A Data Citation Roadmap for Scientific Publishers

Helena Cousijn[1]*, Amye Kenall[2]*, Emma Ganley[3], Melissa Harrison[4], David Kernohan[5], Thomas Lemberger[8], Fiona Murphy[6], Patrick Polischuk[3], Simone Taylor[7] Maryann Martone[9], Tim Clark[10,11]

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. John Wiley & Sons, Inc., Hoboken, NJ, USA
8. EMBO Press, Heidelberg, Germany
9. University of California, San Diego, La Jolla CA, USA
10. Massachusetts General Hospital, Boston MA, USA
11. 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. It can also act as a guide to authors - what to expect from publishers and how to enable their own data citations to gain maximum impact, as well as complying with what will become increasingly common funder mandates on data transparency.

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 accessibility practices for digital repositories.

There is a clear emerging consensus in the scholarly community, including researchers, funders and publishers, supporting the practice of data archiving and citation. This is reflected not only in the broad endorsement of the JDDCP, but also in the increasing proliferation of workshops on this topic. At least one journal, which had earlier published a widely discussed editorial by clinical trialists arguing against openly sharing data, is now leading an effort to help provide institutional incentives for authors to share and cite data [8].

There is also some evidence to suggest that researchers, primarily to enhance the visibility and impact of their work, but also to facilitate transparency and encourage re-use, are increasingly sharing their own data, and are making use of shared data from other researchers [9]. While intellectual property and confidentiality continue to be the main issues researchers mention as prohibitive to sharing, researchers are also concerned about receiving appropriate citation credit or attribution. A standardized route to clear and accessible data citation practices should alleviate this concern.

As participants in the science communications ecosystem (especially funders) increasingly harmonize their views in this direction, publishers are adapting their workflows to enable data citation practices and to provide tools and guidelines that improve the implementation process for authors, and relieve stress points around compliance. Authors will need to know, in general terms, what to expect from publishers supporting data citation when they submit articles based on funding that requires open data. We hope, in the long run, that open data will become a common enough practice so that all authors will eventually expect to provide it and cite it.

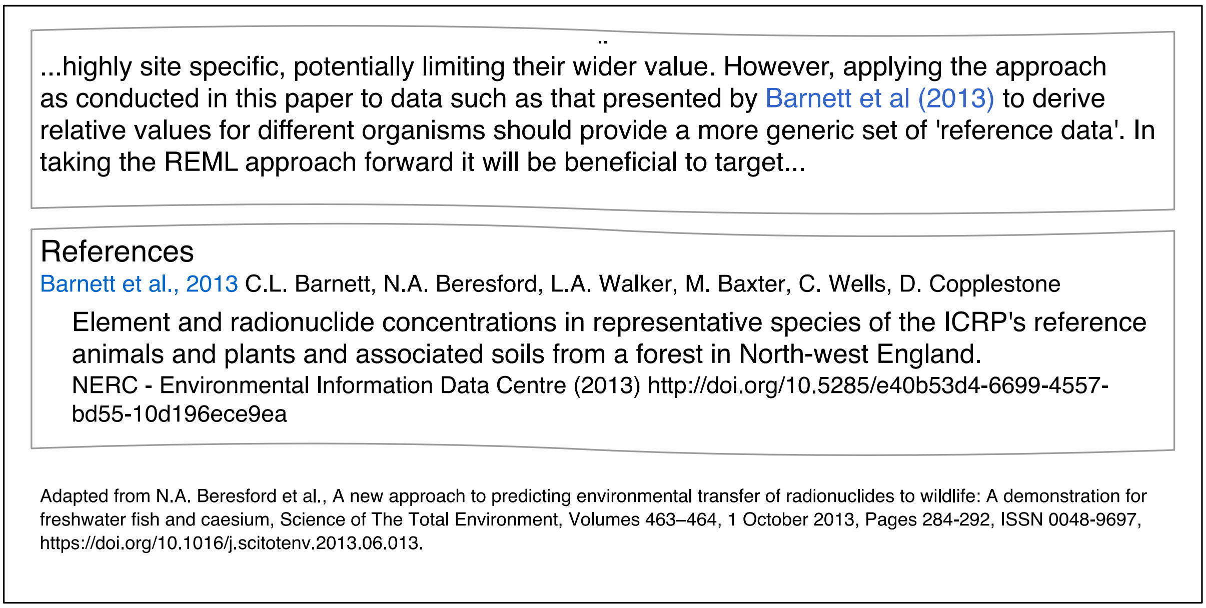
Implementing data citation is not meant to replace or bypass citation of the relevant literature, but rather to ensure we provide verifiable and potentially re-usable backing data for published assertions. It is aimed at significantly improving the robustness and reproducibility of science, which have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

The present document is a detailed roadmap to implementing JDDCP-compliant data citation, prepared by publishers, for an audience of publishers and authors, as part of a larger effort involving roadmap and specification development for and by repositories, informaticians, and identifier / metadata registries [10,11].

Results

This section briefly explains data citations and presents implementation recommendations for publishers, editors and scholarly societies[1]. 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 has been robustly archived in a long-term-persistent repository. This approach supersedes "Supplemental Data" as a home for article-associated datasets. It is designed to make data fully FAIR (Findable, Accessible, Interoperable and Reusable).

Publishers implementing data citation will have domain-specific lists of acceptable repositories for this purpose, or will guide authors to sites that maintain these lists. We provide examples of some of these lists further along in the manuscript. Figure 1 illustrates a data citation. Figure 2 shows the resolution structure from data citations, to dataset landing pages, and to archived data.



...highly site specific, potentially limiting their wider value. However, applying the approach as conducted in this paper to data such as that presented by Barnett et al (2013) to derive relative values for different organisms should provide a more generic set of 'reference data'. In taking the REML approach forward it will be beneficial to target...

References

Barnett et al., 2013 C.L. Barnett, N.A. Beresford, L.A. Walker, M. Baxter, C. Wells, D. Copplestone

Element and radionuclide concentrations in representative species of the ICRP's reference animals and plants and associated soils from a forest in North-west England.

NERC - Environmental Information Data Centre (2013) http://doi.org/10.5285/e40b53d4-6699-4557-bd55-10d196ece9ea

Adapted from N.A. Beresford et al., A new approach to predicting environmental transfer of radionuclides to wildlife: A demonstration for freshwater fish and caesium, Science of The Total Environment, Volumes 463–464, 1 October 2013, Pages 284-292, ISSN 0048-9697, https://doi.org/10.1016/j.scitotenv.2013.06.013.

Figure 1. Data citation example.

Both the dataset reference in the primary article (including its globally resolvable unique identifier), and the archival repository, should follow certain conventions. These are ultimately based upon the JDDCP's eight principles. Initial conventions for Repositories were developed in *Starr et al 2015* [7], and are presented in more depth and detail in the Roadmap for Repositories provided as a companion article to this one [10]. Another companion article outlines special identifier conventions for biomedical data repositories utilizing the "compact identifier" (prefix:accession) format in lieu of Digital Object Identifiers (DOIs) [11].

[1] Although throughout this roadmap we refer to implementation falling under the remit of the publisher, due to the diversity of publishing models, this might not always be the case. Where an aspect of implementation falls to another party (eg, a society journal where journal policy would often be set by the society), approval of and participation in implementation from that party would be needed. In addition, relevant research communities that use specific journals will often have a role in shaping those journals' policies so that they support rather than frustrate the overall objectives (more sharing of better quality datasets that are properly accredited).

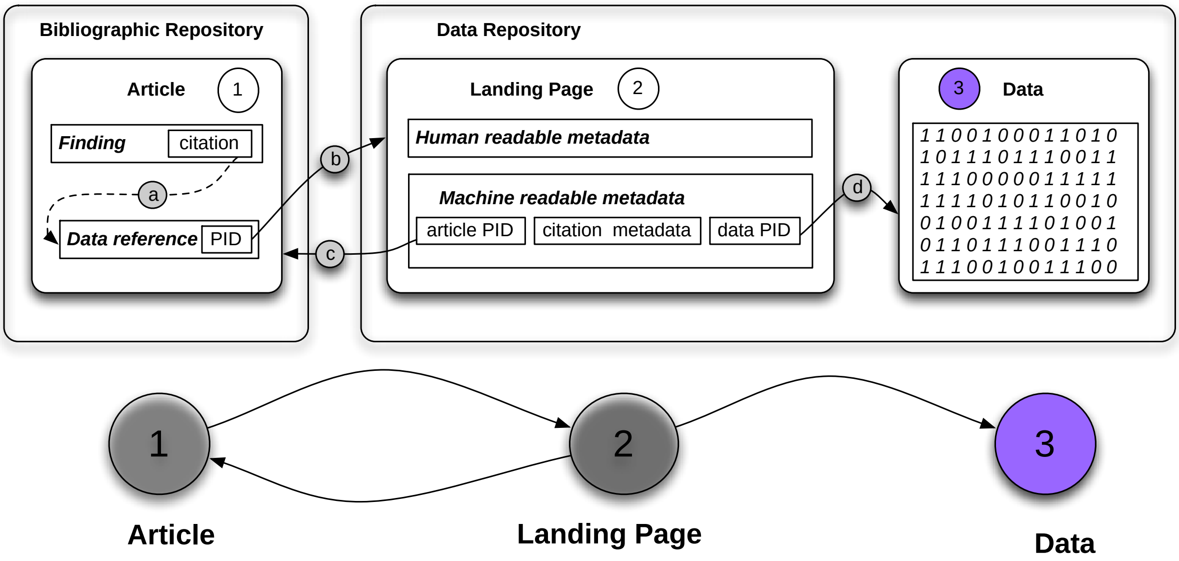


Figure 2. Data citation resolution structure. Articles (1) link to robustly archived datasets, on which their conclusions are based, through citation to a dataset (a), whose persistent unique identifier (PID) resolves (b) to a landing page (2) in a well-supported data repository. The data landing page contains human- and machine-readable metadata, to support search and to resolve (c) back to the citing article, and (d) a link to the data itself (3).

The 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

1.1 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

1.2 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

1.3 Update information for authors

Provide guidance on author responsibilities.

Data citation is based on the idea that the data underlying scientific findings or assertions should be treated as a first-class research object. This begins with author responsibility to properly manage their own data prior to submission.

Specify a policy for data citation.

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, since it is discipline-specific, 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 (http://www.springernature.com/gp/group/data-policy/policy-types). Both Wiley and Elsevier are in the process of rolling out a multi-level data sharing policy for its titles. In addition, data policies can also be defined at the domain level (http://www.copdess.org/copdess-suggested-author-instructions-and-best-practices-for-journals/).

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.

Ask authors for a Data Availability Statement (DAS).

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. In addition DASs can increase transparency by providing a reason why data cannot be made available. 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 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 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 the tag "[dataset]" within the reference citation so that it becomes easily recognizable within the production process. This additional tag does not have to 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 the the in-text citation and of the 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.

[dataset] [28] D. Deng, C. Xu, P.C. Sun, J.P. Wu, C.Y. Yan, M.X. Hu, , N. Yan, Crystal structure of the human glucose transporter GLUT1, Protein Data Bank, 21 May 2014. http://identifiers.org/pdb:4pyp.

Harvard style:

[dataset] Farhi, E., Maggiori, M., 2017. "Replication Data for: 'A Model of the International Monetary System'", Harvard Dataverse, V1. https://doi.org/10.7910/DVN/8YZT9K.

[dataset] Aaboud, M, Aad, G, Abbott, B, Abdallah, J, Abdinov, O, Abeloos, B, AbouZeid, O, Abraham, N, Abramowicz, H, Abreu, H., 2017. Dilepton invariant mass distribution in SRZ. HEPData, 2017-02-08. https://doi.org/10.17182/hepdata.76903.v1/t1.

Vancouver style:

[dataset] [52] Wang G, Zhu Z, Cui S, Wang J. Data from: Glucocorticoid induces incoordination between glutamatergic and GABAergic neurons in the amygdala. Dryad Digital Repository, August 11, 2017. http://dx.doi.org/10.5061/dryad.k9q7h.

[dataset] [17] Polito VA, Li H, Martini-Stoica H, Wang B et al. Transcription factor EB overexpression effect on brain hippocampus with an accumulation of mutant tau deposits. Gene Expression Omnibus, December 19, 2013. http://identifiers.org/GEO:GDS5303.

APA style:

[dataset] Golino, H., Gomes, C. (2013). *Data from the BAFACALO project: The Brazilian Intelligence Battery based on two state-of-the-art models: Carroll's model and the CHC model*. Harvard Dataverse, V1, https://doi.org/10.7910/DVN/23150,

[dataset] Morris, P. (2017) *Head Start CARES Demonstration: National Evaluation of Three Approaches to Improving Preschoolers' Social and Emotional Competence, 2009-2015*. *ICPSR35510-v3.* Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2017-03-06. https://doi.org/10.3886/ICPSR35510.v3

AMA style:

[dataset] 12. Kory Westlund, J. Measuring children's long-term relationships with social robots. Figshare, v2; 2017. https://doi.org/10.6084/m9.figshare.5047657.

[dataset] 34. Frazier, JA, Hodge, SM, Breeze, JL, Giuliano, AJ, Terry, JE, Moore, CM, ... Makris, N. CANDI Share Schizophrenia Bulletin 2008 data; 2008. Child and Adolescent NeuroDevelopment Initiative. http://dx.doi.org/10.18116/C6159Z.

Provide guidance around suitable repositories (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.). A list of recommended repositories is provided by FAIRsharing.org.

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 unique identifier - ideally a DataCite DOI where possible - 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.

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
- http://emboj.embopress.org/authorguide#datadeposition
- https://www.elsevier.com/authors/author-services/research-data/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

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 globally 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 has produced a draft recommendation for tagging data availability statements (http://jats4r.org/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. Publishers should be mindful that they do not reveal the identity of the authors in cases where peer-review is double-blind.

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.

DOIs and Compact Identifiers

Digital Object Identifiers (DOIs) are well understood by publishers based on their role as document identifiers. DOIs are also assigned by many repositories to identify datasets. When available, they should be included in the data reference similarly to the use of DOIs for article references. An advantage to DOIs for data is that the associated metadata is centrally managed by the DataCite organization, similarly to how Crossref manages article metadata. DataCite and Crossref collaborate closely.

Many domain-specific repositories in biomedical research do not issue DOIs, but instead issue locally-assigned identifiers (“accessions”, “accession numbers”). Funders of biomedical research may require data to be deposited in domain specific repositories e.g. GEO, dbGAP, and SRA, many of which use such locally resolvable accession numbers in lieu of DOIs.

Prior informal practice had been to qualify these by a leading prefix, so that the identifier becomes unique. In 2012 the European Molecular Biology Laboratory-European Bioinformatics Institute (EMBL-EBI) began tracking and issuing formal namespace prefixes to avoid collisions and support formal resolution on the Web [12]. Subsequent efforts developed a collaborative curation model [13].

Now EMBL-EBI and the California Digital Library (CDL) maintain a common shared namespace registry and resolvers capable of interpreting and resolving PREFIX:ACCESSION patterns, or “compact identifiers”, hosted at these leading institutions [10]. Technical work to develop the common repository and resolution rules was coordinated with the work of our Publishers Roadmap team.

This means that compact identifiers have now been formalized, are institutionally supported in the U.S. and in Europe**,** and may be used by in place of DOIs. We recommend this be done (1) where the repository does not issue DOIs for deposited datasets and (2) where the repository’s prefix has been registered. Similarly to DOIs, compact identifiers are dereferenced by resolvers hosted at well known web addresses: http://identifiers.org (EMBL-EBI) and http://n2t.net (CDL).

Examples:

https://identifiers.org/GEO:GDS5157

and https://n2t.net/GEO:GDS5157

both resolve the Gene Expression Omnibus local accession number GDS5157.

While these resources are still under active development to resolve an increasing number of identifiers, ensuring that either a DOI or a Compact Identifier is associated with data references will be important to support automatic resolution of these identifiers by software tools, which benefits authors, data providers and service providers.

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.

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 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 identifiers 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.

Data Availability Statements

If a journal has implemented Data Availability Statements (DAS) as part of its required declarations, ensure this is rendered in the article. A JATS4R DAS subgroup has recently put forward a recommendation to the JATS Standing Committee to provide a tag specifically for the DAS to ensure it's contained within open parts of content for subscription controlled journals and so the content is machine readable.This recommendation has been published in Draft format and is available here {http://jats4r.org/data-availability-statements}.

Example:

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

Downstream delivery to Crossref

Crossref ensures that links to scholarly literature persist over time through the Digital Object Identifier (DOI). They also provide infrastructure to the community by linking the publications to associated works and resources through the metadata that publishers deposit at publication, making research easy to find, cite, link, and assess. Links to data resources (i.e., data citations) are a core part of this service.

Publishers deposit the data citations by adding them directly into the standard metadata deposit via references and/or relation type. This is part of the existing content registration process and requires no new workflows. All data citations across journals (and publishers) are then aggregated and made freely available for the community to retrieve and reuse in one shared location, removing the need for access channels to every individual publisher.

In the metadata deposit, publishers can deposit the literature-data links in two places: bibliographic references and/or relationships component. Since each has its own distinct advantage, both are encouraged where possible (see the Data Citations Deposit Guide linked below).

1. Bibliographic references: Publishers include the data citation into the deposit of bibliographic references for each publication. Here, publishers follow the general process for depositing references and apply tags to structure the metadata as applicable.
2. Relation type: Publishers assert the data citation in an existing section of the metadata deposit dedicated to connecting the publication to a variety of research objects associated with it (e.g., data and software, supporting information, protocols, videos, published peer reviews, preprint, conference papers, etc.). In addition to structured metadata about the data, this also allows publishers to identify any data linked as a direct output of the research results (viz., for scientific validation) if this is known.

Should all of this information be sent to Crossref, extended opportunities for data mining and building up pictures of data citations, linking, and relationships will be possible. Whether the data citations come from the authors in the reference list or the DAS, or whether they are extracted by the publisher and then deposited, Crossref collects them across publishers. It then makes the aggregate set freely available in multiple interfaces and formats through its existing metadata delivery channels. This also means that the information about links between articles and datasets becomes available to other services that increase the discoverability of data and potential for reuse. For example making data citations available in a Scholix compliant way (www.scholix.org) enables their retrieval through the Data Literature Interlinking service (https://dliservice.research-infrastructures.eu/#/) - an easy way for both publishers and repositories to retrieve information about associations between articles and datasets.

More detail can be found in the Data & Software Citations Deposit Guide: https://support.crossref.org/hc/en-us/articles/215787303-Crossref-Data-Software-Citation-Deposit-Guide-for-Publishers.

Next steps

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.

Discussion

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 Expert Groups, respectively, with the aim of creating clear recommendations for implementing data citation in line with the JDDCP.

Once the steps outlined in this roadmap are implemented, authors will be able to cite datasets in the same way as they cite articles. In addition to 'author', 'year', and 'title', they will need to add the data repository, version and persistent unique identifier to ensure other researchers can unambiguously identify datasets. Publishers will be able to recognize the references as data references and process these accordingly, so that it becomes possible for data citations to be counted and for researchers to get credit for their work. These are essential steps for substaintially increasing the FAIRness [6] of research data. We believe this will in turn lead to better, more reproducible and re-usable science and scholarship, with many benefits to society.

Methods

In a series of teleconferences over a period of a year, 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.

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*, *Thomas Lemberger*, *Simone Taylor* 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); Johanna McEntyre (European Bioinformatics Institute); Ashlynne Merrifield (Taylor and Francis); Eleonora Presani (Elsevier); Perpetua Socorro (Frontiers); Caroline Sutton (Taylor and Francis); Michael Taylor (Digital Science).

References

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

2. Board on Research Data and Information, Policy and Global Affairs & National Research Council (U.S.). (2012) 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 .

3. CODATA-ICSTI Task Group on Data Citation Standards and Practices. (2013) 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 https://doi.org/10.2481/dsj.OSOM13-043.

4. Hodson S, and Molloy L. (2015, August 13). Current Best Practice for Research Data Management Policies. Zenodo. https://doi.org/10.5281/zenodo.27872

5. National Academies of Science, Committee on Ensuring the Utility and Integrity of Research Data in a Digital Age. (2009) 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 .

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

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

8. Longo, DL and Drazen JM (2016) Data Sharing, *New England Journal of Medicine*, **374**, 276-277. PMID**:** 26789876 https://doi.org/10.1056/NEJMe1516564

9. Vocile, B (2017) Open Science Trends You Need To Know About (https://hub.wiley.com/community/exchanges/discover/blog/2017/04/19/open-science-trends-you-need-to-know-about?referrer=exchanges).

10. Fenner M, Crosas M, Grethe J, Kennedy D, Hermjakob H, Rocca-Serra P, Berjon R, Karcher S, Martone M and Clark T (2016) A Data Citation Roadmap for Scholarly Data Repositories, *bioRxiv*. 097196; https://doi.org/10.1101/097196 .

11. Wimalaratne SM, Juty N, Kunze J, Janée G, McMurry JA, Beard N, Jimenez R, Grethe J, Hermjakob H and Clark T (2017) Uniform Resolution of Compact Identifiers for Biomedical Data, *bioRXiv*. 101279; https://doi.org/10.1101/101279.

12. Juty N, Le Novère N and Laibe C (2012) Identifiers.org and MIRIAM Registry: community resources to provide persistent identification, *Nucleic Acids Research*, **40**, D580-D586. https://doi.org/10.1093/nar/gkr1097.

13. Juty N, Le Novère N, Hermjakob H, Laibe C: Towards the Collaborative Curation of the Registry underlying identifiers.org. Database 2013, 2013:bat017-bat017 https://doi.org/10.1093/database/bat017