

Data Citation Index: Promoting attribution, use and discovery of research data

Megan M. Force^{a,*} and Daniel M. Auld^b

^a *Thomson Reuters, Philadelphia, PA, USA*

E-mail: megan.force@thomsonreuters.com

^b *Thomson Reuters, Enterprise House, Innovation Way, Heslington, UK*

E-mail: daniel.auld@thomsonreuters.com

Abstract. The Data Citation Index forms part of the Thomson Reuters Web of Science research platform. Thomson Reuters makes partnerships with data repositories to create indexed data object metadata records integrated within the larger platform. The core features of the Data Citation Index strengthen and standardize citation and re-use of data across scientific disciplines.

Keywords: Data citation, data discoverability, data repositories

1. Introduction

The analysis tools and bibliographic information provided by the Web of Science (WoS) help to measure the productivity and impact of research in the wider academic community. The platform indexes scholarly journals, books, conference proceedings and patents, covering the sciences, social sciences, arts and humanities. The Data Citation Index (DCI; http://wokinfo.com/products_tools/multidisciplinary/dci/) has been developed to provide WoS users with a tool to search and discover research data associated with the most influential published research. Indexed data sets are linked with bibliographic literature citations in the Web of Science, creating a more holistic view of the scholarly research process. Another of the product's primary aims is to provide metric information on the re-use of data; this will increase as citations to datasets are identified.

The product was initiated after Thomson Reuters became aware of an increased visibility of data in the literature and in the research community, with many institutions devoting more resources to data. Activities were seen to increase due to data management plan requirements from funders and institutions, as well as ever more prominent government mandates. In addition, various strands of academic discussion became focused on methods to elevate the status of data within scholarly research. These discussions mainly concerned the topics of discoverability (encouraging researchers to deposit their datasets in publicly accessible digital repositories), scholarly attribution and citability (developing robust data citation standards) [2].

*Corresponding author: Megan M. Force, Thomson Reuters, 1500 Spring Garden Street, Philadelphia, PA 19130, USA. Tel.: +1 215 823 6194; E-mail: megan.force@thomsonreuters.com.

2. Selection criteria

Not all data repositories identified by Thomson Reuters are selected for inclusion in the DCI; repositories must meet certain criteria with respect to their scope, maintenance, and usefulness to the scholarly community [3]. These criteria include demonstrable stability of the data objects and the repository which oversees their curation, as well as data publication/curation standards and established links to scholarly research. Age of material is not a restricting factor in the acceptance or rejection of a possible data resource to the DCI; historical data remains a useful if not essential tool in many areas, particularly in studies of long-term phenomena such as climate change.

3. Product records and features

Repositories that agree to participate in the DCI provide metadata of their data holdings for Index record creation. Thomson Reuters works closely in partnership with the data repository to help create metadata which can be used by DCI through provision of a data dictionary and XML schema. This enables repositories to be confident in the representation of their content in the DCI records. In its initial phase, the DCI has employed metadata provided by repositories to build the infrastructure to track data citations. With this infrastructure in place, development will focus on gaining data citations for deposited data.

Article records published in scholarly journals and included in the WoS Core Collection can be linked to DCI data records for the underlying data. Similarly, records in DCI for data repositories, data studies, and data sets may contain links to published scholarly articles in the WoS Core Collection which have cited the data, or to articles which the published data set itself cites. This improves researchers', funders' and other interested stakeholders' understanding of research impact, and allows them to determine whether the experiment or study is reproducible. Each data resource included in the platform features a recommended data citation, based on the DataCite data citation guidelines [1], in order to encourage onward citation by DCI users. The inclusion of metadata enriched with indexing terms permits improved data set discoverability, and provides the ability to discern new connections and relationships between data from various disciplines and authors.

4. Conclusion

Currently, disciplinary data repositories often differ in the metadata elements made available, and the appropriate level of granularity with respect to data citation. Thomson Reuters works closely with representatives of data repositories to coordinate an approach which will accurately represent the data while standardizing indexing fields for dataset citation, making data citable and discoverable for the various stakeholders in the data lifecycle.

References

- [1] DataCite, Why cite data?, available at: <https://www.datacite.org/whycitedata>, accessed 2 June 2014.
- [2] T. Green, We need publishing standards for datasets and data tables, *Learned Publishing* 22(4) (2009), 325–327.
- [3] Repository evaluation, selection, and coverage policies for the Data Citation Index Within Thomson Reuters Web of Science, available at: http://wokinfo.com/products_tools/multidisciplinary/dci/selection_essay/, accessed 29 May 2014.