Research Data in Journals and Repositories in the Web of Science: Developments and Recommendations

Megan Force
Thomson Reuters
Philadelphia, PA, USA
megan.force@thomsonreuters.com

Nigel Robinson
Thomson Reuters
York, United Kingdom
nigel.robinson@thomsonreuters.com

Mark Matthews
Thomson Reuters
York, United Kingdom
mark.matthews@thomsonreuters.com

Daniel Auld
Thomson Reuters
York, United Kingdom
daniel.auld@thomsonreuters.com

Mariana Boletta
Thomson Reuters
Philadelphia, PA, USA
mariana.boletta@thomsonreuters.com

ABSTRACT

Data is indexed in the Web of Science^{TM} through the inclusion of data journal content in Web of Science journal databases and data repository content in the Data Citation Index^{SM}. Here we detail new developments in data journal selection and inclusion, as well as recommended practices for research authors, data providers, journal publishers, and funding organizations, to improve citation and attribution for deposited scholarly research data.

Keywords

Data Citation Index, Web of Science, data citation, data publication, data journals, data repositories, metadata

1. INTRODUCTION

The publication of an author's work in a scholarly journal has long been seen as a primary mechanism for sharing scientific and scholarly discoveries [1]. Yet in recent years there have emerged increasing efforts to improve sharing of the data which underpins research conclusions. This comes in response to issues of reproducibility and reuse of research results, as well as the proper presentation of credit to research authors, where data is created outside of the traditional publication system [2,3]. With the advent of ever more powerful computing systems for analyzing and storing data, and with an eye to future preservation, increasing amounts of research data are today deposited and housed in discipline-specific and/or multidisciplinary data repositories dedicated to the wealth of worldwide scholarly information [4]. At the same time, a new form of publication, the data journal, is developing as a bridge between traditional journal publication and data set submission [5].

The research community has identified stakeholder groups with a role to play in the research data landscape, including research authors, publishers, funders, scholarly societies, and universities [6]. These entities require a means to gather metric information on these forms of scholarly publication, with citations viewed as a primary measure of impact [7]. While citations and indexing of data papers in data journals may follow a more traditional form, data objects published to online data repositories present new challenges with respect to accurate citation and attribution. These include unique identification, persistence, and the establishment of technology, policy, and reward infrastructures [8].

In recognition of the importance of data citation as a part of the evolving research data ecosystem, Thomson Reuters seeks to address this emerging area by indexing data journals as well as data repositories in the Web of Science, a research platform for finding, analyzing and sharing curated content in the sciences, social sciences, arts, and humanities. Review of data journal content for inclusion in the Web of Science follows established evaluation criteria, with particular attention to the significant details which distinguish this new form of publication. In evaluating data repository content for inclusion, the Data Citation Index looks to build a solid foundation for data citation and improved metrics for the extended scholarly portfolio by selecting those resources which display the greatest commitment to future persistence, citability, discoverability, and reuse. The Data Citation Index now includes a growing selection of over 300 data sources, with over 5 million data objects indexed to date; data repository selection criteria for the resource may be found on the Thomson Reuters website [9].

These new developments necessitate contributions from stakeholders at each step in the data lifecycle. In the following sections, we present practices in the promotion of data citation and the dissemination of accurate bibliometric information for data. Section 2 details recommended practices for creators of data objects and metadata contributors, and in Section 3 we discuss data publisher responsibilities. Section 4 provides guidelines for literature publishers and funders, and Section 5 an overview of data journal coverage in the Web of Science.

2. DATA AUTHORS

Recent requirements for research authors to make data and research results publicly available may be unspecific with respect to data deposition [10]. As a recent survey by Wiley, the publisher, has shown, many researchers throughout the world still do not share their research, for a variety of reasons including lack of credit and concerns about confidentiality and misuse [11]. Yet even for those researchers who report sharing data, their methods of doing so may vary dramatically. Privacy issues may prevent certain data objects, such as patient data, from being openly available [12]. The data sharing practices that researchers adopt affect long-term data availability and reproducibility of results, as well as the accurate citation necessary for authors to receive credit for their work [13].

2.1 Data Product Equality and Data Citation Practices

It is a helpful first step for research authors to regard data products equally with other citable research output. Concurrent with the consideration of data objects as primary records of research, such as journal articles, is the careful practice of formal data citation in data and publications, including the citation of dataset permanent identifiers. These recommendations are in keeping with the FORCE 11 joint declaration of data citation principles, which has been endorsed by Thomson Reuters [14]. Additionally, newly created data objects should include citations to the code, data, and parent literature publications which lead to their creation [15]. Recommended data citation formats vary, often due to the particular requirements of a specific academic discipline; in order to promote the increased practice of data citation by researchers, each record in the Data Citation Index includes a recommended formal data citation, employing the format proposed by DataCite [16].

2.2 Data Deposition

Research data should be deposited in an established data repository committed to long term data preservation and the use of permanent identifiers for data. Depending on their area of study, authors may have a choice with respect to where they deposit data; there may be a number of discipline-specific data repositories available which provide guidelines for data submission, as well as institutional repositories governed by the affiliated universities of the submitting authors. If the data are deposited in conjunction with the submission of a journal article, the journal or publisher may recommend or even require data submission to a particular repository. In cases where none of these options apply, data may be submitted to a multidisciplinary repository, several of which have been established to fill the need for long-term curation of so-called 'orphan' data sets [17]. Thomson Reuters offers a searchable cross-disciplinary list of data repositories selected for coverage by the Data Citation Index [18].

2.3 Metadata Contribution

At the point of data submission, it is recommended that research authors consider issues related to the future citation and discovery of the data. Providing the metadata elements necessary for a complete data citation is essential for proper attribution as well as future metrics for data objects and software [19]. A general consensus recommends certain minimum bibliographic elements for a data citation [20]. Thomson Reuters defines these components as follows:

- **Author/Creator**: Individuals or organizations that created or contributed to the data set; this metadata element guarantees attribution and credit for data authors, and helps to provide metrics for their non-traditional scholarly output
- **Year**: The year of 'publication' of the data; when it was made publicly available, such as through deposition in an appropriate repository
- **Title**: The title of the data object, which may differ from the title of the parent research paper/project
- **Publisher**: The data repository which houses the data and/or the governing organization responsible for publishing the data (making them available)
- **Version**: Dynamic data sets or those where new editions may be issued (such as with error corrections or new values) must employ proper version control to guarantee accuracy and uniqueness in data citation
- **Permanent Identifier**: An identifier should be assigned which is unique and persistent; for example, a Digital Object Identifier (DOI); in *Data Citation Index* citations, this bibliographic element may take the form of a unique URL, databank accession number, or other permanent identifier such as Handle (hdl) [21]

Certain conventions for data object attribution may exist within a discipline, whereas other disciplines may differ with respect to appropriate credit for various data production or management roles [22].

Additionally, further metadata may be contributed to advance discovery; creation of enhanced metadata increases the possibility of data discovery through searches in the Web of Science. Data descriptions/abstracts, while not strictly required for citation purposes, provide important context as to the relevance and scope of the data or software object in question. Metadata elements such as keywords and discipline-specific indexing terms provide content details, while information related to data type, representation, methodology and licensing provide avenues for researchers to re-use data and software [8]. Author affiliations, as well as grant and funding agency information, provide for a more detailed assessment of departmental, institutional, and individual researcher output.

3. DATA PUBLISHERS

Due to their unique role in the data lifecycle, a different set of requisites apply to data publishers including discipline-specific, cross-disciplinary, and library and university affiliated data repositories. These organizations have a mandate to preserve research data for the long term, as well as make the data accessible to the scientists and scholars who need this information for further research [19]. The recommendations listed here encourage the establishment of citable, unique, and accessible data for researchers as well as indexing services such as the Data Citation Index. Thomson Reuters evaluates resources for coverage using criteria such as those listed below; however, other factors such as evidence of long-term persistence and mentions of the resource in the literature also contribute to the decision to select a repository for inclusion.

3.1 Permanent Identifiers and Metadata Curation

Research authors generally cannot provide all required metadata elements for a complete data citation. In particular, permanent identifiers applied by the data center or publisher are likely more unique, as well as a better guarantee of precise identification and citation [23]. Those required citation elements that a data repository or publisher cannot provide should be gathered from authors upon data object submission, with metadata curated by the repository for accuracy. Validation checks for completeness and consistency with established metadata requirements help to ensure clear attribution for data objects [24].

3.2 Landing Pages and Versioning

There should be a mechanism in place by which data submitted by research authors to a data repository may be retrieved by outside researchers and machines. The clearest means to accomplish this is for data providers to provide unique descriptive landing pages for data objects, where the unique URIs (such as DOIs) provided in the metadata resolve or link to landing pages where the steps required for data access are detailed [24]. Additionally, accurate reproduction of results requires identification of the precise data set used [6]. This is particularly true in cases where the data set in question is in a constant state of change due to continuous scientific readings, or where a database is regularly updated with new information. This may be achieved through attention to versioning for the data or software, or through providing information as to the date on which the object was accessed or utilized. A dedicated working group at the Research Data Alliance has recently released recommendations for persistent identification of evolving data, including time stamps, versioning, and query storage [25]. By maintaining detailed update information, new versions of data objects may be identified, while noting the importance of clarity in the distinction between version of the data and metadata.

3.3 Metadata Harvesting and Data Resource Types

Indexing is made simpler where the database provides a metadata harvesting facility at a programmatic access point, such as the Open Archive Initiative's Protocol for Metadata Harvesting (OAI-PMH) [26]. Where the repository maintains detailed information regarding new, updated, and deleted records, the content as reflected by indexing services will remain accurate when updates are performed. For repositories which hold multiple types of media,- such as software, data sets, journal publications and theses,- indicating the type of resource, whether in the metadata or at the metadata access point, assists indexing services in identifying and filtering various types of material from other scholarly output. This is particularly important for repositories dedicated to preserving the output of an academic institution, as these may primarily hold published literature.

3.4 Mission Statement and Stewardship

Researchers, publishers, funders, and other entities should be fully informed when depositing data or recommending a repository for deposition. This is assisted by a well-documented repository mission, including long-term plans for data retention and funding, as well as a contingency plan for loss of funding or technical failures. Community guidelines for data publishers to address data reuse issues through improved machine and human accessibility are currently in development [27]. The repository may also wish to document their policies for inclusion, so that researchers may be aware of the standards employed for data acceptance, as well as which formats of data are appropriate for submission. Resources including ISO standards for certification of trustworthy repositories [28] and open archival information systems [29] are available to provide guidance in identifying a resource for data curation and future stewardship, while a Data Seal of Approval-World Data System working group has developed a draft Catalogue of Common Requirements for repository audit and certification [30].

4. LITERATURE PUBLISHERS AND FUNDING ORGANIZATIONS

The drive to share research data is increasing in response to open access mandates from funding organizations throughout the world [31]. These still-developing requirements often vary between funders, as do requirements for data deposition from journals and publishers [32,33]. Collaboration between journals, publishers, funding organizations, and data repositories on data submission and citation standards is necessary to resolve outstanding issues related to technology and policy, including the resolution of different types of permanent identifier and duplicate coverage of data objects [34].

4.1 Guideline Development

Literature publishers, journals, and funders are developing guidelines for data deposition, which may include recommended repositories for data submission, as well as openness and availability of data [35]. Stakeholders are encouraged to define specific data publishing guidelines which make journal priorities clear for authors, as well as to evaluate recommended repositories for their potential for long-term preservation and onward data citation. Scholarly societies are encouraged to contribute to the dissemination of proposed standards, both in associated published journals as well as with their constituent member researchers.

4.2 Citation Policies and Metadata Criteria

Formal data citation practices are encouraged as journals and publishers increasingly devise policies which require the practice from submitting authors [15]. Many of these publisher guidelines currently do not specify a location within the paper for data citations, such as the reference section [36]. Author guidelines should detail preferred data citation formats and metadata requirements as selected by the publisher or journal and necessitated by the data requirements of discipline. Those funding organizations which require data sharing as a condition of grant awards may enforce those requirements by developing checks to assess compliance, with possible consequences for those researchers whose data remain inaccessible.

5. DATA JOURNALS

Data journal articles associated with deposited data may also be indexed in the Web of Science. The Web of Science defines data journals as consisting primarily of articles describing data (data papers), which include a citation or link to the data; the data set may also be included within the article. These articles may contain details on the methods used to collect the data, or remarks supporting the quality of the measurements described. However, some traditional evaluation criteria do not apply to these journals, as these articles may lack such elements as the introduction of scientific hypotheses, new insights, analysis, and conclusions about the material. While some data journals feature a dedicated underlying archive for deposition of data objects, others may have no governance of associated data repositories, and therefore no control over long-term data preservation. The practice of peer review of data is an indicator of quality, yet disciplinary-specific challenges remain with regard to the details of assessment [37]. While Web of Science editors are aware of the differences between a traditional research article and the more recent document type developing as data paper, evaluation and coverage of these articles will for now be the same as for any other article type or journal. As this is a relatively new phenomenon, the content team will continue to monitor developments and collect data on aspects such as impact, citation behavior, and other possible metrics over a period of time, in order to decide if and how this publication model may need special treatment.

6. SUMMARY

Access to the data which underlies scientific and scholarly research is key to both reproducibility of results as well as further discoveries based upon previous observations. Publication of data in a dedicated repository, as well as publication of articles about the data in traditional and data-specific journals, creates a means by which credit may be conferred upon data product authors.The Web of Science is including data journal as well as data repository content with the aim of comprehensive coverage of research data resources. In this emerging landscape, a standardized set of best practices continues to be developed, necessitating careful attention and collaboration between academic, governmental, public, and private stakeholders. This concerted approach will ensure clarity, accuracy, and efficiency in future data publication and indexing.

7. REFERENCES

[1] Committee on Science, Engineering, and Public Policy, Institute of Medicine, Policy and Global Affairs, National Academy of Sciences, National Academy of Engineering. 2009. On Being a Scientist: A Guide to Responsible Conduct in Research: Third Edition. *National Academies Press*. DOI= http://dx.doi.org/10.17226/12192

[2] Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science*. 349, 6521 (Aug. 2015). DOI= http://dx.doi.org/10.1126/science.aac4716.

[3] Costello, M. J., Michener, W. K., Gahegan, M., Zhang, Z. Q., and Bourne, P.E. 2013. Biodiversity data should be published, cited, and peer reviewed. *Trends Ecol. Evol*. 28, 8 (Aug. 2013), 454-461. DOI= http://dx.doi.org/10.1016/j.tree.2013.05.002.

[4] Research Information Network. 2011. Data Centres: their use, value and impact. *Research Information Network*. http://www.rin.ac.uk/data-centres.

[5] Callaghan, S., Murphy, F., Tedds, J., Allan, R., Kunze, J., Lawrence, R., Whyte, A. 2013. Processes and procedures for data publication: A case study in the geosciences. *International Journal of Digital Curation*. 8, 1, 193–203. DOI= http://dx.doi.org/10.2218/ijdc.v8i1.253

[6] Sices, C.-I.T.G. on D.C.S. and P. of C.O. of M.T.C., (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. DOI= http://doi.org/10.2481/dsj.OSOM13-043

[7] Kratz, J. E. and Strasser, C. 2015. Making data count. *Sci. Data*. 2 (Aug. 2015). DOI= http://dx.doi.org/10.1038/sdata.2015.39

[8] National Academy of Sciences. 2012. For Attribution – Developing Data Attribution and Citation Practices and Standards: Summary of an International Workshop. *National Academies Press*. http://www.nap.edu/catalog.php?record_id=13564

[9] Thomson Reuters. 2012. Repository Evaluation, Selection, and Coverage Policies for the Data Citation Index Within Thomson Reuters Web of Science. *Thomson Reuters*. http://wokinfo.com//products_tools/multidisciplinary/dci/selection_essay/

[10] Vines, T. H., Andrew, R. L., Bock, D. G., Franklin, M. T., Gilbert, K. J., Kane, N. C., Moore, J., Moyers, B. T., Renaut, S., Rennison, D. J., Veen, T., Yeaman, S. 2013. Mandated data archiving greatly improves access to research data. *The FASEB Journal*. 27, 4 (April 2013), 1304-1308. http://www.fasebj.org/content/27/4/1304

[11] Ferguson, L. 2014. How and why researchers share data (and why they don't). *Wiley Exchanges*. http://exchanges.wiley.com/blog/2014/11/03/how-and-why-researchers-share-data-and-why-they-dont/

[12] Berman, J. J. 2002. Confidentiality issues for medical data miners. *Artificial Intelligence in Medicine*. 26, 25-36.

[13] Pepe, A., Goodman, A., Muench, A., Crosas, M., Erdmann, E. 2014. How do astronomers share data? Reliability and Persistence of Datasets Linked in AAS Publications and a Qualitative Study of Data Practices among US Astronomers. *PLOS One*. DOI= http://dx.doi.org/10.1371/journal.pone.0104798

[14] Data Citation Synthesis Group. 2014. Joint declaration of data citation principles. Martone M. (ed.). *FORCE11*. San Diego, Ca. https://www.force11.org/group/joint-declaration-data-citation-principles-final

[15] Jayasuriya, H. K., Ritcheske, K. 2015. Big Data, Big Challenges in Evidence-Based Policy Making (American Casebook Series). *West Academic Publishing*.

[16] DataCite. Cite your data. https://www.datacite.org/services/cite-your-data.html

[17] Goodman, A. et al. 2014. Ten simple rules for the care and feeding of scientific data. *PLoS Computational Biology*.10, 4, e1003542. DOI= http://dx.doi/org/10.1371/journal.pcbi.1003542

[18] Thomson Reuters. Master Data Repository List. http://wokinfo.com/products_tools/multidisciplinary/dci/repositories/search/

[19] Corti, L., Van den Eynden, V., Bishop, L., Woollard, M. 2014. Managing and Sharing Research Data: A Guide to Good Practice. *Sage Publications*

[20] Kratz, J. and Strasser, C. 2014. Data publication consensus and controversies. *F1000Research*. 3, 94. DOI= http://dx.doi.org/10.12688/f1000research.3979.3

[21] Corporation for National Research Initiatives.Handl.Net Registry. http://www.handle.net/

[22] Borgman, C. L. 2015. Big Data, Little Data, No Data: scholarship in the networked world. *MIT Press*. Cambridge, Mass.

[23] Altman, M., King, G. 2007. A Proposed Standard for the Scholarly Citation of Quantitative Data. *D-Lib*. 13, 3/4 (March/April 2007).

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

[25] Rauber, A., Asmi, A., van Uytvanck, D., Proll, S. 2015. Data Citation of Evolving Data: Recommendations of the Working Group on Data Citation (WGDC). *Research Data Alliance*. https://rd-alliance.org/system/files/documents/RDA-DC-Recommendations_151020.pdf

[26] Open Archives Initiative. Open Archives Initiative Protocol for Metadata Harvesting. https://www.openarchives.org/pmh/