A Data Citation Roadmap for Scientific Publishers

Helena Cousijn[1]*, Amye Kenall[2]*, Emma Ganley[3], Melissa Harrison[4], David Kernohan[5], Fiona Murphy[6], Patrick Polischuk[3], Maryann Martone[7], Tim Clark[8,9]

1. Elsevier, Amsterdam, Netherlands
2. Springer Nature, London, UK
3. Public Library of Science, San Francisco CA, USA
4. eLife Sciences Publications, Ltd., Cambridge, UK
5. JISC, Bristol, UK
6. University of Reading, Reading, UK
7. University of California, San Diego, La Jolla CA, USA
8. Massachusetts General Hospital, Boston MA, USA
9. Harvard Medical School, Boston MA, USA

*These authors contributed equally to the work.

Corresponding author: h.cousijn@elsevier.com

Abstract

This article presents a practical roadmap for scholarly publishers to implement data citation in accordance with the Joint Declaration of Data Citation Principles (JDDCP) [1], a synopsis and harmonization of the recommendations of major science policy bodies. It was developed by the Publishers Early Adopters Expert Group as part of the Data Citation Implementation Pilot (DCIP) project, an initiative of FORCE11.org and the NIH BioCADDIE program. The structure of the roadmap presented here follows the "life of a paper" workflow and includes the categories Pre-submission, Submission, Production, and Publication. The roadmap is intended to be publisher-agnostic so that all publishers can use this as a starting point when implementing JDDCP-compliant data citation.

Introduction

Over the past several years CODATA, the U.S. National Academy of Sciences, and other groups have conducted in-depth authoritative studies on data practices in the sciences. These studies identify problems in reproducibility, robustness and reusability of scientific data, leading ultimately to problems in the scientific record [2, 3, 4, 5].

These studies uniformly recommend that scholarly articles now treat the primary data upon which they rely as first class research objects; that primary data is robustly archived and directly cited as support for findings just as literature is cited. Archived data is recommended – as a matter of policy and of good scientific practice - to be "FAIR": Findable, Accessible, Interoperable, and Reusable [6]; and to be accessible from the primary article. The method for establishing this accessibility is a data citation.

The Joint Declaration of Data Citation Principles (JDDCP) summarizes the recommendations of these studies, and has been endorsed by over 100 scholarly organizations, funders and publishers [1]. Further elaboration on how to implement the JDDCP was provided in Starr et al. 2015 [7], with an emphasis on practices for digital repositories.

There is a clear emerging consensus in the scholarly community supporting the practice of data citation. This is reflected not only in the broad endorsement of the JDDCP, but also in the increasing proliferation of workshops on this topic.

Implementing data citation is not meant to replace or bypass citation of the relevant literature, but rather to ensure we provide verifiable and potentially re-usable backing data for published assertions. It is aimed at significantly improving the robustness and reproducibility of science, which have been the subject of much recent concern.

The present document is a detailed roadmap to implementing JDDCP-compliant data citation, prepared by publishers, for publishers, as part of a larger effort involving repositories, informaticians, and identifier / metadata registries.

Recommendations

This section briefly explains what a data citation is, and then presents recommendations for publishers to implement it.

Data citations are formal ways to ground the research findings in a manuscript, upon their supporting evidence, when that evidence consists of externally archived datasets. They presume that the underlying data is too large to be presented in a table or figure in the article, and that it has been robustly archived in a long-term-persistent repository. Publishers implementing data citation will have domain-specific lists of acceptable repositories for this purpose. We provide some of these lists further along in the manuscript.

Both the dataset reference in the primary article, and the archival repository, should follow certain conventions. These are ultimately based upon the JDDCP's eight principles. The conventions for Repositories are presented in *Starr et al 2015* [7]. A Roadmap for Repositories implementing data citation is provided as a companion article to this one [8].

The present Publishers Roadmap is organized as a set of proposed actions for publishers, applicable to each point in the lifecycle of a research article: Pre-submission, Submission, Production, and Publication.

1. Pre-submission

Revise editor training and advocacy material

Editor advocacy and training material should be revised. This may differ by journal or discipline, and whether there are in-house editors, or academic editors, or both. For example, this might involve updates to the editor training material (internally maintained, for example, on PowerPoint or PDFs or externally on websites) or updates to advocacy material (see examples below). The appropriate material should be revised to enable editors to know what data citation is, why it should be done, what data to cite, and how to cite data. This should equip editors to instruct reviewers and authors on journal policy around data citation.

Examples:

http://blogs.nature.com/scientificdata/2016/07/14/data-citations-at-scientific-data/#more-3779

https://www.elsevier.com/about/open-science/research-data/data-citation

Revise reviewer training material

Reviewer training material should be revised to equip reviewers with the knowledge needed to know what data authors should cite in the manuscript, how to cite this data and how to access the underlying data to a manuscript. Training material should also communicate expectations around data review.

Example:

http://onlinelibrary.wiley.com/journal/10.1002/(ISSN)2049-6060/homepage/guidelines_for_reviewers.htm

Update information for authors

Provide guidance on what data should be shared (e.g. derived dataset versus raw data) and whether this is encouraged, strongly encouraged, or required.

Data citation will need to be implemented at a journal policy level. This should be part of a journal's wider policy on data sharing. It is recommended that this policy is discipline-specific

and should be determined by the journal community (editor, reviewers, etc.) as well as the publisher.

There are multiple levels of data policy (e.g., encouragement of data sharing, strong encouragement, mandatory data sharing).

For example, Springer Nature has implemented a range of policy levels implemented across journals at Springer Nature depending on their specific need (for more details, see http://www.springernature.com/gp/group/data-policy/policy-types). Data policies can also be defined at the publisher level https://www.elsevier.com/about/company-information/policies/research-data or domain level http://www.copdess.org/copdess-suggested-author-instructions-and-best-practices-for-journals/.

Ask authors for a Data Availability Statement (DAS), and point them to standardized text.

It is recommended that as part of data citation implementation publishers adopt standardized Data Availability Statements (DASs). DASs provide a statement about where data supporting the results reported in a published article can be found, including, where applicable, hyperlinks to publicly archived datasets analyzed or generated during the study. Some research funders, including Research Councils UK, require data availability statements to be included in publications so it is important data policies include this. It is highly recommended that publicly available datasets referred to in DASs are also cited in reference lists.

Example:

http://www.springernature.com/gp/group/data-policy/data-availability-statements

Specify a policy for data citation.

Authors can be encouraged to cite data or can be required to cite data. Authors should provide details of previously published major datasets used and also major datasets generated by the work of the paper. The policy should specify which datasets to cite (e.g., underlying data versus relevant data not used for analysis) and how to format data citations. It is recommended if at all possible that data citation occurs either in the standard reference list or (less preferable) in a separate list of cited data, formatted similarly to standard literature references. But regardless of where citations appear in the manuscript, they should be in readily parsable form.

Specify how to format data citations

There are several ways data can be linked from ("cited" in) scholarly articles: reference lists, data availability statements and in-text mention of accession numbers. While a globally unique, machine actionable persistent identifier is needed for all three scenarios, citation metadata (authors, title, publication date, etc.) are specifically recommended for reference lists.

Whilst there are many referencing style guides, including formal standards managed by ISO/BS (ISO 690-2010) and ANSI/NISO (NISO Z39.29-2005 R2010), several of the key style guides provide guidance on how to cite datasets in the reference list. In addition, the reference should also include "[dataset]" within the reference citation so that it becomes easily recognizable within the production process. This additional tag will not be visible within the reference list of the article. It is critical to ensure the recommended format of the data citation also adheres to the Joint Declaration of Data Citation Principles. Publishers should provide an example of a reference to a dataset in their references formatting section, e.g.:

Numbered style:

[dataset] [27] M. Oguro, S. Imahiro, S. Saito, T. Nakashizuka, Mortality data for Japanese oak wilt disease and surrounding forest compositions, Mendeley Data, v1, 2015. http://doi.org/10.17632/xwj98nb39r.1.

Harvard style:

[dataset] Oguro, M., Imahiro, S., Saito, S., Nakashizuka, T., 2015. Mortality data for Japanese oak wilt disease and surrounding forest compositions. Mendeley Data, v1. http://doi.org/10.17632/xwj98nb39r.1.

Vancouver style:

[dataset] [27] Oguro M, Imahiro S, Saito S, Nakashizuka T. Mortality data for Japanese oak wilt disease and surrounding forest compositions, Mendeley Data, v1; 2015. http://doi.org/10.17632/xwj98nb39r.1.

APA style:

[dataset] Oguro, M., Imahiro, S., Saito, S., Nakashizuka, T. (2015). *Mortality data for Japanese oak wilt disease and surrounding forest compositions*. Mendeley Data, v1. http://doi.org/10.17632/xwj98nb39r.1.

AMA style:

[dataset] 27. Oguro M, Imahiro S, Saito S, Nakashizuka T. Mortality data for Japanese oak wilt disease and surrounding forest compositions, Mendeley Data, v1; 2015. http://doi.org/10.17632/xwj98nb39r.1.

Provide guidance around what is a suitable repository (general, institutional, and subject-specific) and how to find one

Publishers should provide or point to a list of recommended repositories for data sharing. Many publishers already maintain such a list. The Registry of Research Data Repositories (Re3Data) is a full scale resource of registered repositories across subject areas. Re3Data provides information on an array of criteria to help researchers identify the ones most suitable for their needs

(licensing, certificates & standards, policy, etc.). For biomedical sciences, a list of recommended repositories is provided by Biosharing.org, here.

Where a suitable repository does not exist for a given discipline or subject area, publishers should provide guidance for the use of a general or institutional repositories by authors where these meet the recommendations of the repository draft roadmap guidance [8] (briefly, by providing authors' datasets with a globally resolvable - ideally DataCite DOI - unique Digital Object Identifier, providing a suitable landing page, and using open licenses).

Some research funders may stipulate that data is deposited in a domain-specific repository where possible; again, publisher lists of recommended repositories should reflect this. Funders of biomedical research may require data to be deposited in domain specific repositories such as GEO, dbGAP, and SRA, which use locally resolvable accession numbers in lieu of DOIs. The European Bioinformatics Institute (EBI) and the California Digital Library (CDL) maintain a common list of namespace prefixes for such repositories, available through GitHub:

https://github.com/identifiers-org/prefix/blob/master/prefix.yaml.

These prefixes should be prepended to the accession number in the following format: prefix:accession, and formed into an http URI by including the resolver address, e.g.

https://identifiers.org/GEO:GDS5157 (EBI resolver)

https://n2t.net/GEO:GDS5157 (CDL resolver)

The goal is that data references should be able to be readily resolved by software agents.

Examples of publisher-maintained recommended repositories include:

- http://journals.plos.org/plosbiology/s/data-availability#loc-recommended-repositories
- http://www.springernature.com/gp/group/data-policy/repositories
- https://www.elsevier.com/books-and-journals/enrichments/data-base-linking/supported-data-repositories
- https://copdessdirectory.osf.io

Consider licensing included under "publicly accessible" and implications (e.g. automated reuse of data).

Publishers should consider the types of licensing allowed under their data policy. It is recommended that data submitted to repositories with stated licensing policies should have licensing that allows for the free reuse of that data, where this does not violate protection of human subjects or other overriding subject privacy concerns.

Update guidelines for internal customer services queries and provide author FAQs

Publishers will need to include a support service around their data policy. This might include a list of author-focused FAQs. Internal FAQs should also be provided to customer services. Alternatively or in addition, publishers might set up a specific email address for queries concerning data. PLOS, Springer Nature and Elsevier provide such email addresses.

Examples of author FAQs:

- PLoS: http://journals.plos.org/plosbiology/s/data-availability#loc-faqs-for-data-policy
- Springer Nature: http://www.springernature.com/gp/group/data-policy/faq
- Elsevier: https://www.elsevier.com/about/company-information/policies/research-data/research-data-faqs

2. Submission

Capture data citations at point of article submission.

At the submission stage it is important that all the elements are captured that are needed to create a data citation: author(s), title, year, version, data repository, persistent unique identifier.

The recommended way of capturing data citations is by asking authors to include these in the reference list of the manuscript

Instructions for formatting can be found in the pre-submission section. Formatting will depend on the reference style of the journal, but in all cases, datasets should be cited in the text of the manuscript and the reference should appear in the reference list.

To ensure data references are recognized, authors should indicate through the addition of [dataset] that this is a data reference.

When data citations are captured through direct in-text links, these should be parsable

It is already common practice to provide direct in-text links to datasets in data repositories.

If datasets are cited in this way, it is important that the links are checked (data citation identifier resolves) as part of the production process and that the links are parsable.

Data availability should be captured in a structured way

In situations where data cannot be made publicly available, authors should be given the option to make a statement about the availability of their data at the submission stage. The JATS4R group is currently working on tagging for data availability statements.

Editors and reviewers are enabled to check the data citation and underlying data at the submission stage

Through the data citation, editors and reviewers can access underlying datasets. Reviewer forms should be updated with information on how to access the data and a question about whether data sharing standards/policies have been met.

Data citations are processed in the same way as other references

When data citations are captured in the reference list of the manuscript, these can be processed in the same way as other references. This means that formatting and quality control will take place at the production stage.

3. Production

The main relevant components of the production process are the input from the peer review process (typically author manuscript in Word or LaTex files), conversion of this to XML and other formats (such as PDF, ePub), and the author proofing stage.

Following all the preceding recommendations for the editorial process, the production process needs to identify relevant content and convert to XML.

Data citations

The production department and their vendor(s) will be required to ensure all data citations provided by the author in the reference list are processed appropriately using the correct XML tags. Typesetters must be provided with detailed instructions to achieve this.

It is out of the scope of this project to provide tools to identify cited datasets that are not also present in the reference list; however, simple search and find commands can be executed using common terms and common database names and lists of these can be provided. This group can also provide boilerplate author queries for common instances.

XML requirements of data citations

For publishers using NISO standard JATS, version 1.1d3 and upwards, JATS4R recommendation on data citations should be followed. The main other publisher-specific DTDs contain similar elements to allow for correct XML tagging.

Examples:

https://github.com/elifesciences/XML-mapping/blob/master/elife-00666.xml

JATS4R recommendation and examples: http://jats4r.org/data-citations

Data availability statement (DAS)

Output format from the editorial process will inform the production department as to how to identify and handle this content. For instance, some publishers require authors to provide the details within the submission screens and thus can output XML to production, others require a separate Word file to be uploaded, and others request the author's manuscript file contains this information. Depending on the method used, production will need to process and convert this content accordingly.

Where the DAS will be contained/displayed within the PDF/ePub format of the article is decided by the individual publisher and this group will not provide recommendations for this.

The XML output of the DAS requires work by the XML component of this Force11 working group. It is anticipated that this work will be carried out by the JATS4R group or as a break-off group of JATS4R, potentially requesting changes to the DTD by the JATS Standing Committee.

4. Publication

Display Data Citations in the article

There are two primary methods of displaying data citations in a manuscript--in a separate data citations section or in the main references section. A separate data citations section promotes visibility, but inclusion in the main references section helps establish equal standing between data citations and standard references and is strongly recommended.

Data citations should include a persistent identifier (such as a DOI) and should ideally include the minimum information recommended by DataCite and the Force11 data citation principles. Where possible, persistent identifier should be favored over URLs, and they should function as links that resolve to the landing page of the dataset.

Optionally, some publishers may choose to highlight the datasets on which the study relies by visualizing these in a side panel or alternate tab.

Data Availability Statements

If a journal has implemented Data Availability Statements (DAS) as part of their required declarations, ensure this is rendered in the article. If persistent IDs are given to datasets in this, they should be expressed as full URLs. Any datasets mentioned in the DAS should link to a citation in the references.

Example:

http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002297

Downstream delivery to Crossref

For any data cited in the reference list or DAS, provided there is a DOI, these can be supplied to Crossref simply as bibliographical data. However, for data citations that do not resolve to a DOI, but another accession number, these can be deposited using the relation type tagging to ensure persistent linking between the items. More information can be found in this blogpost: https://support.crossref.org/hc/en-us/articles/215787303-Crossref-Data-Software-Citation-Deposit-Guide-for-Publishers.

Conclusions

This roadmap originated through the implementation phase of a project aimed at enhancing the reproducibility of scientific research and increasing credit for and reuse of data through data citation. The project was organized as a series of Working Groups in FORCE11 (http://force11.org/), an international organization of researchers, funders, publishers, librarians, and others seeking to improve digital research communication and eScholarship.

The effort began with the Joint Declaration of Data Citation Principles [1, 7], which distilled and harmonized conclusions of significant prior studies by science policy bodies on how research data should be made available in digital scholarly communications. In the implementation phase (the Data Citation Implementation Pilot, https://www.force11.org/group/dcip), repositories, publishers, and data centers formed three working groups, respectively, with the aim of creating clear recommendations for implementing data citation in line with the JDDCP.

In a series of teleconferences over 4 months, major publishers compared current workflows and processes around data citation. Challenges were identified and recommendations structured according to the publisher workflow were drafted. In July 2016 this group met with additional representatives from publishers, researchers, funders, and not-for-profit open science

organizations in order to resolve remaining challenges, validate recommendations, and to identify future tasks for development. From this the first full draft of the Publisher Roadmap was created. Feedback was then solicited and incorporated from other relevant stakeholders in the community as well as the other Data Citation Implementation Pilot working groups.

Several publishers are now in the process of implementing the JDDCP in line with the steps described in this roadmap. More work is still needed, both by individual publishers and by this group. This document describes basic steps that should be taken to enable authors to cite datasets. As a next step, improved workflows and tools should be developed to automate data citation further. In addition, authors need to be made aware of the importance of data citation and will require guidance on how to cite data. Ongoing coordination amongst publishers, data repositories, and other important stakeholders will be essential to ensure data is recognized as a primary research output.

Author contributions

Helena Cousijn and *Amye Kenall* co-chaired the DCIP Publishers Expert Group which produced this article. They had primary responsibility for leading regular telecons as well as a face-to-face meeting of participants (see Acknowledgements) at the SpringerNature London campus in July of 2016. Drs. Cousijn and Kenall provided the article structure; organized their Expert Group to collect and integrate information from the participating publishers, including their own organizations; and did the majority of writing for this article. They made equal contributions to the work.

Emma Ganley, Patrick Polischuk, Melissa Harrison, David Kernohan, and *Fiona Murphy* participated in the work of the Publishers Expert Group and co-authored this article. They provided knowledgeable content and input to the work from the perspectives of their respective organizations. In addition, *Melissa Harrison* coordinated and informed this work with the perspective of the JATS4R group (Journal Article Tag Suite for Reuse), which she chairs.

Tim Clark coordinated the work of the Publishers Expert Group with the other DCIP participants (Repositories, Identifiers, JATS, and Primer/FAQ), co-authored sections of this article and edited the whole.

Tim Clark and *Maryann Martone* co-led the Data Citation Implementation Pilot as a whole.

Acknowledgments

Research reported in this publication was supported in part by the National Institutes of Health under award number U24HL126127. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

The authors gratefully acknowledge the following members of the FORCE11/bioCADDIE Data Citation Pilot Publishers Expert Group, who participated in workshops and/or telecons to assist in developing this Roadmap: Helen Atkins (Public Library of Science); Paul Donohoe (SpringerNature); Martin Fenner (DataCite); Scott Edmunds (GigaScience); Ian Fore (National Cancer Institute, National Institutes of Health); Carole Goble (University of Manchester); Florian Graef (European Bioinformatics Institute); Iain Hrynaszkiewicz (SpringerNature); Thomas Lemberger (European Molecular Biology Organization); Johanna McEntyre (European Bioinformatics Institute); Ashlynne Merrifield (Taylor and Francis); Eleonora Presani (Elsevier); Perpetua Socorro (Frontiers); Michael Taylor (Digital Science), Simone Taylor (John Wiley & Sons, Inc).

References

1. Data Citation Synthesis Group. 2014. Joint declaration of data citation principles. http://force11.org/datacitation (2014).

2. Board on Research Data and Information, Policy and Global Affairs & National Research Council (U.S.). For attribution -- developing data attribution and citation practices and standards: summary of an international workshop. *Washington: National Academies Press* http://site.ebrary.com/lib/alltitles/docDetail.action?docID=10863947 (2012).

3. CODATA-ICSTI Task Group on Data Citation Standards and Practices. Out of cite, out of mind: the current state of practice, policy, and technology for the citation of data. *Data Science Journal* **12**, CIDCR1–CIDCR7 (2013) http://doi.org/10.2481/dsj.OSOM13-043.

4. Hodson, S. & Molloy, L., Current best practice for research data management policies. Preprint at https://doi.org/10.5281/zenodo.27872 (2015).

5. National Academies of Science, Committee on Ensuring the Utility and Integrity of Research Data in a Digital Age. Ensuring the integrity, accessibility, and stewardship of research data in the digital age. https://www.ncbi.nlm.nih.gov/books/NBK215264/pdf/Bookshelf_NBK215264.pdf (2009).

6. Wilkinson, M. D.et al. . The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**,160018 (2016) http://doi.org/10.1038/sdata.2016.18.

7. Starr, J. et al. Achieving human and machine accessibility of cited data in scholarly publications. PeerJ Comput. Sci. **1(e1)** PMID: 26167542 (2015) http://doi.org/10.7717/peerj-cs.1.

8. Fenner M & Crosas M.. et al. A Data Citation Roadmap for Scholarly Data Repositories. Preprint at https://doi.org/10.1101/097196 (2017).