PhD

Name of Organization: Federation of American Societies for Experimental Biology (FASEB)

Type of Organization: Professional Org or Association

Type of Organization-Other:

Role: Scientific Researcher

Comments: **Comments also attached in PDF format on organizational letterhead**

The Federation of American Societies for Experimental Biology (FASEB) appreciates the opportunity to provide feedback on the National Institutes of Health (NIH) plan to increase findability and transparency of research results through the use of metadata and persistent identifiers as published in the NIH Guide on December 17, 2024. We applaud the agency's commitment to enhance public access to NIH-supported research and ensure transparency of research findings. FASEB's comments on specific sections of the plan are provided below.

Section I.D., Reporting PIDs to NIH (pages 5 – 6)

FASEB supports NIH's expectation for NIH-supported institutions and NIH intramural investigators to include PIDs in proposals for funding and research performance progress reports. This will facilitate proper attribution of prior works and increase the agency's ability to link investments with research outputs. While the majority of NIH-funded investigators are already reporting PubMed Central Identifiers (PMcIDs) in their grant applications and progress reports, PMcIDs do not fulfill the interoperability requirements included in the 2022 White House Office of Science and Technology Policy Memorandum on Ensuring Free, Immediate, and Equitable Access to Federally Funded Research. Therefore, FASEB encourages the use of digital object identifiers (DOIs) and ORCID identifiers in grant applications and progress reporting.

Section I.E., Citing and Cross-Linking Metadata and PIDs (page 6)

The NIH Plan encourages researchers to add their research outputs to their ORCID records. While FASEB appreciates the sentiment of this recommendation, relying only on investigator inputs is not a best practice within the ORCID community. For publications, data deposition, and employment, the authoritative sources to write to an ORCID record are the publisher, repository, or employer/institution - not the individual. The authoritative source is defined as the entity that completes the action (e.g., publishes, posts datasets, employs the researcher at their institution). Therefore, we suggest that NIH update this section to encourage researchers to opt in with the publisher, data repository, and their institution to write to their ORCID record. If not already the case, FASEB urges NIH to serve as the authoritative source for confirming grant awards reported to ORCID records.

Section II.B., Collecting and Making Metadata and PIDs Publicly Available (page 7)

FASEB applauds NIH's decision to develop a minimum set of metadata standards for scientific data repositories to collect and make publicly available, echoing our prior comments on the draft NIH

Strategic Plan for Data Science, 2023 - 2028. This first set of metadata standards are broadly applicable, addressing general needs such as the researchers who generated the data, their affiliated institutions, and associated funding and publications. FASEB encourages NIH to continue development of additional specific metadata standards for the various types of scientific data being reported. The loss of information between data collection and data reporting data leads to slower uptake of research reuse projects and limits the usability of data being reported. To facilitate this critical step, FASEB strongly recommends NIH issue Notices of Funding Opportunities (NOFOs) to support workshops or related convenings to establish metadata standards that are broadly applicable to specific research domains. Improving the types and quality of metadata reported with data files will enhance the return on NIH's research investments.

Section III, Assigning Identifiers for NIH Awards and NIH-Conducted Research Projects (page 8)

FASEB recognizes that NIH grant award numbers have been in existence for a long time, and have a meaningful structure that provides staff and researchers with relevant information about the grant, including the type of application, category of support, institute or center associated with the grant, unique identifier for the individual grant, current year of support, and suffixes for supplements, amendments, or fellowship institutional allowances. This same type of information could also be captured in parts of the digital object identifier (DOI), a unique number designed to be used by humans as well as machines. DOIs are persistent identifiers with a set structure that provides reasonable flexibility for various use cases, are broadly indexed and globally adopted as a default identifier, and enable citation and linking between publications, datasets, and software. The DOI suffix can be almost any string of characters and symbols so long as those characters are allowed in a URL. Adopting DOIs for research grant awards would provide the global community with the most rapid integration of grants with publications, datasets, and software, enabling a clear path for maximizing access to, use of, and connections between these output types, improving the ability to track research outcomes and impact. FASEB encourages that NIH adopt the digital object identifier (DOI) as the persistent identifier for NIH awards and NIH-conducted research projects.

Thank you for providing the research community with an opportunity to review and comment on the proposed plan.

Uploaded File: FINAL_FASEB-Comments-on-NIH-Metadata-and-PID-Plan-RFI_20250205.pdf

Description: FASEB comments on NIH plan to increase findability and transparency of research results through the use of metadata and persistent identifiers, formatted on organizational letterhead.



February 5, 2025

Lyric Jorgenson, PhD
Associate Director for Science Policy
National Institutes of Health
6705 Rockledge Drive
Bethesda, MD 20817

Transmitted electronically via [online form](#)

The Federation of American Societies for Experimental Biology (FASEB) appreciates the opportunity to provide feedback on the National Institutes of Health (NIH) plan to increase findability and transparency of research results through the use of metadata and persistent identifiers as published in the [NIH Guide](#) on December 17, 2024. We applaud the agency's commitment to enhance public access to NIH-supported research and ensure transparency of research findings. FASEB's comments on specific sections of the plan are provided below.

Section I.D., Reporting PIDs to NIH (pages 5 – 6)

FASEB supports NIH's expectation for NIH-supported institutions and NIH intramural investigators to include PIDs in proposals for funding and research performance progress reports. This will facilitate proper attribution of prior works and increase the agency's ability to link investments with research outputs. While the majority of NIH-funded investigators are already reporting PubMed Central Identifiers (PMCID) in their grant applications and progress reports, PMCID do not fulfill the interoperability requirements included in the 2022 White House Office of Science and Technology Policy Memorandum on *Ensuring Free, Immediate, and Equitable Access to Federally Funded Research*. Therefore, FASEB encourages the use of digital object identifiers (DOIs) and ORCID identifiers in grant applications and progress reporting.

Section I.E., Citing and Cross-Linking Metadata and PIDs (page 6)

The NIH Plan encourages researchers to add their research outputs to their ORCID records. While FASEB appreciates the sentiment of this recommendation, relying only on investigator inputs is not a best practice within the ORCID community. For publications, data deposition, and employment, the authoritative sources to write to an ORCID record are the publisher, repository, or employer/institution - not the individual. The authoritative source is defined as the entity that completes the action (e.g., publishes, posts datasets, employs the researcher at their institution). Therefore, we suggest that NIH update this section to encourage researchers to opt in with the publisher, data repository, and their institution to write to their ORCID record. If not already the case, FASEB urges NIH to serve as the authoritative source for confirming grant awards reported to ORCID records.

Section II.B., Collecting and Making Metadata and PIDs Publicly Available (page 7)

FASEB applauds NIH's decision to develop a minimum set of metadata standards for scientific data repositories to collect and make publicly available, echoing our prior comments on the draft [NIH](#)

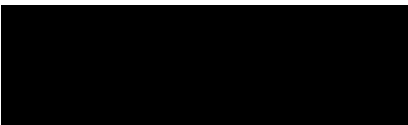
[Strategic Plan for Data Science, 2023 - 2028](#). This first set of metadata standards are broadly applicable, addressing general needs such as the researchers who generated the data, their affiliated institutions, and associated funding and publications. FASEB encourages NIH to continue development of additional specific metadata standards for the various types of scientific data being reported. The loss of information between data collection and data reporting data leads to slower uptake of research reuse projects and limits the usability of data being reported. To facilitate this critical step, FASEB strongly recommends NIH issue Notices of Funding Opportunities (NOFOs) to support workshops or related convenings to establish metadata standards that are broadly applicable to specific research domains. Improving the types and quality of metadata reported with data files will enhance the return on NIH's research investments.

Section III, Assigning Identifiers for NIH Awards and NIH-Conducted Research Projects (page 8)

FASEB recognizes that NIH grant award numbers have been in existence for a long time, and have a meaningful structure that provides staff and researchers with relevant information about the grant, including the type of application, category of support, institute or center associated with the grant, unique identifier for the individual grant, current year of support, and suffixes for supplements, amendments, or fellowship institutional allowances. This same type of information could also be captured in parts of the digital object identifier (DOI), a unique number designed to be used by humans as well as machines. DOIs are persistent identifiers with a set structure that provides reasonable flexibility for various use cases, are broadly indexed and globally adopted as a default identifier, and enable citation and linking between publications, datasets, and software. The DOI suffix can be almost any string of characters and symbols so long as those characters are allowed in a URL. Adopting DOIs for research grant awards would provide the global community with the most rapid integration of grants with publications, datasets, and software, enabling a clear path for maximizing access to, use of, and connections between these output types, improving the ability to track research outcomes and impact. FASEB encourages that NIH adopt the digital object identifier (DOI) as the persistent identifier for NIH awards and NIH-conducted research projects.

Thank you for providing the research community with an opportunity to review and comment on the proposed plan.

Sincerely,



Beth A. Garvy, PhD
FASEB President

Submit date: 2/7/2025

I am responding to this RFI: Behalf of an Organization

Name: Roy Kaufman

Name of Organization: Copyright Clearance Center

Type of Organization: Other

Type of Organization-Other: Non profit engaged in research support, PIDs, and copyright licensing.

Role: Member of the Public

Comments: Please see the attached document. Thank you.

Roy Kaufman

Uploaded File: CCC-comments-to-NIH-PID-plan-with-attachment-final.pdf

Description: CCC comments NIH Plan to Increase Findability and Transparency of Research Results Through the Use of Metadata and Persistent Identifiers (PID)



Response of Copyright Clearance Center to NIH Plan to Increase Findability and Transparency of Research Results Through the Use of Metadata and Persistent Identifiers.

Copyright Clearance Center applauds the NIH plan to increase findability and transparency of research results through the use of metadata and persistent identifiers (the “[Metadata Plan](#)”), which generally aligns without our comments (attached hereto) to Question 4 of the NIH’s Request for Information on the NIH Plan to Enhance Public Access to the Results of NIH-Supported Research (the “Public Access Plan”). More importantly, we welcome NIH’s continued interest in PIDs and metadata.

Background on CCC.

CCC is a not-for-profit organization founded in 1977 at the suggestion of Congress to facilitate collective copyright licensing for the text sector. Presently, among other lines of business, CCC provides licenses to content from over 10,000 rightsholders for whom we serve as an agent. We provide these licenses to more than 35,000 business organizations (Business Users) around the world. CCC is a supplier of knowledge management software called RightFind®, which is used by a subset of these Business Users to manage and access content. We also provide (1) other software services, (2) library staffing, (3) content enrichment, data and metadata services, and (4) content delivery.

Our fastest growing business is managing the agreement- and fee-administration process on behalf of publishers who collect fees or otherwise track usage from authors, institutions, consortia, government and other funding bodies for immediate open access (OA). We do this primarily through our [RightsLink® for Scientific Communications](#) software platform (RLSC). RLSC is by far the market leader in managing open access agreements and payments, doing so for many of the top publishers of NIH-funded research.

In 2022 we acquired [Ringgold](#), which we had previously adopted as our preferred organizational PID. We did this to support the ongoing development of RLSC and also because of Ringgold’s widespread adoption in the publishing industry,

Comments.

CCC addressed our comments to the Public Access Plan specifically to the metadata question in Question 4. Given the relevance of our comments to the Public Access Plan, we herein restate selected comments submitted by CCC in response to the Public Access Plan, followed by statement of how there are addressed by Metadata Plan:

Through both our knowledge management work with Business Users and our work on behalf of publishers, CCC experiences firsthand the promise of persistent identifiers (PIDs) when applied early, consistently and persistently. We are also

keenly aware of the problems related to the entropy that results from lack of early, consistent, and persistent application thereof.

A healthy research and publishing ecosystem requires PIDs and robust, rich, high quality metadata to make connections among people, organizations, places, and digital objects. For example, in RLSC alone, we depend on dozens of author, institution, and manuscript metadata elements to apply the appropriate business logic and workflows necessary to automate and scale OA on the path toward open science.

The Metadata Plan recognizes the importance of PIDs and requires researchers to register for ORCID iDs (as is required for many journals). Use of these PIDs (or other sector accepted PIDs such as ISNI) will aid author disambiguation and make research more findable and usable for all users.

On the issue of organizational IDs, we stated:

[E]ven within a seemingly unified sector such as scientific publishing, it is sometimes necessary to accommodate multiple PIDs serving the same purpose, such as organizational identifiers. While in some ways accommodating multiple PIDs increases work and decreases interoperability, PIDs have different scope, attributes, and audiences. Some users prefer PIDs with ISO certification, while others prefer PIDs with established business models to ensure sustainability and maintenance, and others focus on ability to use without cost to access PIDs. When one PID has been selected for use by a stakeholder as part of master data management, being forced to accommodate a different PID can have significant costs and introduce unnecessary friction. Accordingly at CCC, we accommodate a variety of organizational IDs in RLSC and have long preferred the features of Ringgold for our primary use.

The Metadata plan expresses a desire for use of organizational IDs, without mandating a specific PID (Ringgold, ISNI, RoR, etc.). This reflects the reality that different organizational IDs have different core users and that mandates favoring one will increase costs and friction. We support this conclusion, while noting that as Ringgold is (1) free to the researcher, (2) a registration agency for the ISO-approved ISNI standard, with which it is interoperable, and (3) ubiquitous in publishing. As such, we recommend listing it with ISNI and RoR in footnote 1.

On the issue of identifying grants and awards, we said:

As a final recommendation, we suggest that NIH follow the lead of Wellcome Trust and the Bill and Melinda Gates Foundation, among others, in registering grants for DOIs. This will help enable connectivity of PIDs and the discoverability of the grants, maximizing return to US taxpayers.

The Metadata Plan states that “NIH plans to consider exploring avenues to identify NIH awards and NIH-conducted research projects with globally unique, machine resolvable PIDs. NIH plans to coordinate this exploration with efforts of other federal agencies and relevant communities to assess how to best develop a robust, connected ecosystem where institutions, researchers, research outputs, and funding sources are linked consistent with Findable, Accessible, Interoperable, and Reusable (FAIR) Principles.” We agree with this approach.

Finally, we again recommend that NIH periodically review the chart CCC maintains at <https://www.copyright.com/stateofmetadata/>. This has been updated since our response to the Public Access Plan.

Respectfully submitted for Copyright Clearance Center by,

A solid black rectangular box used to redact the signature of the sender.



Response of Copyright Clearance Center (CCC) to Request for Information on the NIH Plan to Enhance Public Access to the Results of NIH-Supported Research (RFI)

Notice Number:
NOT-OD-23-091

CCC welcomes the opportunity to submit this response to Question 4 of the NIH's [Request for Information on the NIH Plan to Enhance Public Access to the Results of NIH-Supported Research](#). More importantly, we welcome NIH's interest in the use of PIDs and metadata to increase findability and transparency of scientific research.

Background on CCC.

CCC is a not-for-profit organization founded in 1977 at the suggestion of Congress to facilitate collective copyright licensing for the text sector. Presently, among other lines of business, CCC provides licenses to content from over 10,000 rightsholders for whom we serve as an agent. We provide these licenses to more than 35,000 business organizations (Business Users) around the world. CCC is a supplier of knowledge management software called RightFind®, which is used by a subset of these Business Users to manage and access content. We also provide (1) other software services, (2) library staffing, (3) content enrichment, data and metadata services, and (4) content delivery. On October 19, 2021, [U.S. Secretary of Commerce Gina Raimondo announced](#) that we were awarded a Market Development Cooperator Program grant, administered by the Commerce Department's International Trade Administration, to support our work with standards development organizations.

Our fastest growing business is managing the agreement- and fee-administration process on behalf of publishers who collect fees or otherwise track usage from authors, institutions, consortia, government and other funding bodies for immediate open access (OA). We do this primarily through our RightsLink® for Scientific Communications software platform (RLSC). RLSC is by far the market leader in managing open access agreements and payments, doing so for many of the top publishers of NIH-funded research.

PIDs and Metadata.

Through both our knowledge management work with Business Users and our work on behalf of publishers, CCC experiences firsthand the promise of persistent identifiers (PIDs) when applied early, consistently and persistently. We are also painfully aware of the problems

related to the entropy that results from lack of early, consistent, and persistent application thereof.

A healthy research and publishing ecosystem requires PIDs and robust, rich, quality metadata to make connections among people, organizations, places, and digital objects. For example, in RLSC alone, we depend on dozens of author, institution, and manuscript metadata elements to apply the appropriate business logic and workflows necessary to automate and scale OA on the path toward open science.

Even within a seemingly unified sector such as scientific communications, it is sometimes necessary to accommodate multiple PIDs serving the same purpose, such as organizational identifiers. While in some ways accommodating multiple PIDs increases work and decreases interoperability, PIDs have different scope, attributes, and audiences. Some users prefer PIDs with ISO certification, while others prefer PIDs with established business models to ensure sustainability and maintenance, while others focus on ability to use without cost to access PIDs. When one PID has been selected for use by a stakeholder as part of master data management, being forced to accommodate a different PID can have significant costs and introduce unnecessary friction. Accordingly at CCC, we accommodate a variety of organizational IDs in RLSC and have long preferred the features of Ringgold for our primary use.¹

Review of data quality of bibliographic records from the MEDLINE database

In 2022, three CCC colleagues reviewed the data quality of bibliographic records in the Medline database. A paper detailing the results of their research have been posted on bioRxiv and is attached to this document (Bramley, R, Howe, S, Marmanis, H 2022, Notes on the data quality of bibliographic records from the MEDLINE database, doi: <https://doi.org/10.1101/2022.09.30.510312>; hereafter, “Bramley, et al”). As noted in the paper:

[T]he PubMed database, which contains over 33.8 million records collected over many decades, suffers from several data quality issues. These issues relate to, in part, character encodings, the absence of persistent identifiers, differences in human languages, and schema changes. These shortcomings should not be surprising since

¹ CCC adopted Ringgold as its preferred organizational PID approximately 8 years ago. CCC acquired Ringgold in 2022 so that we could ensure its continued viability given its importance to ourselves and our clients.

PubMed aggregates information produced by different publishers and XML providers, a fact that leads naturally to the presence of “multi-source” problems.

Among the conclusions of the paper are (1) “[g]iven the incompleteness and uniqueness of identifying fields, the disambiguation of author names remains a significant problem for PubMed, particularly for records dating before 2014, and (2) [o]verall, there is an improvement in the use of identifiers; in particular, records created since 2015 exhibit an increase in external identifiers. However, the data quality for institutional identifiers is poor and their use has been diminishing over time.”

Mapping metadata management across the research lifecycle.

In late 2022, CCC and Media Growth Strategies undertook a thorough examination of metadata management across the research lifecycle. This review builds on an existing body of work to uncover multiple system complexities and breakages, which – separately and together – create missed opportunities for the communities for whom OA and open science models are designed to serve.

CCC has made this information publicly available in interactive infographic form at <https://www.copyright.com/stateofmetadata/>, and we have attached a chart summarizing where metadata breakages occur throughout the research lifecycle and how they impact various stakeholder groups. Drawn directly from research interviews, the infographic depicts the significant economic impact that a fragmented metadata supply chain is having today on researchers, institutions, funders, and publishers. Researchers in particular shoulder a significant administrative burden that ultimately disrupts and delays the process of scientific discovery.

The infographic is a living document which will be updated and modified based on ongoing community feedback.

As the scholarly communications community continues its shift to OA and open science, stakeholders require a robust network of interoperable systems for making critical and necessary improvements, and much progress is underway. In that environment, a dedication to data stewardship across each stakeholder group, and the service providers supporting them, will lead to greater data sharing; reliable, trustworthy metrics on research impact; and a responsive, equitable rewards system. NIH can lead the way.

Question 4 of the RFI states: “NIH seeks suggestions on any specific issues that should be considered in efforts to improve use of PIDs and metadata, including information about experiences institutions and researchers have had with adoption of different identifiers.”

First, we recommend that NIH review the research, findings and recommendations set forth in Bramley, et al.

Second, NIH, as the premier funder of biomedical research in the US, is well positioned to help research and lead by example by requiring PIDs at appropriate points. As can be seen in the above-referenced infographic, grant application is one of the first organized parts of the lifecycle where PIDs can be effectively mandated. Once mandated and used, PIDs can flow throughout the lifecycle to improve everything from grant management to expression in PubMed. We urge NIH to review the infographic, sign up for updates, and provide feedback should NIH believe there are amendments and changes needed.

We have three specific recommendations with respect to mandated use of PIDs:

- 1. NIH should mandate that grant applications include organizations IDs for the institutions(s) affiliated with each researcher listed on the grant application, and Funder Registry IDs for the distinct funders of the grant. The requirement should insist that grant applications include at least one of the following organizational identifiers used in the scholarly publishing ecosystem and NIH should make metadata fields available for all four:**

A. Ringgold- a proprietary global organization identifier system owned by CCC with over 600,000 unique records and rich hierarchical metadata used today by (1) most large and mid-sized commercial and non-commercial publishers, and (2) a range of critical infrastructure providers in the publishing ecosystem. For publishers, Ringgold often is part of a master data management strategy. Ringgold is also used by some funders, academic institutions, and consortia. Ringgold maps one-to-one with ISNI and the Funder Registry.

B. ISNI- ISO standard name identifier system with 1,697,000 unique organizational records of which a minimum of 500,000 are relevant to the research sector. ISNI is free to use and has been adopted by many national libraries. It lacks the hierarchical metadata of Ringgold but enjoys the rigor and authority of ISO accreditation. The

relevant organization records in ISNI map one-to-one with Ringgold.

C. ROR- Research Organization Registry (ROR) is a global, community-led registry of open persistent identifiers for research organizations. ROR is free to use and has been adopted by some publishers, institutions, and overseas funders. It contains 104,000 unique identifiers and some hierarchical metadata. It can map to ISNI and the Funder Registry, but not on a one-to-one basis.

D. Funder Registry (formerly known as FundRef) –Funder Registry is an open registry of grant-giving organization names and identifiers, with 32,000 unique identifiers for funders. It is donated by Elsevier to CrossRef and is updated approximately every 4-6 weeks. The Funder Registry ID can be used for author affiliations where the funder and affiliation are one and the same.

2. NIH should mandate that grant applicants include one or both of the following individual identifiers for all researchers in grant applications, and NIH should make metadata fields available for both.

a. ORCID- ORCID, which stands for Open Researcher and Contributor ID, is a global, not-for-profit organization sustained by fees from member organizations. ORCID is the most broadly adopted identifier system for individuals in scientific publishing.

b. ISNI- While not as well adopted as ORCID in research and science, ISNI has been broadly adopted in adjacent and non-adjacent fields.

3. NIH should mandate that appropriate PIDs be used at each stage reporting, while remaining flexible as to which PIDs it mandates, and should reevaluate its mandated PIDs on an ongoing basis. New PIDs such as RAiD (Research Activity Identifier) and DataCite (DOI-based system for research outputs) are being developed regularly and can help connect people, places and research. Likewise, other existing PIDs such as, e.g., Scopus Affiliation ID (AF-ID) and Author ID (AU-ID) are currently used in certain relevant applications. Appropriate PIDs should be mandated at each stage of the workflow, while recognizing that the needs of researchers and the availability of PIDs change over time.

As a final recommendation, we suggest that NIH follow the lead of Wellcome Trust and the Bill and Melinda Gates Foundation, among others, in registering grants for DOIs. This will help enable connectivity of PIDs and the discoverability of the grants, maximizing return to US taxpayers.

Respectfully submitted for Copyright Clearance Center by,



Notes on the data quality of bibliographic records from the MEDLINE database

Robin Bramley* Stephen Howe† Haralambos Marmanis†

August 17, 2022

Abstract

The US National Library of Medicine has created and maintains the PubMed® database, a collection of over 33.8 million records that contain citations and abstracts from the biomedical and life sciences literature. That database is an important resource for researchers and information service providers alike. As part of our work related to the creation of an author graph for coronaviruses, we encountered several data quality issues with records from a curated subset of the PubMed database called MEDLINE. We provide a data quality assessment for records selected from the MEDLINE database and report on several issues ranging from parsing issues (e.g., character encodings and schema definition weaknesses) to low scores against several data quality metrics (e.g., identifier completeness, validity, and uniqueness).

1 Introduction

PubMed is an enormously valuable resource for the biomedical and health fields. The PubMed database is a voluminous collection of medical literature citations that is free, easily accessible, and has been a data source for many works in the information retrieval and life sciences communities. As machine learning becomes more prevalent in various branches of the life sciences, the number of works that rely on the PubMed database increases. Many papers that cited PubMed have appeared within the proceedings of The International Conference on Data and Text Mining in Biomedicine series e.g., DTMBIO '10 [1]. In ACM's Digital Library[2], the year 2021 was a new high point at 235 for computing research articles that mentioned PubMed in the full-text collection, up from 1 in 1998 and 115 in 2010. Many information providers utilize the PubMed database, and there are a variety of machine learning models trained on PubMed[3]. It should be no surprise that, during the COVID-19 pandemic,

*Copyright Clearance Center Limited, London, United Kingdom

†Copyright Clearance Center Inc., Danvers, Massachusetts, United States of America

the PubMed database has been crucial in providing timely and frictionless access to the scientific literature[4].

However, the PubMed database, which contains over 33.8 million records [5] collected over many decades, suffers from several data quality issues. These issues relate to, in part, character encodings, the absence of persistent identifiers, differences in human languages, and schema changes. These shortcomings should not be surprising since PubMed aggregates information produced by different publishers and XML providers, a fact that leads naturally to the presence of “multi-source problems” [6].

MEDLINE is a curated subset of PubMed, its records are indexed with a controlled vocabulary called MeSH [7] and include information regarding funding, genetic, chemical, and other metadata. Articles in MEDLINE predominantly come from a set of indexed journals and a reference data file of these journals is available separately [8]. MEDLINE was made available online, through PubMed, in 1997.

In this article, we will provide an account of our experience in working with the curated MEDLINE records and report on the data quality issues that we encountered. We will describe, at length, the problem of Author Name Disambiguation, which is widely acknowledged as a source of errors when processing bibliographic databases in general, due to the challenges of synonyms (e.g., “John Doe”, “John T Doe”, and “JT Doe” referring to the same individual) and homonyms (i.e., two different people who share the same name such as “John Smith”) [9]. Other problem areas that we will discuss include issues with character encodings, date related issues, the presence of persistent identifiers (and lack thereof), affiliation disambiguation, language related data issues, and schema data quality issues. Knowing how to address these challenges is valuable for practitioners who need to work with MEDLINE (or databases like MEDLINE) and process its records so that they can be used in their information systems.

1.1 PubMed data

The PubMed database is available as XML, based on a DTD (currently the 2019 version) [10]. The compressed files are made available via an FTP server (they are also accessible by HTTPS) and each one of them contains up to 30,000 citation records. Every year, in mid- December, the data are consolidated and an annual baseline is produced. This is followed by incremental daily update files that include deletions.

A PubMed XML file has a root element of `PubmedArticleSet` that contains 1, or more, `PubmedArticle` or `PubmedBookArticle` children. The DTD also permits 0 or 1 `DeleteCitation` elements, and these can be seen in the update files. The elements of the `PubmedArticle` are divided into the `MedlineCitation` and the optional `PubmedData` - we have colloquially referred to these as the “front” and “back” matter respectively.

The description of the XML elements [11], also outlines potential discrepancies caused by schema changes, or policy changes to the collected data. For

example, records created before 2002 only contained author initials instead of full, first or middle, names; moreover, records between 1988 and 2013 only included the affiliation for the first author.

1.1.1 Known DTD shortcomings

There are two known problems with the DTD that have not yet been addressed. The first known problem is that authors cannot be linked to their `CollectiveName`. Some publishers have tried to work around this by interspersing `CollectiveName` elements and `Author` elements. In a wheat genome sequencing consortium paper (PMID 30115783), one of the contributors was a member of 12 groups, so that person appears as an `Author` record 12 times. This multiplicity complicates the author name disambiguation, as it may be impossible to distinguish a duplicate author entry from a valid homonym.

The second problem is related to a shortcoming in the 2019 DTD. Specifically, the back matter `PubMedData` element may contain a `ReferenceList` with many `Reference` elements, but it doesn't prevent the presence of many `ReferenceList` elements each with one `Reference`. Consequently, extraction must be able to handle both because both have been observed in the records. Furthermore, the `ReferenceList` definition permits deeply nested `ReferenceList` elements, as shown below:

```
<!ELEMENT ReferenceList (Title?, Reference*, ReferenceList*) >
```

1.1.2 Escape characters

Escape sequence characters may appear within text fields such as the article title or abstract text. For example, if you wanted to represent a record in JSON, then you would have to beware of trailing backslashes and double quotes. Backslashes can also be problematic for the language used to parse the record. Furthermore, it may be necessary to remove other special characters such as new line characters (e.g., carriage return, line feed), tabs, and so on.

1.1.3 Extended characters

PubMed encompasses articles published in many different languages, sometimes multiple languages. Consequently, fields such as the affiliation string, or parts of the author's name, may contain extended characters. This is an important consideration for the disambiguation of author names.

1.2 Open Source libraries

Since PubMed has been a canonical source of biomedical citations, there are open source libraries to assist with parsing the records. Whilst none of these libraries were appropriate for our needs, they are included here for completeness.

For Python, `pubmed_parser` [12] is an active project, but only handles a constrained field list. The `pymed` [13] project, which is now archived, only

parsed and cleansed a limited subset of the fields. It also seems that the design was intended to wrap the API.

For Java, there is `pubmed-parser` [14], which is based around the Java Architecture for XML Binding (JAXB). This project only had a short flurry of commits over 6 days in April 2021, consequently it is unclear whether this is actively maintained.

2 Materials and Methods

This work will identify challenges that can be faced when working with the MEDLINE data and categorize them along several dimensions of data quality [15].

2.1 Data acquisition

The PubMed baseline files were downloaded from their respective NLM FTP folders [16][17] and uploaded to separate folders on an S3 bucket.

2.2 Data processing

Figure 1 illustrates our data processing approach. The PubMed gzipped XML files were processed using Apache Spark 3.1.1 on Amazon EMR 6.3.1. The initial ingestion process extracted a few key properties, such as the PMID and DOI (from the `PubmedData` if present), before splitting the XML into two fragments representing the front matter (bibliographic metadata) and the back matter (references).

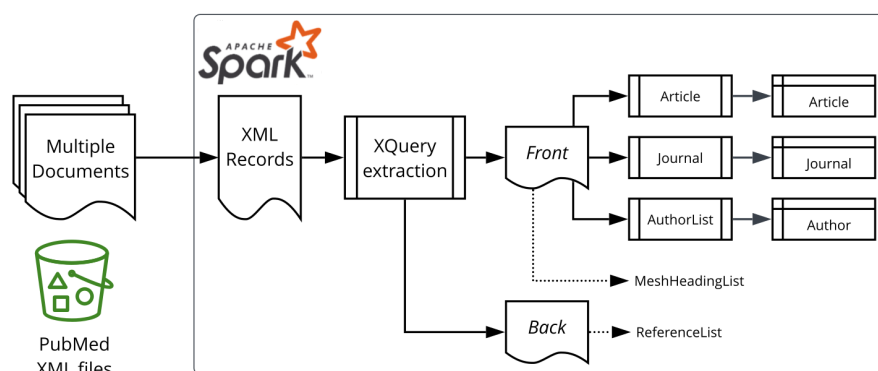


Figure 1: Data processing overview.

The baseline files were ingested first, then the update files were subsequently processed to apply updates, inserts and deletions. Record updates were applied

by sorting the records by their PMID in conjunction with the `DateRevised` property; only the newest records were retained. Note that the PMID Version attribute is not suitable for this purpose as it is only used by Public Library of Science (PLOS) records [11].

Spark SQL [18] is designed for tabular data, with the key construct being the `DataFrame`. Whereas XML documents are represented using a hierarchical structure that allows for repeating elements (a one-to-many relationship). This leads to an inherent mismatch between the two data formats that requires data transformation.

There is a `spark-xml` module [19], but we discovered during our initial experiments that the PubMed XML was too complex for `spark-xml`, as it resulted in heavily nested `DataFrames`, and incorrect query results. Consequently, we solved the XML to `DataFrame` impedance mismatch by performing an XQuery [20] operation per target entity type (e.g. Article, Author, etc.) as shown on the right-hand side of Figure 1.

The `spark-xml` `XmlInputFormat` class was retained for loading the XML files into Spark, with the ingestion and extraction utilizing XQuery queries to extract properties, via the Saxon-HE [21] library as provided by the Elsevier Labs `spark-xml-utils` [22] module.

To ease maintenance of the complex XQuery queries, we adopted a pattern whereby the XQuery output produces a JSON document. This makes the target property for a particular XPath or XQuery expression transparent (Figure 2) and inserting new elements does not break downstream code because it does not rely on positional information. The last part of that transformation phase is to leverage the `read` method of the `SparkSession` object which parses the JSON documents to `DataFrame` records. Note that Figure 2 also represents the handling of escape characters using the XQuery `replace` function.

```
""forename"": "", replace(replace($x/ForeName, '\\', '\\\\'), '&quot;', '\\&quot;'), '"', ', ',
""initials"": "", replace(replace($x/Initials, '\\', '\\\\'), '&quot;', '\\&quot;'), '"', ', ',
""lastname"": "", replace(replace($x/LastName, '\\', '\\\\'), '&quot;', '\\&quot;'), '"', ', ',
""suffix"": "", normalize-space(replace(replace(replace($x/Suffix, '^[,\\. ]+', ''), '[,\\. ]', '\\&quot;'), '\\&quot;', '\\'), '\\&quot;'), '"', ', ';
```

Figure 2: JSON representation within XQuery.

2.3 Data analysis

The resulting `DataFrames` were analyzed using Spark SQL in Apache Zeppelin [23]. For string fields, we consider the length in characters and in words (by splitting on spaces). Metrics were rounded to 3 decimal places (or less).

The plots were produced in R, with the box plots using log-scale for the y-axis.

2.4 Definitions

- \mathcal{N} = number of records
- \mathcal{M} = number of records missing a value for the target property
- \mathcal{D} = distinct values of those present (excludes null / blank)
- \mathcal{V} defined by count of records matching a regex for identifiers (Table 1)
- \mathcal{P} = present = $\mathcal{N} - \mathcal{M}$
- *Completeness* metric = $\mathcal{P} / \mathcal{N}$
- *Validity* metric = $\mathcal{V} / \mathcal{P}$
- *Uniqueness* metric = $\mathcal{D} / \mathcal{P}$

Identifier	Regular expression
DOI [24]	"^10.\d{4,9}/[-. ; ()/:a-zA-Z0-9]+\$" ¹
ORCID [25]	"^\d{4}-\d{4}-\d{4}-\d{3}X\d{4}"
ISNI [26] (presentation)	"[0-9]{4} [0-9]{4} [0-9]{4} [0-9]{3} [0-9X]"
ISNI (compact)	"[0-9]{15} [0-9X]"
GRID [27]	"grid\.\d{4,6}\.[0-9a-f]{1,2}"

Table 1: Regular expressions for identifier validation

2.5 Limitations of the study

The source dataset comprises the PubMed 2022 baseline plus daily update files to 1252 (30th March 2022).

It should be noted that our study includes only the `PubMedArticle` records, not the `PubMedBookArticle` records. The `PubMedArticle` records are only those from the MEDLINE subset (based on the `Status` attribute), and further excludes news articles, and those articles without a title; this gives a total of 28,986,590 article records. News articles were excluded from extraction because journalists, anecdotally those from the British Medical Journal, skew attempts to identify prolific authors through aggregation.

Other applied constraints are as follows:

- Only `Author` records with the `ValidYN` attribute of `Y` have been extracted, not `Investigator` records. For these 120,191,520 authors, only the first `Affiliation` element is considered.
- The `DataBank` element provides links to external datasets such as clinical trials. These identifiers were not investigated as part of the reported study.

¹Adapted from <https://www.crossref.org/blog/doi-and-matching-regular-expressions/>

- For alternative article identifiers, we did not extract the ELocationID element nor Publisher Item Identifiers (PII) from the PubmedData.
- For Journals, ISSNs were not analyzed.

2.5.1 Approximation

Five number summary information is produced using Spark's `DataFrameStatFunctions` `approxQuantiles` method with an error margin of 0.0001, an example is shown below:

```
articleDF.stat.approxQuantile("doi_len", Array(0.0,0.25,0.5,0.75,1.0), 0.0001)
```

However, the distinct counts do not leverage the Spark SQL `approx_count_distinct` function, rather the `dataframe.select("column").distinct.count` approach was used.

3 Results and discussion

In this section, we'll present our results related to data quality for the entities and fields shown in Figure 3. The PubMed XML data model is article-centric, but we will work our way from left to right.

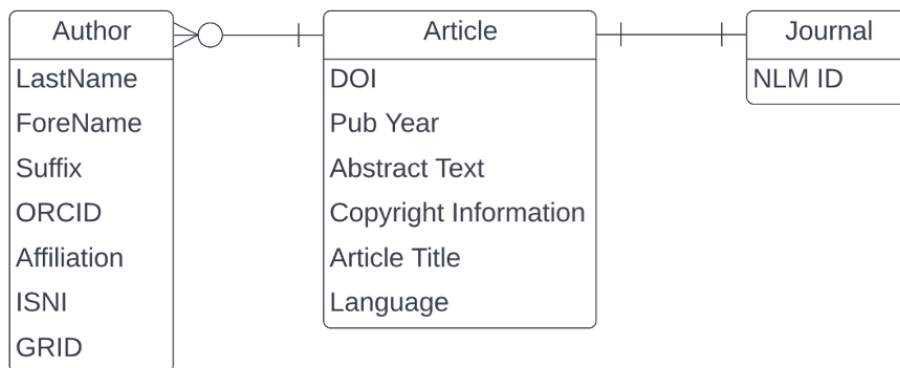


Figure 3: Entity Relationship Diagram for a subset of PubMed.

3.1 Data quality issues related to author names

One of the important considerations regarding author records is that PubMed has not always recorded all the authors of a paper. The number of authors was limited to 10 between the years 1984 and 1995, and to 25 between the years 1996 and 1999 [11].

The most common last names in MEDLINE are Romanized Chinese names (Table 2), which can be very challenging to disambiguate. Looking at the length

characteristics (Figure 4), there are a few obvious problems, namely pollution of the author elements by incorrectly entered collective names (Table 3), and single character last names potentially caused by name transposition errors (Table 4).

LastName	Occurrences
Wang	1,086,073
Li	895,976
Zhang	878,544
Chen	722,753
Liu	703,743
Lee	547,636
Kim	523,687
Yang	433,439
Wu	360,532
Huang	309,375

Table 2: Top 10 LastName values.

LastName	Length
Endocrinology Genetics And Metabolism Group Pediatric Branch Of Chinese Medical Association Neonatal Screening Group Specialist Committee For Prevention And Control Of Birth Defects Chinese Association Of Preventive Medicine Prevention And Control Committee Of Birth Defects Pediatric Branch Of Chinese Medical Association	322
The Group Of Minimally Invasive Spinal Surgery And Enhanced Recovery Professional Committee Of Orthopedic Surgery And Enhanced Recovery Association Of China Rehabilitation Technology Transformation And Promotion	211
Genetic Disease Society Guangdong Precision Medicine Application Association Prenatal Diagnosis Group Maternal And Child Health Care Society Guangdong Medical Association Expert Committee Of Prenatal Diagnosis	209
Arir Associazione Riabilitatori dell'Insufficienza Respiratoria Sip Società Italiana di Pneumologia Aifi Associazione Italiana Fisioterapisti And Sifir Società Italiana di Fisioterapia E Riabilitazione	201
This Paper Is A Co-Publication Between European Journal Of Preventive Cardiology European Heart Journal Acute Cardiovascular Care And European Journal Of Cardiovascular Nursing	176
Committee For Birth Defect Prevention And Control Chinese Association Of Preventive Medicine Genetic Testing And Precision Medicine Branch Chinese Association Of Birth Health	174
Consensus Group Of Experts On Application Of Metagenomic Next Generation Sequencing In The Pathogen Diagnosis In Clinical Moderate And Severe Infections	152
Expert Committee Of The Inter-Laboratory Quality Assessment Of Prenatal Screening And Diagnosis Clinical Test Center Of The National Health Commission	150
For The Antimalarial Therapeutic Efficacy Monitoring Group National Malaria Elimination Programme The Federal Ministry Of Health Abuja Nigeria	142
On Behalf Of The Association Of Rural Surgeons Of India-Lancet Commission On Global Surgery Consensus Committee Arsi-LCoGS Consensus Committee	142

Table 3: Ten longest LastName values.

LastName	Occurrences
S	756
A	704
E	636
M	592
O	563
K	497
R	453
P	363
G	306
V	279

Table 4: Top 10 shortest LastName values.

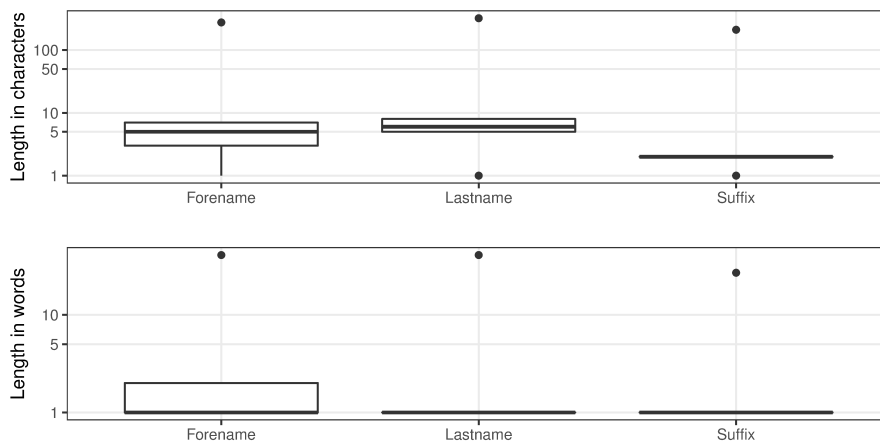


Figure 4: Author name character / word distributions.

The author forename field is 99.913% complete. Regarding the length, before 1945, the longest value in the forename field was 3 characters long, which reflects the policy to only hold author initials. The distributions, in Figure 4, clearly show that there are outliers. As shown in Table 5, these are primarily for working groups (a validity error), but the first row represents a different form of data preparation error where the affiliation has been concatenated with the forename.

PMID	LastName value	ForeName value	Length
34313229	Choi	Moon Hyung Department Of Radiology Eunpyeong St Mary's Hospital College Of Medicine The Catholic University Of Korea Seoul Republic Of Korea Catholic Smart Imaging Center Eunpyeong St Mary's Hospital College Of Medicine The Catholic University Of Korea Seoul Republic Of Korea	276
33145749	En Representación Del Grupo de Trastornos de la Conducta Y Del Movimiento Durante El Sueño de la Sociedad Española de Sueño	En Representación Del Grupo de Trastornos de la Conducta Y Del Movimiento Durante El Sueño de la Sociedad Española de Sueño	123
32329046	En Representación Del Grupo de Estudio de Enfermedades Desmielinizantes de la Comunidad Autónoma de Madrid	En Representación Del Grupo de Estudio de Enfermedades Desmielinizantes de la Comunidad Autónoma de Madrid	106
32433836	Pharmakopsychiatrie	The Therapeutic Drug Monitoring Task Force Of The Arbeitsgemeinschaft Für Neuropsychopharmakologie Und	102

Table 5: ForeName values over 100 characters.

Completeness does not apply to author suffixes since not everyone has a suffix to their name. In terms of uniqueness there are 823 distinct values across 483,541 entries. There are also consistency issues, examples of which can be observed in Table 6 (e.g., Jr, Junior, Júnior). Figure 4 shows the range of suffix lengths and clearly indicates that there is something wrong with at least some records. When we look at the longest values for author suffixes (Table 7) and the most common single character values (Table 8), it becomes clear that there are multiple data issues related to the author suffix field; the general theme of misplaced values, or value “pollution”, occurs across fields and is a major data quality weakness for the MEDLINE records.

Suffix value	Occurrences
Jr	374,510
3rd	74,260
2nd	20,364
4th	5,828
Sr	4,075
Junior	535
Júnior	380
Filho	241
PhD	238
5th	204
Neto	200
III	199
Dr	146
6th	129
MD	99

Table 6: Top 15 suffixes.

Suffix value	Length
Brian Buckley Caitlin Cornell Alyssa Fuller Eric Hojnowski Ryan LaFollette Yelena Livshits Todd Michaelis Claire Motyl Tarakad Ramachandran Devan Rahmachandrin Sofia Seckler Evaline Tso And Kate Zmijewski-Mekeem	211
European Society Of Clinical Microbiology And Infectious Diseases Escmid Vaccine Study Group Evasg	98
(Conceptualization; Review and editing; Read and approved final version of manuscript)	86
Faculty of Bioscience and Bioindustry, Tokushima University, Tokushima, Japan	77
BA, MBBS (Hons), FRANZCP, PhD, Dip Psychodynamic Psychotherapy, Cert ATP	72
on behalf of the Portuguese visual impairment study group (PORVIS-group)	72
(Writing original draft; Read and approved final version of manuscript)	71
RN, Cert Psych Nurs, BA (Hons), Dip Ed, B Ed, M Ed, PhD, FACMHN	63
DVM, PhD, Diplomate ABVP (Dairy Practice), SFHEA, NVS, MRCVS	60
B Phil (Hons), B Soc & Comm Stud (Community Development)	60

Table 7: Ten longest suffixes.

Suffix value	Occurrences
*	32
S	12
K	11
W	11
J	8
F	8
†	8
A	7
P	7
M	5

Table 8: Top 10 shortest suffixes.

The PubMed DTD does not have a dedicated field for an email address. From 1996, NLM included “the first author’s electronic mail (e-mail) address at the end of <Affiliation>, if present in the journal. Furthermore, as of October 1, 2013, NLM no longer edits affiliation data to add e-mail address” [11]

A word of caution about relying on email addresses as a discriminator for author name disambiguation; the most common email address is user@example.com which occurred 2023 times in the MEDLINE dataset of this study. Additionally, there are other non-specific email addresses such as journal editorial mailboxes.

Since 2010, the PubMed DTD has included an **Identifier** element, which has been used from 2013 [11]. However, it has less than 3% completeness (Table 9) and it is worth noting that there are occurrences where the same ORCID identifier has been incorrectly allocated to multiple authors within a paper.

Identifier	Completeness	Validity	Uniqueness
ORCID	2.820%	99.915%	40.921%

Table 9: Author ORCID measures.

3.2 Data quality issues related to affiliation names

An author’s institutional affiliation is a very important information field, but the completeness is only around 42%. We have not derived a validity score, but there are quality problems within that set that are obvious from the length distributions (Figure 5). As previously mentioned, this field may contain values that aren’t written in English as well as non-ASCII characters.

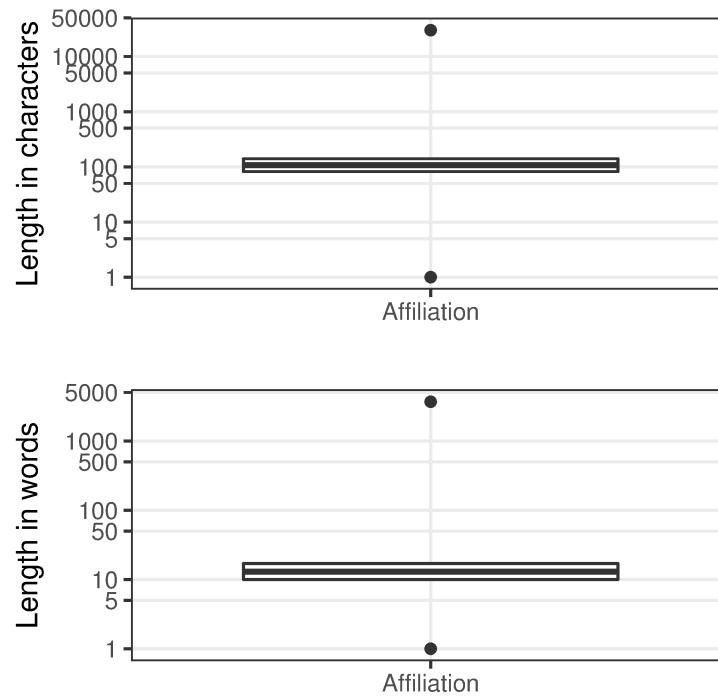


Figure 5: Affiliation character / word distributions.

In Figure 5, the outliers at the top of the range, which we have termed “narrative affiliations”, typically describe the affiliations for many, if not all, of the contributors to the paper (e.g., see Figure 6 where we show the entry from the article with PMID 32308221). These narrative affiliations may also be repeated for all the author entries within the author list. At the other end of the range, there are many incomplete, or indistinguishable entries (Table 10).

Affiliations

- 1 Amy Meyer, is at the University of Missouri School of Medicine, Columbia, Missouri. Hariharan Regunath, MD, MSMA member since 2019, is in the Department of Medicine, Division of Pulmonary, Critical Care and Environmental Medicine, and Division of Infectious Diseases, University of Missouri, Columbia, Missouri.
- 2 Christian Rojas-Moreno, MD, William Salzer, MD, and Gordon Christensen, MD is in the Department of Medicine, Division of Infectious Diseases, University of Missouri, Columbia, Missouri.

Figure 6: An example of narrative affiliations.

Affiliation string	Occurrences
.	5,761
,.	2,463
London, UK.	601
Editor-in-Chief.	468
London.	405
Pathology.	360
GSK, Siena, Italy.	342
Duke University.	341
Harvard University.	332
McGill University.	329
Paris, France.	323
School of Medicine.	303
Yale University.	301
Editor.	295
Radiology.	262

Table 10: Top 15 affiliations under 20 characters long.

Our parsing has not included any special case exclusions. We note that `pubmed_parser` [12] excludes “For a full list of the authors’ affiliations please see the Acknowledgements section.” - though this exact string only occurs once within our selected dataset of over 51 million affiliation strings! It should also be noted that “as of October 1, 2013, NLM no longer performs quality control of the affiliation data” [11].

Whilst multiple affiliations were possible from the 2015 DTD [11], this is a good place to mention how some data providers concatenate multiple affiliations for an author in a single element. Here is an example for Yong-Beom Park (PMID 29465366):

Division of Rheumatology, Department of Internal Medicine, Yonsei University College of Medicine, Seoul; and Institute for Immunology and Immunological Diseases, Yonsei University College of Medicine, Seoul, Republic of Korea.

Affiliation identifiers, such as ISNI and GRID, were possible from the 2015 DTD [11]. We’ve captured values for those too in Table 11.

Identifier	Completeness	Validity	Uniqueness
ISNI	0.002%	99.965%	22.803%
GRID	0.003%	100.000%	23.752%
Affiliation	42.526%	N/A	45.979%

Table 11: Key measures for Affiliations / Affiliation identifiers.

3.3 Data quality issues related to articles

3.3.1 Article persistent identifiers

As can be seen in Table 12, the application of digital object identifiers (DOI), although not perfect, reaches a respectable score in terms of uniqueness but there are issues with validity of those identifiers and a significantly low score in terms of completeness; we'll examine the impact that earlier publications have on DOI completeness.

Identifier	Completeness	Validity	Uniqueness
DOI	71.373%	99.377%	99.949%

Table 12: MEDLINE article identifiers.

3.3.2 Publication year

In the full PubMed database, there are over 100,000 records with a publication year earlier than 1900. In our selected data set from MEDLINE, there are only 3 that are clearly wrong (Table 13). In the first two examples, the publication year has the upper value from the journal pagination range. These erroneous publication years caused Parquet compatibility problems with Spark 3 (see issue SPARK-31404: <https://issues.apache.org/jira/browse/SPARK-31404>) when constructing a Date column, as they pre-date the introduction of the Gregorian calendar in 1582 and Spark implements a Proleptic Gregorian calendar as of version 3.

PMID	Publication Year
11662976	1132
11665278	1041
32422596	1

Table 13: Example of erroneous publication year values.

Figure 7 illustrates the volume of citation records with a valid DOI per publication year with 2022 in progress. Note that as of Q1 2022 there are not yet articles scheduled for publication in subsequent years.

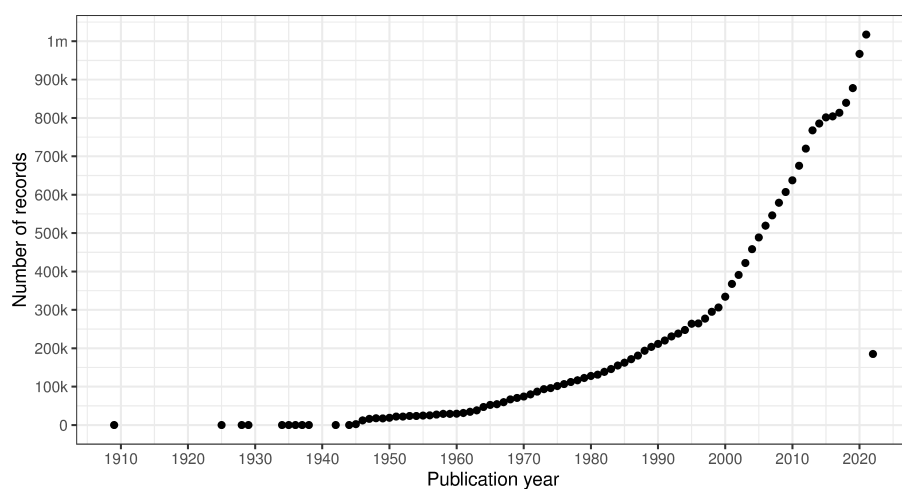


Figure 7: Count of citation records with a valid DOI per publication year (excluding erroneous years).

3.3.3 Abstract

The abstract field was added to the PubMed record in 1975 [11]. The abstract text, which may be subject to copyright restrictions, is a prime candidate for text mining. Consequently, for the two-thirds of records with an abstract, it's useful to understand their length distribution (Figure 8) and the erroneous values that they contain. Whilst the uniqueness is 99.942%, there is still a significant number (over 11 thousand abstracts) with non-unique abstract values. From the length information, we can infer that there are clearly meaningless abstract entries towards the lower end of these ranges, as seen in Table 14.

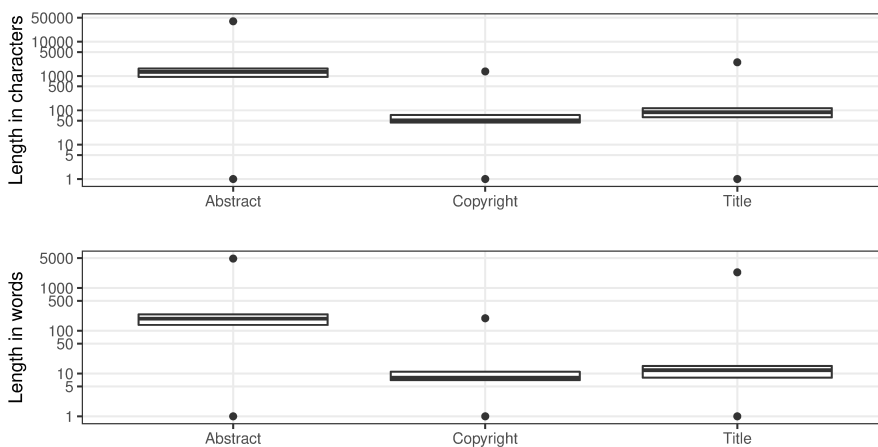


Figure 8: Article character / word distributions.

Abstract text	Occurrences
[Figure: see text].	579
.	182
Not available.	106
N/A.	51
n/a.	50
no summary.	48
Null.	41
NA.	29
No Abstract.	22
<p/>.	20
Editorial.	17
EDITORIAL.	16
	13
No abstract.	13
None.	10

Table 14: Top 15 abstracts under 20 characters long.

3.3.4 Copyright

An important consideration when mining MEDLINE should be whether copyrighted material is being used. The NLM terms and conditions clearly state that they do not provide legal advice [28]. The copyright information field was introduced in 1999 [11], with a completeness measure of almost 22% of the records that have an abstract. From Table 15, it is evident that Elsevier is

most consistent in supplying copyright statements although there is some lack of consistency regarding the actual values. Figure 8 shows the distributions of character length and word tokens, it should be clear that at the low end of the range there must be some invalid values (Table 16).

Copyright information	Occurrences
Copyright © 2020 Elsevier Inc. All rights reserved.	40,773
© 2021. The Author(s).	39,577
Copyright © 2020 Elsevier B.V. All rights reserved.	39,221
Copyright © 2020 Elsevier Ltd. All rights reserved.	39,220
Copyright © 2016 Elsevier Inc. All rights reserved.	38,600
Copyright © 2019 Elsevier Inc. All rights reserved.	37,672
Copyright © 2018 Elsevier Inc. All rights reserved.	37,414
Copyright © 2017 Elsevier Ltd. All rights reserved.	36,833
Copyright © 2017 Elsevier Inc. All rights reserved.	36,817
Copyright © 2018 Elsevier Ltd. All rights reserved.	36,766

Table 15: Top 10 copyright statements.

Copyright information	Occurrences
© 2013.	6,941
excerpt	4,996
© The author(s).	3,193
© FASEB.	1,444
full text	1,238
©2011 AACR.	1,159
©2013 AACR.	1,145
©2012 AACR.	958
Celsius.	956
© 2017 The Authors.	925

Table 16: Top 10 short copyright statements.

3.3.5 Title

MEDLINE has just over 7,500 records without an `ArticleTitle` element, leading to a completeness value of 99.974%. The uniqueness of the title field is approaching 98%. Like our observations for the abstracts, there are standard article titles that relate to the publication type towards the lower end of the character length and number of word token ranges (Figure 8; see also Table 17).

Article title	Occurrences
[Not Available].	13,440
Reply.	1,972
Invited commentary.	1,896
Editorial comment.	1,676
Editorial.	1,465
Response.	1,312
Discussion.	1,052
Editorial Comment.	1,051
Preface.	974
The authors reply.	768
In reply.	714
Introduction.	585
In Reply.	519
Authors' response.	469
Foreword.	428

Table 17: Top 15 article titles under 20 characters long.

3.3.6 Language

Another important consideration for text mining is the language, or languages, that the article is published in. It should be noted that PubMed includes translated titles, in square brackets, where appropriate. The language element contains language codes from the US Library of Congress MARC [29] schema, such as “*chi*” for Chinese. The language code table [30] includes “*und*” for undetermined and “*mul*” for multiple languages. However, language codes can also be concatenated together; for example, “*fregerita*” means the article was published in French, German, and Italian.

The language field is complete for the entirety of the MEDLINE records, but if we treat a solitary value of “*und*” or “*mul*” (238,470 and 1,399 occurrences, respectively) as invalid then the validity of this field is 99.55%. This excludes cases where they are present with other values too. From a recency perspective, “*und*” last occurred in 2002, and that is the only occurrence since 1985; “*mul*” occurred once in both 2016 and 2015, but before that it was last seen in 2011.

The maximum number of languages specified for a record is 6, but the 75th percentile is 1. Considering the values individually by splitting the strings and exploding the resulting array, allows us to produce the top 10 languages (Table 18). Note that almost 84% of records within the MEDLINE sample are published in English. The next most common language, German, only accounts for about 3% of articles.

Language code	Occurrences
eng	24,290,379
ger	861,109
fre	744,111
rus	697,806
jpn	429,283
spa	364,920
chi	329,153
ita	305,526
und	239,588
pol	172,956

Table 18: Top 10 languages.

3.4 Data quality issues related to journals

The key identifier provided in MEDLINE for a journal is the US National Library of Medicine (NLM) identity. When compared to the J_MEDLINE reference data set of MEDLINE indexed journals [8], the NLM identifiers have a referential integrity [15] measurement of 99.989%. There were 146 NLM identifiers that were not included within the J_MEDLINE dataset, affecting 3,045 articles. When considering a graph representation of the dataset, this would result in dangling edges that may not be permitted by some graph storage engines, such as Neo4j.

3.5 Data quality issues related to time evolution

In this section we consider the change over time for some of the key identifiers. Are there any obvious trends in whether identifiers are becoming more pervasive or prevalent in newer citation records? Here are some general observations: DOIs are almost ubiquitous for new articles (Figure 9), ORCIDs have been on the rise to just under 17% of authors per year (Figure 10), but GRID and ISNI usage peaked in 2017, having first appeared in 2015 (Figure 11). That leaves us with the tedious task of disambiguating the affiliation of the authors in the records. As can be seen in Figure 12, the vast majority of recent records contain an affiliation string for all authors; this is due to a policy change in 2014 to collect affiliations for all contributors [11].

4 Conclusions

PubMed is an enormously valuable resource for the biomedical sciences and healthcare, yet, those attempting to identify authors and affiliations, or otherwise use the records from that database, need to be aware of the quality issues within the dataset. This article has highlighted some of those data quality concerns.

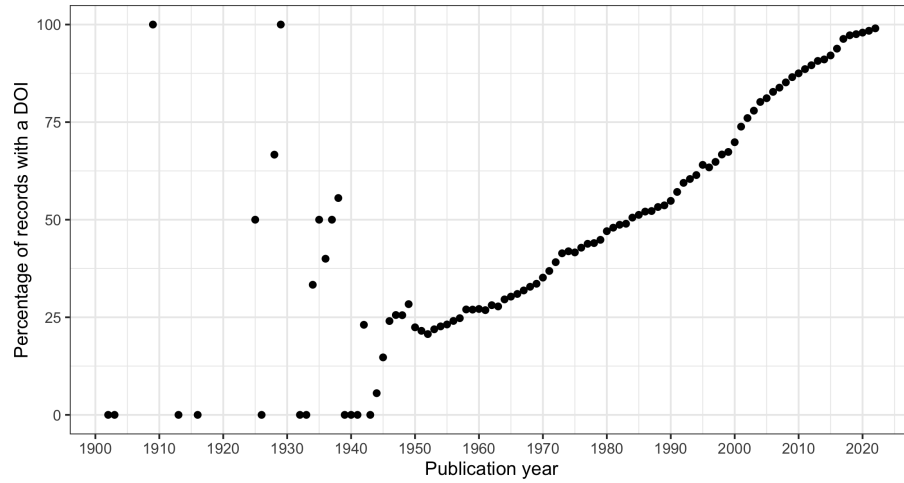


Figure 9: DOI percentage of articles per publication year.

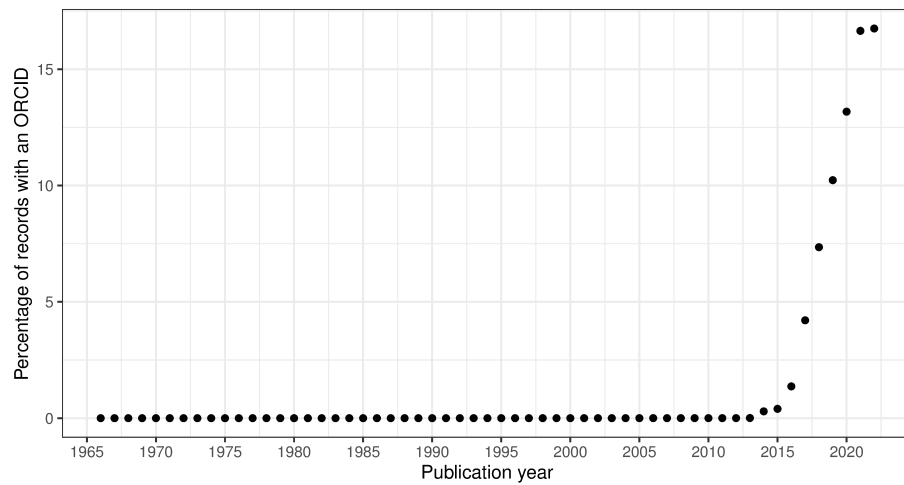


Figure 10: ORCID percentage of authors per publication year.

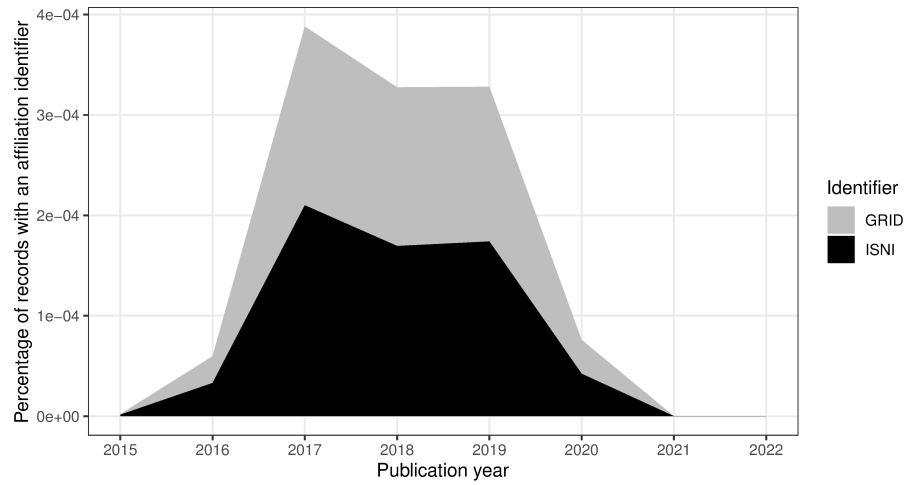


Figure 11: ISNI & GRID percentage of authors per publication year.

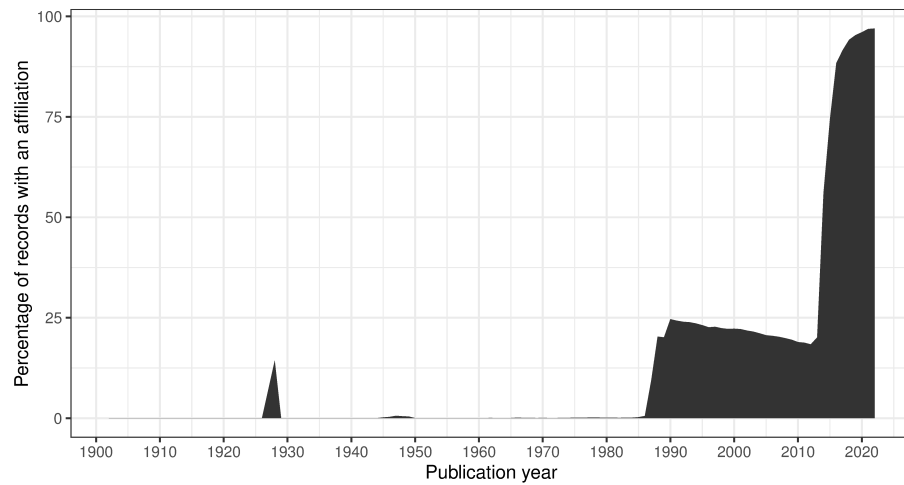


Figure 12: Percentage of authors per publication year with an affiliation string.

The data are subject to many human errors, such as typographical errors, and system related errors such as inconsistent representations of author names (leading to the synonym problem) and affiliations. There is a lack of author identifiers (contributing to the homonym problem) and a significant lack of affiliation identifiers. Being an aggregated source, the PubMed database suffers from multi-source problems such as inconsistent representations from the upstream XML providers that result in a high degree of lexicographic entropy.

In summary, our work supports the following conclusions:

- Given the incompleteness and uniqueness of identifying fields, the disambiguation of author names remains a significant problem for PubMed, particularly for records dating before 2014.
- PubMed has excellent integrity for NLM-internal identifiers (e.g., MeSH), though there is the noted exception around the J_MEDLINE dataset. Beyond the NLM database, the majority of articles are labelled with a DOI, and the DTD provides support for identifiers for authors, institutions, both of which are far from complete. The DTD also caters for grant information, and auxiliary data through the DataBank elements, though these were beyond the scope of our work.
- Overall, there is an improvement in the use of identifiers; in particular, records created since 2015 exhibit an increase in external identifiers. However, the data quality for institutional identifiers is poor and their use has been diminishing over time.

Unless the data quality issues are addressed retroactively, they will weaken (if not entirely distort) any subsequent data analysis. Perhaps, an intervention in current publishing systems, to prevent the data sources of PubMed from manifesting the data quality issues mentioned herein, is the best one can hope for the future. Much like the application of machine learning has been applied within the NLM for indexing (e.g., with the MTI tooling [31]), the NLM could enhance their process with systems that possess a learning architecture to improve and accelerate the curation of the PubMed records. It is also possible that another information provider will provide an open data repository containing cleansed PubMed data, although a proprietary offering is more likely.

Another possibility for better use of the PubMed treasure trove is the creation of an open source library for cleansing the data, or at least properly identify the data quality issues, and optimize the amount of information that one can obtain from processing the PubMed records. Once this is accomplished with one programming language the open source community can augment the library and expand its adoption in other programming languages, for example by porting the library.

Lastly, the community would benefit from the availability of open source libraries that can accurately perform author name disambiguation, or a substantial set of “gold data” that can be used for training and validation; that

dataset, however, should be orders of magnitude larger than the ones that are currently available (e.g., the ‘amorgani/AND’ dataset [32] [33]).

References

- [1] *DTMBIO '10: Proceedings of the ACM fourth international workshop on Data and text mining in biomedical informatics*. URL: <https://dl.acm.org/doi/proceedings/10.1145/1871871>.
- [2] *ACM Digital Library*. URL: <https://dl.acm.org>.
- [3] Lee J et al. “BioBERT: a pre-trained biomedical language representation model for biomedical text mining.” In: *Bioinformatics* 36.4 (2020), pp. 1234–1240. DOI: 10.1093/bioinformatics/btz682.
- [4] *LitCovid, NLM*. URL: <https://www.ncbi.nlm.nih.gov/research/coronavirus/>.
- [5] *About PubMed, NLM*. URL: <https://pubmed.ncbi.nlm.nih.gov/about/>.
- [6] Rahm and Do. “Data Cleaning: Problems and Current Approaches”. In: *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering* (2000).
- [7] *MeSH: Medical Subject Headings*. URL: <https://www.nlm.nih.gov/mesh/meshhome.html>.
- [8] *Dataset of MEDLINE indexed journals*. URL: https://ftp.ncbi.nlm.nih.gov/pubmed/J_Medline.txt.
- [9] Sanyal DK, Bhowmick PK, and Das PP. “A review of author name disambiguation techniques for the PubMed bibliographic database”. In: *Journal of Information Science* 47.2 (2021), pp. 227–254. DOI: 10.1177/0165551519888605.
- [10] *PubMed 2019 DTD*. URL: http://dtd.nlm.nih.gov/ncbi/pubmed/out/pubmed_190101.dtd.
- [11] *MEDLINE PubMed XML Element Descriptions and their Attributes*. URL: https://www.nlm.nih.gov/bsd/licensee/elements_descriptions.html.
- [12] Titipat Achakulvisut, Daniel E. Acuna, and Konrad Kording. “Pubmed Parser: A Python Parser for PubMed Open-Access XML Subset and MEDLINE XML Dataset XML Dataset”. In: *Journal of Open Source Software* 5.46 (2020), p. 1979. DOI: 10.21105/joss.01979. URL: <https://doi.org/10.21105/joss.01979>.
- [13] *pymed*. URL: <https://github.com/gijswobben/pymed>.
- [14] *pubmed-parser*. URL: <https://github.com/thecloudcircle/pubmed-parser>.

- [15] DAMA International. *DAMA - DMBOK Data Management Body of Knowledge*. 2nd Edition. New Jersey, USA.: Technics Publications, 2017.
- [16] *PubMed baseline download files*. URL: <https://ftp.ncbi.nlm.nih.gov/pubmed/baseline/>.
- [17] *PubMed daily update files*. URL: <https://ftp.ncbi.nlm.nih.gov/pubmed/updatefiles/>.
- [18] Michael Armbrust et al. “Spark SQL: Relational Data Processing in Spark”. In: *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. SIGMOD '15. Melbourne, Victoria, Australia: Association for Computing Machinery, 2015, pp. 1383–1394. ISBN: 9781450327589. DOI: 10.1145/2723372.2742797. URL: <https://doi.org/10.1145/2723372.2742797>.
- [19] *The spark-xml project*. URL: <https://github.com/databricks/spark-xml>.
- [20] *XQuery 3.0 standard*. URL: <https://www.w3.org/TR/2014/REC-xquery-30-20140408/>.
- [21] *Saxon-HE library, Saxonica*. URL: <http://www.saxonica.com/>.
- [22] *The spark-xml-utils project*. URL: <https://github.com/elsevierlabs-os/spark-xml-utils>.
- [23] *Apache Zeppelin*. URL: <https://zeppelin.apache.org>.
- [24] *Digital Object Identifier (DOI)*. URL: <https://www.doi.org>.
- [25] *Open Researcher and Contributor ID (ORCID)*. URL: <https://orcid.org>.
- [26] *ISO 27729, International Standard Name Identifier (ISNI)*. URL: <https://isni.org>.
- [27] *Global Research Identifier Database (GRID), Digital Science*. URL: <https://www.grid.ac>.
- [28] *NLM terms and conditions*. URL: https://www.nlm.nih.gov/databases/download/terms_and_conditions.html.
- [29] *MARC definition, US Library of Congress*. URL: <https://www.loc.gov/marc/faq.html#definition>.
- [30] *MEDLINE/PubMed Language Table*. URL: https://www.nlm.nih.gov/bsd/language_table.html.
- [31] *NLM Medical Text Indexer (MTI)*. URL: <https://lhncbc.nlm.nih.gov/ii/tools/MTI.html>.
- [32] Vishnyakova D, Rodriguez-Esteban R, and Rinaldi F. “A new approach and gold standard toward author disambiguation in MEDLINE.” In: *J Am Med Inform Assn* 26 (2019), pp. 1037–1045. DOI: 10.1093/jamia/ocz028.
- [33] *Vishnyakova D et al., AND - Author Name Disambiguation corpus*. URL: <https://github.com/amorgani/AND/>.

The State of Scholarly Metadata: 2023



In late 2022, CCC and Media Growth Strategies undertook a thorough examination of metadata management across the research lifecycle.

This in-depth review builds on an existing body of work to uncover multiple policy and system complexities and breakages, which – separately and together – create missed opportunities for the communities for whom Open Access (OA) and Open Science models are designed to serve.

CCC is sharing this analysis with the scholarly communications community to spark dialogue and to drive action. Drawn directly from our research interviews, this living infographic depicts the significant economic and social impact that a fragmented metadata supply chain has today on researchers, institutions, funders, and publishers. Researchers, in particular, shoulder a significant administrative burden that ultimately disrupts and delays the process of scientific discovery.

As the scholarly communications community continues its shift to full OA, stakeholders recognize that new strategies, inclusive policies, and a robust network of interoperable data and systems are essential for making critical infrastructure improvements, and much progress is underway. In that environment, a dedication to data stewardship across each stakeholder group, and the service providers supporting them, will lead not only to a smoother OA transition, but also to greater research integrity; data sharing; reliable, trustworthy metrics on research impact; and a responsive, equitable rewards and recognition system.

<p>Research stage Idea Development</p> <p> RESEARCHER Researcher seeks collaborators; meets with colleagues and library / research office staff</p>	<p>CHALLENGES</p> <p>Underutilization of ORCID Some institutions don't require researchers to use ORCID; records can be outdated if authors don't consistently update; ORCID may not be accessible to authors in some geographies.</p>	<p>↔ IMPACT</p> <p>If authors can't be identified with a standard ID, they may not be able to authenticate to content, get credited appropriately for their work, secure OA funding, or complete downstream processes without unnecessary manual effort. Costly manual effort is also required of publishers, institutions, and funders to disambiguate authors retrospectively.</p>
<p>Research stage Proposal Submission</p> <p> RESEARCHER Researcher submits application for funding</p> <p> FUNDER Funder selects reviewers and begins application review</p> <p> FUNDER Funder logs funding terms in grant management system</p>	<p>CHALLENGES</p> <p>Inconsistent Metadata Capture Variability across grant application process/systems results in possible loss of metadata necessary to determine OA funding entitlements at a later stage, e.g., institutional affiliations.</p> <p>CHALLENGES</p> <p>Legacy System Limitations Low adoption of standardized PIDs (FundRef, RAiD, Ringgold, ISNI, ROR) due to limitations of legacy systems and/or lack of awareness.</p> <p>CHALLENGES</p> <p>Low-Quality Data Free text fields are great for gathering feedback; they're not designed to capture granular data like an organizational identifier. Researchers often confuse proposal numbers with grant IDs later in the publication process--they need structure to improve the accuracy of data capture.</p>	<p>↔ IMPACT</p> <p>Without disambiguated grant and funder details, grants may not be effectively utilized in later publication stages, leaving OA funding unclaimed and shifting coverage to research institutions. In an ecosystem that values a sustainable OA shift, this impacts everyone.</p> <p>↔ IMPACT</p> <p>Hindered conflict of interest management among peer reviewers threatens research integrity, and low-quality data results in low accuracy of later-stage funding identification, tracking, and analysis of research output.</p> <p>↔ IMPACT</p> <p>Lack of registered grant DOIs makes it difficult and costly to link funding to particular research outputs, resulting in missed OA opportunities as well as incomplete analysis to inform future funding investments.</p>
<p>Research stage Research & Authoring</p> <p> RESEARCHER Researcher conducts literature review</p> <p> RESEARCHER Researcher posts pre-print / shares early outputs</p> <p> RESEARCHER Researcher selects publication for submission</p>	<p>CHALLENGES</p> <p>Researcher Inequities & Research Barriers</p> <ul style="list-style-type: none"> Valid research coming from under-represented researchers is hard to find due to lack of metadata, including DOIs. Search and discovery are difficult due to inconsistency in identifying the user and enabling appropriate access to research. Authors from under-represented areas may not have equitable access to search and discovery services or equitable opportunities for publication. <p>↔ IMPACT Global inequities hinder scientific progress.</p> <p>CHALLENGES</p> <p>Poor Connections Across Research Outputs Lack of persistent identifiers (PIDs) and inconsistent application of PIDs across research outputs e.g., data sets, equipment, setting(s), samples, software</p> <p>CHALLENGES</p> <p>Risk of OA non-compliance Metadata lost upstream makes managing funding compliance onerous.</p>	<p>↔ IMPACT</p> <p>Inability to easily find, verify, and reuse the data and artifacts underlying research, making it difficult to accurately interpret, cite and reproduce research findings.</p> <p>↔ IMPACT</p> <p>Lack of available information about both corresponding author and all co-authors leads to manual input to identify funder and institutional mandates at best and missed funding requirements at worst.</p>

Research stage
Publication

RESEARCHER
Researcher submits article

INSTITUTION
Institution funds OA publication

PUBLISHER
Publisher indexes metadata to enable search & discovery

CHALLENGES

Missed Funding Opportunities

- Under-utilization of metadata validation services
- If the researcher has submitted before, outdated information from their existing profile can be pulled into the submission
- Inconsistency between journal policies and metadata procedures
- Lack of funding information captured at submission and validated at acceptance
- Demand for increased interoperability between IDs

CHALLENGES

Missed Funding Opportunities & Costly Billing Complications

If funder/institution information manually input by the author does not use a standardized name or PID (e.g., abbreviations, nicknames), this can interfere with matching to the correct OA funding source.

CHALLENGES

Unnecessary Manual Intervention

Publishers are sometimes manually entering PIDs prior to registering DOIs for a more complete publication record.

IMPACT

Without granular, accurate organizational affiliation identifiers for a manuscript, coupled with incomplete funding details, authors may miss the opportunity to get OA funding or miss the chance to opt into OA due to affordability concerns. OA initiatives driven by institutions and funders may lack uptake as a result. Publishers are also unable to automate processes that reduce the cost of business model transformation. Manual effort is required to retrospectively cover the publication with proper funding sources, driving up the cost of publishing. No one benefits in this scenario.

IMPACT

Publishers and institutions take on the time and expense of manually finding the papers that should have matched to an agreement and collaborating on a resolution. Funding decisions cannot be based on abbreviations or free-form data.

IMPACT

This is a laborious practice with high economic and opportunity costs that could be reduced with earlier, automated PID assertion and/or validation.

Research stage
Reuse & Measurement

RESEARCHER
Researcher evaluates research impact

INSTITUTION
Institution assesses historical subscription & publication data to inform institutional deals

FUNDER
Funder evaluates research impact

PUBLISHER
Publisher assesses historical subscription and publication to inform institutional deals

CHALLENGES

Problematic Research Impact Measurement

Difficult to track research/researcher impact due to lack of adoption of metadata standards.

CHALLENGES

Problematic Deal Modeling

- Lack of consistent affiliation and funding data makes modeling future agreements hard for institutions.
- Data is not standardized across publisher platforms, creating unnecessary manual work to gather and normalize data for analysis.

CHALLENGES

Problematic Research Impact Measurement

Difficult to track funder impact due to lack of adoption of metadata standards.

CHALLENGES

Problematic Deal Modeling

Lack of consistent affiliation and funding data makes modelling future agreements difficult for publishers and institutions.

IMPACT

Researcher rewards and recognition decisions, or future opportunities for funding, may be based on incomplete or inaccurate data, affecting reputation and career advancement.

IMPACT

The transition to modern models of OA publication is delayed, prolonging a mixed-model landscape and the availability of open outputs to advance science.

IMPACT

Incomplete analysis to support future funding investments and to report activities to the public.

IMPACT

The transition to OA is delayed, putting some publishers at risk of losing authors to funding mandates and losing revenue that is necessary to sustain operations.

To view the interactive map, visit stateofmetadata.com

About CCC

A pioneer in voluntary collective licensing, CCC (Copyright Clearance Center) helps organizations integrate, access, and share information through licensing, content, software, and professional services. With expertise in copyright, information management, artificial intelligence, and machine learning, CCC and its subsidiary RightsDirect collaborate with stakeholders to design and deliver innovative information solutions that power decision-making by harnessing information from a wide variety of data sources and content assets.



Submit date: 2/20/2025

I am responding to this RFI: Behalf of an Organization

Name: Katherine Eve

Name of Organization: Elsevier

Type of Organization: Other

Type of Organization-Other: Publisher

Role: Member of the Public

Comments: As a global leader in information and analytics, Elsevier helps researchers and healthcare professionals to advance science and improve health outcomes, striving to create a better future worldwide.

Elsevier shares the Office of Science and Technology Policy (OSTP) and National Institutes of Health (NIH) goals to increase the discoverability, transparency and impact of research, and we welcome the opportunity to respond to the “NIH Plan to Increase Findability and Transparency of Research Results Through the Use of Metadata and Persistent Identifiers (PIDs),” [the Plan]. We provide our comments for each section in turn, where relevant, under the relevant heading drawn from the Plan.

We appreciate your consideration of our below comments. As this Plan continues to be refined, we remain willing and open to supporting NIH, including via the Generalist Repository Ecosystem Initiative (GREI) where Traci Snowden, Product Manager at Elsevier is currently serving as the GREI Year 3 Co-Chair.

For further queries, please contact Katherine Eve, Policy Director for Open Science, Elsevier. Email: k.eve@elsevier.com.

Uploaded File: Elsevier-Response_NIH-Plan-for-metadata-and-PIDs.pdf

Description: Elsevier’s Response: NIH Plan to Increase Findability and Transparency of Research Results Through the Use of Metadata and Persistent Identifiers (PIDs)

Elsevier's Response: NIH Plan to Increase Findability and Transparency of Research Results Through the Use of Metadata and Persistent Identifiers (PIDs)

As a global leader in information and analytics, [Elsevier](#) helps researchers and healthcare professionals to advance science and improve health outcomes, striving to create a better future worldwide.

Elsevier shares the Office of Science and Technology Policy (OSTP) and National Institutes of Health (NIH) goals to increase the discoverability, transparency and impact of research, and we welcome the opportunity to respond to the “NIH Plan to Increase Findability and Transparency of Research Results Through the Use of Metadata and Persistent Identifiers (PIDs),” [the Plan]. We provide our comments for each section in turn, where relevant, under the relevant heading drawn from the Plan.

We appreciate your consideration of our below comments. As this Plan continues to be refined, we remain willing and open to supporting NIH, including via the Generalist Repository Ecosystem Initiative (GREI) where Traci Snowden, Product Manager at Elsevier is currently serving as the GREI Year 3 Co-Chair.

For further queries, please contact Katherine Eve, Policy Director for Open Science, Elsevier. Email: k.eve@elsevier.com.

I. Using and Submitting Metadata and PIDs

I.B. Including Metadata and PIDs when Submitting Manuscripts:

We note that the plan expects researchers to use ORCID identifiers when submitting new manuscripts to journal submission systems, as allowed. We welcome the recognition (via the addition “as allowed”) that not all editorial submission platforms offer functionality to capture ORCID identifiers, and reiterate the importance of flexibility as a component of any policy. The Editorial Manager submission system, used by Elsevier and other publishers, provides the option for authors to add their ORCID identifier. Given we serve researchers globally, wherein other identifiers may be preferred and adopted, we note that we will be unable to make this a requirement at submission.

The ORCID is displayed alongside the article on ScienceDirect and surfaced in the article metadata, where provided by the author. Where PDFs or NIH-funded articles are provided to NIH, NIH is responsible for extraction of any corresponding metadata.

I.C. Submitting Metadata and PIDs when Depositing Scientific Data in Repositories

Similarly, we note the plan expects researchers to use ORCID identifiers when submitting scientific datasets to repositories, as allowed. Mendeley Data, a Generalist Repository Ecosystem Initiative (GREI)-supported repository provides the option for authors to add their ORCID identifier. For the same reasons outlined above, we welcome recognition of the practical realities that will face researchers, and the flexibility afforded via the addition of “as allowed.”

The ORCID is displayed alongside the dataset in Mendeley Data. As a GREI-supported repository, Mendeley Data stands ready and willing to explore developing ORCID integration in metadata, should this be a priority under the GREI initiative.

We fully support the notion that researchers should be supported to choose the repository that represents the best and most appropriate home for their data. For some researchers, cost will be important, and Elsevier is proud that Mendeley Data, a GREI-supported repository, provides a mechanism for researchers to deposit datasets at no cost. We welcome continuation of GREI, including associated support under the initiative, for any developments that may be required to meet the expectations outlined in the draft Plan.

I.E. Citing and Cross-Linking Metadata and PIDs

Elsevier has long encouraged connection of datasets and their corresponding publications to enhance the discoverability, utility and impact of each component. We co-founded Scholix (now functionality that sits within Crossref) which provides a technical framework to link publications and datasets through their metadata.

During the submission process we encourage researchers to include dataset DOIs in their publication, either as part of data availability statements and/or in the reference lists. We also encourage them to update their datasets to include their publication DOI, when available. We recommend adding this additional straightforward and practical example for grantees as part of the Plan.

II. Collecting and Making Metadata and PIDs Publicly Available

II.A. Collecting Publicly Available Metadata and PIDs for Publications at NIH

Regarding the following sentence: “Publications resulting in whole or part from NIH support are required to be deposited into PMC,” where publications are to be made immediately available on PMC, we offer the gold open access (pay-to-publish) model. We respectfully remind NIH that the majority of journals and publishers (including Elsevier) are unable to support approaches that aim to make subscription articles immediately and freely available. Please refer to Elsevier’s response to the Request for Information on the National Institutes of Health Draft Public Access Policy (89 FR 51537) for further information.

Elsevier surfaces metadata fields and PIDs to support discoverability, transparency and impact assessment by research institutes and funders, ambitions shared by OSTP and NIH. We open a number of metadata fields for articles and their references within Crossref.

Establishing additional metadata feeds beyond our existing offering would require additional investment given the value-added services involved in providing and enhancing these feeds. We understand that the expectation for providing or extracting metadata to NIH-supported repositories sits with the agency, researcher, or researcher’s institution. We would encourage further consultation with researchers and research institutions to confirm feasibility and understand any researcher burdens or institutional overheads this could generate. We also note that while individual authors will be free to share their own original research metadata, there will be restrictions on sharing metadata that has been enhanced by publishers, as a result of publishers’ investments in technology and expertise.

II.B. Collecting Publicly Available Metadata and PIDs for Scientific Data in NIH-Supported Repositories

Similar to the above, we provide a feed of dataset metadata to DataCite. As a GREI-supported repository, Mendeley Data will explore any additional requirements on metadata delivery and search functionality under the direction of, and with support from GREI, should this be deemed a priority under the program.

III. Assigning Identifiers for NIH Awards and NIH-Conducted Research Projects

We look forward to hearing more of NIH's plans to develop a persistent identifier for NIH awards and conducted projects. Elsevier's submission system provides existing fields to capture author-provided funder/ grant IDs associated with publications. Likewise, Mendeley Data tracks NIH-funded awards and offers a field for grant IDs.

For further queries, please contact Katherine Eve, Policy Director for Open Science, Elsevier.
Email: k.eve@elsevier.com.

Submit date: 2/21/2025

I am responding to this RFI: Behalf of an Organization

Name: Shawna Sadler

Name of Organization: ORCID

Type of Organization: Nonprofit Research Organization

Type of Organization-Other:

Role: Institutional Official

Comments: Thank you for the opportunity to provide feedback to the NIH Plan to Increase Findability and Transparency of Research Results Through the Use of Metadata and Persistent Identifiers (PID). Please see our feedback as an attached file.

Uploaded File: ORCID_Response_to_NIH.pdf

Description:



21 February 2025

National Institutes of Health
9000 Rockville Pike, Bethesda, Maryland 20892, United States

RE: Comment Form: NIH Plan to Increase Findability and Transparency of Research Results Through the Use of Metadata and Persistent Identifiers (PID)

To the National Institutes of Health,

Thank you for the opportunity to provide feedback to the NIH Plan to Increase Findability and Transparency of Research Results Through the Use of Metadata and Persistent Identifiers (PID). We would like to thank NIH for including the following points in this plan:

- Senior and key personnel will be required to use their ORCID iD in the SciENcv system to submit Common Forms (for the Biographical Sketch, Current and Pending (Other) Support) and the NIH Biographical Sketch Supplement
- Link eRA profiles with an ORCID iD
- All NIH intramural principal investigators will also be required to obtain an ORCID iD, with the requirement for other scientists as determined by the Deputy Director for Intramural Research.
- Expect NIH-supported institutions to ensure all authors who are named senior and key personnel use ORCID iDs when submitting new research manuscripts to the NIH Manuscript Submission system and the journal submission system, as allowed.
- Expect NIH intramural principal investigators to use ORCID iDs when submitting new research manuscripts to the NIH Manuscript Submission system and the journal submission system, as allowed.
- Encourage NIH-supported institutions to consider the use of ORCID iDs for all other authors of a manuscript that was generated with NIH support.
- Expect institutions to ensure that submissions of scientific data that were generated with NIH support include the following metadata: ORCID iDs/PIDs and names for contributing senior and key personnel, affiliations (or other PIDs for affiliations¹) for contributing senior and key personnel, and funding sources.
- Encourage researchers to add their research outputs to their ORCID iD records.
- Require NIH-supported scientific data repositories to collect and make publicly available the following minimum set of metadata, which may include:



- submitter name, PID/ORCID iD for submitter, and affiliation (or PID for affiliation) or submitting organization name or PID,
- names and PIDs/ORCID iDs for all research contributors
- affiliations (or PIDs for affiliations) for all research contributors

ORCID supports NIH plans to explore the most suitable option for a PID that could be issued for NIH awards and conducted research projects.

To further maximize the benefits ORCID participation to NIH and NIH funded researchers and institutions, ORCID encourages NIH to:

- Write funding awards to ORCID records, with a grant DOI
- Write Peer Review activities to the peer reviewer's ORCID record
- Display ORCID iDs in citations PubMed Central® (PMC)
- Recommend that NIH-supported institutions write affiliations to their researchers ORCID records, specifically,
 - Employment
 - Education
 - Visiting scholar
- Recommend that professional societies and conference organizers write citations to ORCID records for publications and presentations. They can also write service membership roles to ORCID records, to facilitate disclosure of the researcher's professional affiliations.

I stand ready to speak with you further about our recommendation should this be useful,

Thank you,



7AF6654653DF4F0...
Chris Shillum

Executive Director, ORCID

Submit date: 2/21/2025

I am responding to this RFI: Behalf of an Organization

Name: Kacy Redd

Name of Organization: APLU, ARL, AAU, and COGR

Type of Organization: Professional Org or Association

Type of Organization-Other:

Role: Member of the Public

Comments:

Uploaded File: 2025-APLU-ARL-AAU-COGR-Comment-NIH-Metadata-Plan.pdf

Description: The Association of Public and Land-grant Universities (APLU), the Association of Research Libraries (ARL), the Association of American Universities (AAU), and COGR's comments on NIH's plan to increase findability and transparency of research outputs throu



To: NIH Office of Science Policy

From: Association of Public and Land-grant Universities (APLU): Contact – Kacy Redd, kredd@aplu.org
Association of Research Libraries (ARL): Contact - Marcel LaFlamme, marcel@arl.org
Association of American Universities (AAU): Contact – Kate Hudson, kate.hudson@aau.edu
COGR: Contact – Krystal Touns, ktouns@cogr.edu

Date: February 21, 2025

RE: Comments on NIH Plan to Increase Findability and Transparency of Research Results Through the Use of Metadata and Persistent Identifiers (PID) - NOT-OD-25-050
Submitted by online form

The Association of Public and Land-grant Universities (APLU), the Association of Research Libraries (ARL), the Association of American Universities (AAU), and COGR appreciate NIH's commitment to increasing findability and transparency of research outputs through metadata and persistent identifiers (PIDs). This plan represents an important step toward creating an interconnected knowledge network that will accelerate scientific discovery, enhance research security, and maximize the impact of federal research investments.

Our associations strongly support NIH's vision for standardizing metadata and PIDs. The research community has already made significant progress in establishing best practices for PIDs. In 2019, a National Science Foundation-sponsored conference convened by APLU, ARL, and other partners produced recommendations for adopting core PIDs, including Digital Object Identifiers (DOIs), ORCID iDs, and Research Organization Registry (ROR) IDs.¹ This framework provides a foundation for the knowledge network NIH envisions.

While supporting the plan's direction, we recommend the following refinements to ensure successful implementation.

Ensuring Trusted and Interconnected PID Infrastructure

The open PID infrastructure represents a core community asset that requires sustained financial and operational support. As NIH mandates ORCID usage and supports the use of ROR, we recommend specific actions to strengthen and maintain these critical research infrastructures.

We recommend that NIH establish and maintain automated systems that write publication and data output metadata from PubMed Central and other NIH-supported repositories directly to

¹ Chodacki, John, Cynthia Hudson-Vitale, Natalie Meyers, Jennifer Mulenburg, Maria Praetzellis, Kacy Redd, Judy Ruttenberg, Katie Steen, Joel Cutcher-Gershenfeld, and Maria Gould. *Implementing Effective Data Practices: Stakeholder Recommendations for Collaborative Research Support*. Washington, DC: Association of Research Libraries, September 2020. <https://doi.org/10.29242/report.effectivedatapactices2020>.

researchers' ORCID profiles. This bidirectional integration would enable NIH systems to both read from ORCID profiles and automatically update them when new public research outputs are deposited. Such automation improves data accuracy and demonstrates NIH's commitment to enriching the scholarly record.

Building on this infrastructure, NIH should convene federal funding agencies, publishers, institutions, and repositories to promote systems and practices for maintaining ORCID profiles with trusted, authoritative information. This coordinated approach would help create an interconnected PID ecosystem that strengthens research security through better visibility of research relationships and funding sources, while reducing the burden on researchers.

We appreciate NIH's movement towards an ideal PID landscape in U.S. research, which would be a fully interoperable and widely adopted system where researchers, datasets, publications, institutions, and funding sources are seamlessly connected through standardized, open PIDs like ORCID, DOI, and ROR. This would enhance research integrity, reproducibility, and accessibility while reducing administrative burden, ensuring long-term discoverability of scholarly outputs, and enhancing research security.

Reducing Institutional and Researcher Burden while Enhancing Research Security

We strongly support the adoption of PIDs, and we recommend refining the implementation approach regarding institutional oversight. The current plan states that NIH expects "institutions to ensure all authors who are named senior and key personnel use ORCID iDs when submitting manuscripts." This expectation is not aligned with established institutional roles and responsibilities and would create a significant administrative burden.

Institutions lack the mechanisms to monitor real-time manuscript submissions to journals and repositories. For example, when researchers from multiple institutions collaborate on NIH-funded research, no single institution has visibility into all submission activities. A paper might be submitted by a co-author at another institution without advance notice to the primary institution. Additionally, institutions would need to create new tracking systems and staffing to monitor thousands of annual submissions across hundreds of journals and repositories - a costly and inefficient approach. Creating systems to track compliance at the institutional level could also have the effect of being perceived as a barrier to publishing and academic freedom.

These institutional oversight challenges extend beyond manuscript submissions. The plan includes multiple expectations for institutions to ensure ORCID iD usage and metadata submission across various research outputs. For example, institutions are expected to ensure metadata submissions for scientific data include ORCID iDs, affiliations, and funding sources for all key personnel. As with manuscript submissions, institutions often lack visibility and control over researcher interactions with third-party systems, making such oversight impractical.

While the administrative burden concerns are significant, it is important to note that a well-implemented PID system serves broader national interests. The NSPM-33 Implementation Guidance specifically identifies PIDs as a crucial tool for research security, enabling transparent documentation of research relationships and collaborations. An automated, system-based

approach to PID implementation would both reduce administrative burden and better achieve these research security objectives compared to institution-by-institution monitoring.

Under NIH guidelines, institutions, through their Authorized Organizational Representatives (AORs), certify compliance with award terms and conditions, while Program Directors/Principal Investigators (PD/PIs) are responsible for the proper conduct of research activities, including publication submissions.

Instead of requiring institutional monitoring of individual manuscript submissions, we recommend NIH:

- Build ORCID iD requirements into existing manuscript submission systems (i.e., PubMed Central), as applicable by law
- Allow PD/PIs to certify compliance through standard progress reports
- Maintain institutional responsibility at the policy and certification level rather than individual transaction monitoring

This approach achieves increased PID adoption while preserving established administrative structures and avoiding unsustainable oversight burdens on institutions.

We stand ready to work with NIH to develop a robust, trusted, and impactful PID infrastructure. This will allow us to better track research outcomes and impacts, strengthen research security, and accelerate science as research outputs become easier to find and reuse.

Submit date: 2/21/2025

I am responding to this RFI: Behalf of an Organization

Name: Anurupa Dev

Name of Organization: Association of American Medical Colleges

Type of Organization: Professional Org or Association

Type of Organization-Other:

Role: Institutional Official

Comments:

Uploaded File: AAMC-Plan-to-Increase-Findability-and-Transparency-of-Research-Results.pdf

Description: AAMC Response to NIH



**Association of
American Medical Colleges**
655 K Street, N.W., Suite 100, Washington, D.C. 20001-2399
T 202 828 0400 F 202 828 1125
www.aamc.org

February 20, 2025

NIH Office of Science Policy
9000 Rockville Pike
Bethesda, Maryland 20892

Re: NIH Plan to Increase Findability and Transparency of Research Results Through the Use of Metadata and Persistent Identifiers (PID)

Submitted online at <https://osp.od.nih.gov/comment-form-nih-plan-to-increase-findability-and-transparency-of-research-results-through-the-use-of-metadata-and-persistent-identifiers-pids/>

The Association of American Medical Colleges (AAMC) appreciates the opportunity to provide feedback to the National Institutes of Health (NIH) on the agency’s plan to increase the findability and transparency of research results through the use of metadata and persistent identifiers (PIDs).

The AAMC is a nonprofit association dedicated to improving the health of people everywhere through medical education, health care, medical research, and community collaborations. Its members are all 158 U.S. medical schools accredited by the Liaison Committee on Medical Education; 13 accredited Canadian medical schools; nearly 500 academic health systems and teaching hospitals, including Department of Veterans Affairs medical centers; and more than 70 academic societies. Through these institutions and organizations, the AAMC leads and serves America’s medical schools, academic health systems and teaching hospitals, and the millions of individuals across academic medicine, including more than 201,000 full-time faculty members, 97,000 medical students, 158,000 resident physicians, and 60,000 graduate students and postdoctoral researchers in the biomedical sciences. Following a 2022 merger, the Alliance of Academic Health Centers International broadened participation in the AAMC by 70 international academic health centers throughout five regional offices across the globe.

The AAMC strongly supports efforts to increase the findability of research outputs and promote transparency in the research process, as well as the ability for NIH to track the outputs of its investment in research. As previously noted in comments to the White House Office of Science and Technology Policy (OSTP)¹ and NIH², “Making these outputs more readily available advances science by enabling further validation of experimental results, facilitating reuse of hard to-generate data, catalyzing new research and scientific collaboration, and generally promoting more responsible

¹ AAMC Comments to OSTP. Request for Information: Public Access to Peer-Reviewed Scholarly Publications, Data and Code Resulting from Federally Funded Research (85 FR 9488). May 6, 2020.
<https://www.aamc.org/media/44641/download?attachment>

² AAMC Comments to NIH. Re: NOT-OD-20-013: Request for Public Comments on a DRAFT NIH Policy for Data Management and Sharing and Supplemental Draft Guidance. Jan. 10, 2020.
<https://www.aamc.org/media/40536/download?attachment>

stewardship of federal resources.” We are pleased to offer the following comments to the NIH as it develops a policy for the use of PIDs and metadata.

I. Using and Submitting Metadata and PIDs

The use of ORCID as a unique, persistent identifier for researchers is a key step to improving the connectedness and findability of research outputs (as defined in the Plan, publications and scientific data). The AAMC has previously noted the critical importance of ORCID as part of its Credit for Data Sharing initiative, which is focused on linking researchers and organizations to shared datasets³. We also note the central role that ORCID plays in research security efforts and agency implementation of National Security Presidential Memorandum-33 (NSPM-33)⁴, by strengthening the linkage and documentation between researchers and funding sources and allowing for easier tracking of any potential conflicts of commitment. NSPM-33 was issued in Jan. 14, 2021 by the Trump Administration, and the work initiated by OSTP during that time to carry out its implementation continues as initially envisioned.

To optimize the use of ORCID, we strongly encourage the NIH to work with ORCID to establish and maintain automated systems that populate publication and data metadata from PubMed Central and other NIH-supported repositories directly into researchers' ORCID profiles. Establishing a robust automated process would improve data integrity, reduce institutional and researcher burden, and maximize the NIH's role in process improvement. Additionally, NIH should work with other federal research agencies to promote a set of unified practices which would maintain ORCID profiles with trusted, authoritative information. Such coordination would strengthen the overall PID ecosystem and ensure that the NIH's requirement for ORCID is valuable to both the agency and the broader research ecosystem.

Finally, NIH should continue to produce guidance and trainings on the use of ORCIDs and any other selected PIDs and ensure that these are easily findable and usable. We also suggest that the agency continue to engage with institutions, particularly libraries and sponsored program offices, to understand the challenges which might arise from the use or implementation of ORCID or standardized metadata.

II. Collecting and Making PIDs and Metadata Publicly Available

The AAMC recognizes and supports the value of PubMed and PubMed Central (PMC) continued efforts to collect and make publicly searchable all available metadata submitted to PMC, including authors' names, affiliations, funding information, publication dates and DOI. We also appreciate efforts to utilize these same metadata fields across other NIH sites such as REPORTER and iCite, to further improve data linkage.

³ Pierce, H.H., et al. Credit data generators for data reuse. *Nature* 570, 30-32 (2019). <https://doi.org/10.1038/d41586-019-01715-4>

⁴ Presidential Memorandum on United States Government-Supported Research and Development National Security Policy. Issued on: January 14, 2021. <https://trumpwhitehouse.archives.gov/presidential-actions/presidential-memorandum-united-states-government-supported-research-development-national-security-policy/>


AAMC recognizes the need for standardized metadata as a necessary component of effective research data sharing. As noted in previous comments to NIH on the draft scientific data sharing policy, and to OSTP⁵ on the desired characteristics of research repositories, “In order for data to be successfully reused, it must not only be deposited in an appropriate repository, but also meet several other criteria, including adequate metadata, curation, and the use of common standards.” We encourage NIH to work with community partners and repositories to inform best practices for metadata standards and curation, and to maintain a list of federally supported repositories which are available to NIH grantees.

III. Assigning Identifiers for NIH Awards and NIH-Conducted Research Projects

As NIH notes, PIDs are most useful when they can be linked in standardized ways, and we encourage NIH to collaborate not only with other federal agencies, but also with community organizations, institutions, and societies as it determines the most suitable PIDs for research awards and conducted research projects. Cross-stakeholder groups such as the Research Data Alliance and FORCE11 have already developed protocols and standards to be used for both PIDs and metadata that align with the FAIR⁶ and TRUST⁷ principles for data and repositories and should be utilized as a resource during this process.

We are very appreciative of the work NIH has undertaken to improve the findability and transparency of research products. The AAMC looks forward to continued engagement with NIH as the process of policy development progresses. Please feel free to contact me or my colleague Anurupa Dev, PhD, Director of Science Policy and Strategy (adev@aamc.org), with any questions about these comments.

Sincerely,



Elena Fuentes-Afflick, MD, MPH
Chief Scientific Officer

cc: David J. Skorton, MD, AAMC President and Chief Executive Officer

⁵ AAMC Comments to OSTP Re: Request for Public Comment on Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting from Federally Funded Research (85 FR 3085). March 10, 2020.

<https://www.aamc.org/media/42891/download?attachment>

⁶ FAIR Principles. <https://www.go-fair.org/fair-principles/>

⁷ Lin, D., et al. The TRUST Principles for digital repositories. *Sci Data* 7, 144 (2020). <https://doi.org/10.1038/s41597-020-0486-7>

Submit date: 2/21/2025

I am responding to this RFI: Behalf of an Organization

Name: J. Carl Maxwell

Name of Organization: Association of American Publishers

Type of Organization: Other

Type of Organization-Other: Trade Association-Publishers

Role: Member of the Public

Comments: Please find attached a comment from the Association of American Publishers

J. Carl Maxwell

Senior Vice President, Public Policy

Association of American Publishers

Uploaded File: AAP-NIH-PIDS-NOT-OD-25-050_Final.pdf

Description: Comment from Association of American Publishers Attached

February 21, 2025

National Institutes of Health
Office of Science Policy
6705 Rockledge Drive
Suite 630, MSC-7985
Bethesda, MD 20892

Submitted via electronic form.

Re: Written Comments in Response to “Request for Public Comment: NIH Plan to Increase Findability and Transparency of Research Results Through the Use of Metadata and Persistent Identifiers (PIDs)” (NOT-OD-25-050)

To Whom It May Concern,

The Association of American Publishers (AAP) welcomes this opportunity to provide written comments in response to “Request for Public Comment: NIH Plan to Increase Findability and Transparency of Research Results Through the Use of Metadata and Persistent Identifiers (PIDs).” AAP represents over 80 Professional and Scholarly Publishers, including dozens of scholarly societies representing over a million scientists, engineers, researchers, and other members of the academy. A full list of [AAP members](#) may be found on our website: [publishers.org](#)

Scientific publishing has been a critical part of the scientific method for centuries. AAP members take deep pride in their significant contributions to advancing science and engineering, economic prosperity, and public welfare. Our work in article selection, curation, peer, and editorial review, as well as publication form a trusted foundation of modern biomedicine. Many of the advancements enabling modern open science and information sharing, including online publication, pre-print servers, archiving, persistent identifiers, rigorous standards, and metadata, are a direct result of the publishing industry investments. We believe a free and competitive marketplace for scholarly publishing is critical for sustaining the high quality of scholarly communication, which benefits authors, funders, the scientific community, and society at large.

AAP presents two general recommendations for the National Institutes of Health (NIH) as it considers future plans on metadata and PIDs:

NIH policies should center and empower researchers, enabling them to communicate the results of their research in the venue of their choice for maximum impact, without burdensome compliance regimes or unfunded mandates. At the core of scientific integrity is allowing researchers freedom and space to conduct the scientific process and communicate their findings to academia and public without restriction or reservation. Researchers should be able to decide how and where they report and publish their findings and interact with their community and the broader public. This includes ensuring their freedom

of choice in publication outlets and the licenses that apply to their work. Researchers should have final say about who can modify and commercialize their work.

Regarding publications based on research funded by the National Institutes of Health, NIH is unique among federal science agencies in having specific, targeted statutory authorization from Congress on the issue of public access ([P.L. 110-161](#)):

SEC. 218. The Director of the National Institutes of Health shall require that all investigators funded by the NIH submit or have submitted for them to the National Library of Medicine's PubMed Central an electronic version of their final, peer-reviewed manuscripts upon acceptance for publication, to be made publicly available no later than 12 months after the official date of publication: *Provided*, That the NIH shall implement the public access policy in a manner consistent with copyright law.

AAP recommends any future NIH plans appropriately align with Congress's detailed statutory authorizations, and we would direct NIH to AAP's response to the Request for Information on the National Institutes of Health (NIH) Draft Public Access Policy (89 FR 51537).

Additionally, publishers seek to explore partnerships with NIH to boost innovation, information sharing, and scientific discovery. As NIH considers future changes and improvements, we hope to bring publishers from across the research spectrum together with NIH to revolutionize biomedical research. AAP would be interested in hosting a series of conversations with NIH to encourage this discussion.

AAP appreciates this opportunity to comment on NIH's plans to improve the health information ecosystem and looks forward to future opportunities to partner and dialogue with the agency.

With Regards,

J. Carl Maxwell
Senior Vice President, Public Policy
Association of American Publishers

Submit date: 2/21/2025

I am responding to this RFI: Behalf of an Organization

Name: Meagan Phelan

Name of Organization: AAAS

Type of Organization: Nonprofit Research Organization

Type of Organization-Other:

Role: Institutional Official

Comments: AAAS Response to RFI on the NIH Plan to Increase Findability and Transparency of Research Results Through the Use of Metadata and Persistent Identifiers

The American Association for the Advancement of Science (AAAS) welcomes the NIH's efforts to enhance public access, in line with OSTP guidance aimed at making federally funded research publications and supporting data publicly available. Open and accessible data are essential to scientific integrity and reproducibility.

AAAS, a multi-disciplinary non-profit association of scientists at all levels of the scientific enterprise, publishes the Science family of journals. Our mission is to advance science and innovation throughout the world for the benefit of all.

The Science family of journals is open to the public without embargo using green open access models for five of our journals and a gold open access model for one.

Our journals require published authors to make their data immediately accessible in approved repositories and authors may share their author accepted manuscripts immediately upon publication.

AAAS applauds the NIH for balancing readers' need for access to published work with authors' ability to publish, in its approach to public access policy development. AAAS is committed to collaborating with NIH, other federal research agencies, and OSTP to develop public access policies that achieve this balance and is pleased to offer its response to the NIH's RFI in this document.

Response:

Access and transparency are foremost considerations at AAAS, where our mission includes communicating science accurately, broadly, and in such a way to ensure the scientific community can reanalyze and reproduce new works. In recognition, AAAS supports the final peer-reviewed author-accepted manuscript (AAM) version of a paper being broadly and immediately shared and the flexibility afforded by NIH's acceptance of this approach as a means of complying with its updated public access policy. At AAAS, however, we believe that publisher oversight of a final version (the version of record, or VOR) is essential – not only to maintaining the quality and accuracy of scientific research but also to advancing the subsequent work from which new research stems. Only the final version of a manuscript overseen by a publisher committed to maintaining the accuracy of the scientific record can be counted on to be corrected, retracted or otherwise updated with clear notation for the global scientific research

community. Ensuring that publication repositories clearly distinguish between multiple versions of articles (i.e., ensuring that singular publication records point to the VoR, where the AAM is deposited first) will be critical, as NIH moves forward. The NIH may wish to implement guidelines requiring that authors depositing their AAMs provide a DOI (digital object identifier) pointing to the VOR. Indeed, at AAAS, our instructions for authors depositing AAMs require them to include a link to the VOR.

With respect to metadata, linkages between publishers and organizations such as the Research Organization Registry (ROR), Open Researcher and Contributor ID (ORCID), Crossref, and data repositories are aimed at increasing robustness of metadata by providing persistent identifiers and connecting them to research outputs. As a publisher, AAAS monitors and implements best practices for both metadata collection (e.g., on institutions and funders) and metadata propagation in the VOR and associated research objects.

All Science journal papers include details about funding, author contributions, competing interests, data and materials availability, and license information. The publisher oversees accuracy of important associated metadata after publication, including in cases where authors request to change their names in previously published papers, as one example. As a criterion to publish, AAAS requires authors to make their data publicly accessible. AAAS has also partnered with Dryad, an international open-access data repository; we encourage such partnerships because they help ensure that publishers and repositories share the same metadata, thus providing better linkage between the data and the research paper. NIH may wish to consider implementing guidelines for data availability in publications. These guidelines could include a clear set of criteria for data deposition and ease of linking to that data, which publishers could help enforce. As a best practice, NIH could also encourage connections between publishers and data repositories of various kinds (general or field-specific, or both).

In its “Plan to Increase Findability and Transparency of Research Results Through the Use of Metadata and Persistent Identifiers,” NIH is seeking to uphold a standardized minimum set of metadata. This is important. We recommend this standardized minimum aim to have a common language/shared terminology. Further, we recommend the metadata required describe not only datasets but also code underlying research papers. To minimize researcher burden for compliance, it would be ideal if metadata could be entered centrally via a common metadata app/other system. Finally, if there is a metadata change at one source—one of the linked places (i.e., journal, data repository)—that change needs to get connected to the other locations where the metadata is referenced.

Description: AAAS Response to RFI on the NIH Plan to Increase Findability and Transparency of Research Results Through the Use of Metadata and Persistent Identifiers