PeerJ Preprints NOT PEER-REVIEWED

Open Science Strategies for NIH Data Management, Sharing, and Citation: A Response to NIH Request for Information (RFI) NOT-OD-17-015

Tim Clark[1,2], Helena Cousijn[3], Daniel S. Katz[4] and Martin Fenner[5]

[1] Massachusetts General Hospital, Boston MA
[2] Harvard Medical School, Boston MA
[3] Elsevier BV, Amsterdam NL
[4] University of Illinois, Urbana-Champaign IL
[5] DataCite, Hannover DE

This document summarizes a series of authoritative views on research data management, sharing and citation, developed in Expert Groups, Working Groups, and other activities organized through FORCE11 (http://force11.org), an international community of over 2,000 members dedicated to advancing research communications and e-scholarship.

It was prepared in response to NIH Request for Information (RFI) **NOT-OD-17-015 (**URL: https://grants.nih.gov/grants/guide/notice-files/NOT-OD-17-015.html), and has been reviewed and approved by the FORCE11 Board of Directors on January 19, 2017.

I. Data Sharing Strategy Development

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

Concerns of science policy bodies such as CODATA, the Royal Society and the U.S. National Academies about reliability of published scientific findings and reusability of research data [1-3] led to the development of the Joint Declaration of Data Citation Principles (JDDCP) [4, 5], which has been subsequently endorsed by over 100 scholarly organizations.

The JDDCP require archiving of primary research data in persistent stores, and its citation and inclusion in a reference list, wherever it is the basis for a published research finding [4, 5]. This occurs: (1) where findings or claims are based on the authors' primary research data; and (2) where data from other sources is input to the authors' analysis.

In our view, the first use case is critically important because mandating it will force data into persistent archives and thus support validation and verification of many research results. This is already done for certain specialist archives such as sequence, expression and protein structure data. We propose it be expanded, as proposed in the JDDCP, to all research data. Generalist data archives such as Dryad, Figshare, and Dataverse, already exist to handle such data (see re3data for a service describing these and similar data archives [6]). The second use case is important as well, e.g. for meta-analysis studies. However, it is currently a less common use case than the first and the second use case presupposes the first already being implemented in practice. Therefore, the first use case is in our view the immediate policy priority. Ultimately, both use cases must be supported.

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

Maintaining data availability for secondary research imposes both costs and benefits on the research ecosystem. This is ultimately a judgment and cost-benefit determination that varies with the type of data. Like books in libraries, data may be de-accessioned if no longer relevant.

The JDDCP determined that, regardless of the concrete persistence policy for any given dataset, the metadata and its landing page in the data archive should persist. Likewise, a data citation should always resolve by default to such a landing page, from which point further navigation to the underlying data can be requested either manually or in the case of software agents, by content resolution [4, 5] .

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

There are development, implementation, and maintenance burdens for data stewardship and sharing; and these can each be subdivided into policy, infrastructure, operational, and cultural burdens or costs.

FORCE11 and its stakeholder communities have already undertaken very significant policy development work enabling common definition of data citation practices [1, 2, 4, 5, 7-9], along with implementation preparation and pilot studies work, partly funded by NIH and outlined here:

(a) document model revisions to the NISO Journal Article Tag Suite, now standardized in ANSI/NISO Z39.96-2015 [10];
(b) human and machine accessibility guidelines for cited data [5];
(c) a Data Citation Roadmap for Scholarly Data Repositories [11];
(d) a Data Citation Roadmap for Scientific Publishers [12];
(e) a model for Uniform Resolution of Compact Identifiers, which are commonly used in biomedical repositories [13];
(f) a set of Software Citation Principles [14].

These studies outline a path to achieve comprehensive research transparency as called for in the FAIR Principles [15].

Additionally, core work undertaken by the bioCADDIE program to build a data discovery index, produced

(g) the DATS vocabulary of data discoverability metadata [16]; and
(h) DataMed, a pilot data discovery index [17].

Additional costs will be associated with rollout of a production version of DataMed. The bioCADDIE team, in coordination with NLM and NCBI staff, can best estimate these costs.

Early-adopter publishers such as Elsevier [9, 18, 19] (https://www.elsevier.com/about/press-releases/science-and-technology/elsevier-implements-data-citation-standards-to-encourage-authors-to-share-research-data) and SpringerNature have helped to define practical requirements at the journal level. This requires supplementation by continued outreach to other publishers and repositories. FORCE11 has developed a three-year plan to accomplish this, outlined in a recent U13 submission to NIH.

Through NIH policies for those it funds, and working with publishers, authorship practices can be transformed. This is a cultural transformation requiring material incentives similar to those developed to populate PubMed Central.

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

We believe data citation and sharing will have many profound benefits, including significant improvements to our ability to validate and verify results, reuse data, and translatability of research results to successful pharmaceutical development. At the same time, it should be remembered that validity of data depends upon the methods used to obtain, transform and analyze it. Citation,

identification and sharing of software [14] and key biological research reagents [20, 21] are essential elements in a larger sharing and joint validation culture that needs the support and sustaining focus of NIH and other research funding and policy bodies.

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

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

FORCE11 supports increased reporting requirements of data and software sharing in RPPRs and competing grant applications as a means to incentivize data sharing. This includes coordination of standards for data sharing / data management plans with the practices we outline here. Too often such plans are mere decoration and boilerplate. Adherence should be closely verified and this can happen if NIH requires that the data needed for supporting the findings of all grant funding publications is archived in repositories that conform to JDDCP requirements and publisher requirements based on JDDCP. The FORCE11 Publishers Roadmap to Data Citation [12] discusses how publishers can determine what repositories meet archival requirements and provides sources where conformance lists can be found.

We strongly support the use of preprint repositories such as bioRXiv (http://biorxiv.org); in particular, where early deposition of preprints is accompanied by coordination of data and software archiving and citation policies with such archives.

2. Important features of technical guidance for data and software citation in reports to NIH, which may include:

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

   The FORCE11 recommendations for data repositories recommends that all datasets and software intended for citation must have a globally unique persistent identifier that resolves to a landing page specific for that resource [11]. For software citations, the identifier should resolve to a landing page referencing both the specific version, and the software project as a whole [14].

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

   We recommend that data and software citations are included in reports in the same way as in articles, with a persistent unique identifier as described in (a) above. This means that the data and software citations have the same structure as other kinds of citations/references, and include the following elements: author(s), title, repository, year, version and persistent identifier. More information and examples of formatted data references can be found in the FORCE11 recommendations for publishers [12].

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

   We recommend that the authors should be included in the metadata for data and software, using ORCID IDs or other appropriate persistent identifiers for authors. We recommend that

all contributors to a specific cited version of software be identified in both the landing pages and the citation.

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

Both use cases are common with data citation and therefore need to be supported. While a study should cite the underlying data as specifically as possible (as recommended in the JDDCP), we also need to be able to cite a collection of data from one or more databases that was used in a study. An example use case for this is a very large number of datasets that is impractical to cite individually. Services such as Biostudies (PMID: 26700850) provide this functionality. As discussed in 2a above, individual software releases and the overall software development project should be able to be cited, with appropriate metadata that can be used to link them together for understanding their interaction and gathering project-wide citation statistics.

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

This information should be part of the required metadata for the data/software resource – the *publisher* property – as described in the FORCE11 recommendations for data repositories [11].

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

NIH can exercise a significant influence on publishers of biomedical research articles, as recent history has demonstrated; and on data archives. For data citation to be comprehensively adopted, an ecosystem without major gaps needs to be developed.

(a) NIH should provide sufficient funding to support required transformations of this ecosystem – including **adequately funding the enhancement of its own data repositories** to comply with the landing page metadata requirements detailed in *A Data Citation Roadmap for Scholarly Data Repositories* [11].
(b) NIH should provide adequate funding for continued outreach to non-NIH repositories, publishers and identifier/metadata providers to fully adopt and support JDDCP compliant data citation practices. This can be done through U13 mechanisms, initially through RFA-CA-16-020 BD2K Support for Meetings of Data Science Related Organizations. Needs for continuation of support and coordination meetings should be continually reassessed.
(c) NIH should work closely with ELIXIR, the EMBL-EBI, and corresponding bodies in Asia, to develop joint funding models for data archiving and citation infrastructure.
(d) NIH should consider expanding its Public Access policy beyond PubMed Central deposition of articles, to include data and software deposition in recognized JDDCP-compliant repositories. Wherever possible it should find ways to strongly incentivize archived data and software citation in primary research articles.
(e) NIH should continue to support JATS related development and reuse activities.
(f) NIH should actively fund extramural activities to further and incentivize software and research resource identification, archiving / cataloguing, and citation.

(g) In addition to requiring Data Management plans, NIH should require a specific management plan for software where it is a major project component of a funded effort. Such plans should include a requirement for reporting on reuse of the datasets and software, thereby incentivizing good data and software management and the sharing of high-quality digital research materials.

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

None at present.

References

1. CODATA/ITSCI Task Force on Data Citation: **Out of cite, out of mind: The Current State of Practice, Policy and Technology for Data Citation**. *Data Science Journal* 2013, **12**:1-75. http://doi.org/10.2481/dsj.OSOM13-043
2. RoyalSociety: **Science as an Open Enterprise**. In. London: The Royal Society Science Policy Center; 2012. https://royalsociety.org/~/media/policy/projects/sape/2012-06-20-saoe.pdf
3. Uhlir Pe: **Developing Data Attribution and Citation Practices and Standards**. In. Washington DC: National Academies; 2012. http://www.nap.edu/download.php?record_id=13564.
4. Data Citation Synthesis Group: **Joint Declaration of Data Citation Principles**. In. Edited by Martone M. San Diego CA: Future of Research Communication and e-Scholarship (FORCE11); 2014 https://http://www.force11.org/datacitation.
5. Starr J, Castro E, Crosas M, Dumontier M, Downs RR, Duerr R, Haak LL, Haendel M, Herman I, Hodson S, JH, Kratz JE, Lin J, Nielsen LH, Nurnberger A, Proell S, Rauber A, Sacchi S, Smith A, Taylor M, Clark T: **Achieving human and machine accessibility of cited data in scholarly publications**. *PeerJ* 2015, **1**: **PMID:** 26167542
6. Pampel H, Vierkant P, Scholze F, Bertelmann R, Kindling M, Klump J, Goebelbecker H-J, Gundlach J, Schirmbacher P, Dierolf U: **Making Research Data Repositories Visible: The re3data.org Registry**. *PLOS ONE* 2013, **8**(11):e78080. http://dx.doi.org/10.1371%2Fjournal.pone.0078080.
7. Uhlir P: **For Attribution - Developing Data Attribution and Citation Practices and Standards: Summary of an International Workshop (2012)** In.: The National Academies Press; 2012: 220 http://www.nap.edu/catalog.php?record_id=13564.
8. Altman M, King G: **A Proposed Standard for the Scholarly Citation of Quantitative Data**. *DLib Magazine* 2006, **13**(3/4):march2007-altman. http://www.dlib.org/dlib/march07/altman/03altman.html.
9. Altman M, Borgman C, Crosas M, Martone M: **An introduction to the joint principles for data citation**. *Bulletin of the Association for Information Science and Technology* 2015, **41**(3):43-45. http://doi.org/ 10.1002/bult.2015.1720410313
10. NISO: **JATS: Journal Article Tag Suite, version 1.1**. In., vol. ANSI/NISO Z39.96-2015. Baltimore MD, USA: National Information Standards Organization; 2015 http://www.niso.org/apps/group_public/download.php/15933/z39_96-2015.pdf.
11. Fenner M, Crosas M, Grethe J, Kennedy D, Hermjakob H, Rocca-Serra P, Berjon R, Karcher S, Martone M, Clark T: **A Data Citation Roadmap for Scholarly Data Repositories**. *bioRXiv* 2016. https://doi.org/10.1101/097196
12. Cousijn H, Kenall A, Ganley E, Harrison M, Kernohan D, Murphy F, Polischuk P, Martone M, Clark T: **A Data Citation Roadmap for Scientific Publishers** *bioRXiv* 2017. https://doi.org/10.1101/100784

13. Wimalaratne SM, Juty N, Kunze J, Janée G, McMurry JA, Beard N, Jimenez R, Grethe J, Hermjakob H, Clark T: **Uniform Resolution of Compact Identifiers for Biomedical Data**. *bioRXiv* 2017. https://doi.org/10.1101/101279
14. Smith AM, Katz DS, Niemeyer KE: **Software citation principles**. *PeerJ Computer Science* 2016, **2**:e86. https://doi.org/10.7717/peerj-cs.86.
15. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PAC, Hooft R, Kuhn T, Kok R, Kok J *et al*: **The FAIR Guiding Principles for scientific data management and stewardship**. *Scientific Data* 2016, **3**:160018. http://dx.doi.org/10.1038/sdata.2016.18.
16. Gonzalez-Beltran A, Rocca-Serra P: **WG3-MetadataSpecifications: DataMed DATS specification v2.1 - NIH BD2K bioCADDIE** *Zenodo* 2016: **PMID:**
17. Ohno-Machado L, Sansone S-A, Alter G, Fore I, Grethe J, Xu H, Gonzalez-Beltran A, Rocca-Serra P, Soysal E, Zong N, Kim H-e: **DataMed: Finding useful data across multiple biomedical data repositories**. *bioRXiv* 2016. https://doi.org/10.1101/094888
18. Taylor M: **Data citation is becoming real with FORCE11 and Elsevier**. In: *Research Data.* 2016. http://www.elsevier.com/connect/data-citation-is-becoming-real-with-force11-and-elsevier.
19. Cousijn H, Ash E: **Making data citation a reality**. In: *Elsevier Connect.* Elsevier; 2016. http://www.elsevier.com/connect/making-data-citation-a-reality.
20. Bandrowski A, Brush M, Grethe JS, Haendel MA, Kennedy DN, Hill S, Hof PR, Martone ME, Pols M, Tan SS, Washington N, Zudilova-Seinstra E, Vasilevsky N, Initiative RRI: **The Resource Identification Initiative: A Cultural Shift in Publishing**. *Neuroinformatics* 2016, **14**(2):169-182: **PMID:** 26589523 http://www.ncbi.nlm.nih.gov/pubmed/26589523.
21. Bandrowski A, Tan S, Hof PR: **Promoting research resource identification at JCN**. *The Journal of comparative neurology* 2014, **522**(8):1707: **PMID:** 24723247 http://www.ncbi.nlm.nih.gov/pubmed/24723247.