ANDS and Thomson Reuters collaboration: Data Citation Index June 2014

Introduction

Following the agreement between the Australian National Data Service (ANDS) and Thomson Reuters to work together to streamline the flow of Australian data into the Data Citation Index (DCI) to allow tracking of data citation and reuse, trials have been carried out converting incoming metadata received by ANDS from early adopting data providers and partners into a form suitable for ingestion and tracking.

This document has been prepared by Thomson Reuters to allow ANDS partners and data providers to understand requirements needed for compliance and inclusion in this process.

Metadata

ANDS partners provide metadata to ANDS to create a catalogue record for data objects curated by them or their data providers. Review of these objects confirms that information is given to ANDS relating to data objects held in repositories under the control of the ANDS partner or data provider, or can occasionally represent secondary catalogue records for data held elsewhere in repositories outside ANDS coverage or even outside Australia. For example, data produced by Australian researchers, but deposited in a non Australian repository as part of a wider study.

In tracking the citation of Australian data objects, it is important to distinguish these scenarios as the data citation should, and will most likely, be made to the location on the web where the data can be obtained [1]. Hence when Thomson Reuters discover a data citation, this will be to the actual repository providing data access, not to a secondary catalogue record or web page.

There are also a number of records in samples reviewed which link to institutional web pages, rather than data. These records are not suitable for DCI and if they occur frequently in the metadata of a given partner, may result in the partner omitted from DCI.

In creating records in DCI, Thomson Reuters aim to:

- Provide attribution for the data object to the person(s), department(s) or institution(s) creating the data
- Provide a standard form of data citation for each data object to encourage citation. The format of data citation currently recommended by Thomson Reuters follows the DataCite guidelines. Elements needed to create the data citation must be present in the metadata.
- Track citations and reuse of the data in the scientific literature, and provide bidirectional links between the research articles and the data they use or generate.
- Provide a means to discover data associated with the most influential research publications

In order to do this, certain metadata are needed to create a data citation which can be matched to a data citation in the literature and which provide access to the actual data in the repositories to allow reuse and citation as part of the data lifecycle. The threshold of metadata needed to do this is relatively low:

- Author/creator: the person(s), department(s) or institution(s) who created the data and should receive attribution; data objects without attribution of this type can be assigned as having 'Anonymous' authorship which may result in lack of attribution to the actual authors of the data. A number of author formats have been observed in incoming data, and

Thomson Reuters is working with ANDS to normalise these data. Fully parsed and fielded data for each author is preferred rather than including all authors in a single field

- Data object title
- Source: the repository or data provider enabling access to the data themselves. The source listed in ANDS metadata should be the data provider or repository where the data are actually held and which should be included in a data citation, thereby enabling Thomson Reuters to track the reuse of the data through citation in the scientific research literature; the clickable link provided via a persistent identifier or URL should provide access to the full data deposited, not to a secondary catalogue record or institutional web page with no data access. Where the source of a catalogue record is included as the source of the actual data, these records are not suitable for DCI.
- Source location: URL, or preferably DOI (or some other persistent identifier) which can be tracked and used to link the user to the source of the data in the repository which offers access
- Publication Year: the year the data were published – made available for reuse

Note that at all points in this process, the actual research data reside with the repository which allows the repository and data owner to govern the use and licensing of the data themselves. Thomson Reuters only receive and use the metadata description of the data to create a record which provides access to the actual data, via the data provider/repository, for reuse and allows provision of a standard data citation and a point to attach metrics for citation and reuse.

Selection & benefits

ANDS partners who are able to provide these metadata and fulfil the DCI selection criteria are eligible for inclusion in DCI and citations to the data objects can be tracked. In return, if selected for inclusion, the data provider will have free access to DCI to enable them to review the implementation of their data. Throughout the process, Thomson Reuters works closely with the data provider to ensure correct representation of repository information as all material deposited in a given repository is linked to that repository record, thereby raising the visibility of the repository within the Web of Science and positioning data as a first class research object, alongside the scientific research literature to which they relate.

References

[1] Ball, A. & Duke, M. (2012). 'How to Cite Datasets and Link to Publications'. *DCC How-to Guides*. Edinburgh: Digital Curation Centre. Available online: http://www.dcc.ac.uk/resources/how-guides