

Data citation

Who should read this?

This guide is intended for eResearch infrastructure support providers and researchers. It is not so much a guide to how to cite data, but a guide to the issues around it, and activities underway to change the culture around data citation in order to support improved data management and sharing.

What do we mean by data citation?

Data citation refers to the practice of providing a reference to data in the same way as researchers routinely provide a bibliographic reference to printed resources. The need to cite data is starting to be recognised as one of the key practices underpinning the recognition of data as a primary research output rather than as a by-product of research. While data has often been shared in the past, it is rarely, if ever, cited in the same way as a journal article or other publication might be. If data sets were cited, they would achieve a validity and significance within the cycle of activities associated with scholarly communications and recognition of scholarly effort.

How do you cite data?

At present, there is no generally recognised way of citing data, despite the growing need.

ISO690 2 dates from 1997. This 'specifies the elements to be included in bibliographic references to electronic documents. It sets out a prescribed order for the elements of the reference and establishes conventions for the transcription and presentation of information derived from the source electronic document. [It] is intended for use by authors and editors in the compilation of references to electronic documents for inclusion in a bibliography, and in the formulation of citations within the text corresponding to the entries in that bibliography. It does not apply to full bibliographic descriptions as required by librarians, descriptive and analytic bibliographers, indexers, etc.'¹

A recent *OECD Publishing White Paper*² by Toby Green sets out the need for a recognised standard and proposes a model which will be used by the OECD for its own data and data tables.

Altman and King³ proposed a standard for citing quantitative data in 2007. This contains many of the elements common to print citations, to which are added components specific to quantitative data sets. Similar to the recommendations of the OECD White Paper and the citation supplied by ICPSR, their standard includes a permanent identifier (whether DOI or other) as an essential element. Their minimum citation includes only six elements, including the permanent identifier.

Various data repositories provide a recommended format for citing data from that repository. For example: ICPSR and other social science data centres provide a citation for each of their datasets as follows:

Kessler, Ronald C. National Comorbidity Survey: Baseline (NCS-1), 1990-1992 (Restricted Version) [Computer file]. ICPSR25381-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2009-05-11. doi:10.3886/ICPSR25381

The following citation comes from PANGAEA, the Publishing Network for Geoscientific & Environmental Data in Germany, and covers both the publication and the data on which it was based.

Nishioka, J et al. (2008): Profiles of iron concentration from GoFlow bottles during the CARUSO-EISENEX experiment, doi:10.1594/PANGAEA.701305, *Supplement to*: Nishioka, Jun; Takeda, Shigenobu; de Baar, Hein JW; Croot, Peter L; Boyé, Marie; Laan, Patrick; Timmermans, Klaas R (2005): Changes in the concentration of iron in different size fractions during an iron enrichment experiment in the open Southern Ocean, *Marine Chemistry*, 95(1-2), 51-63, doi:10.1016/j.marchem.2004.06.040



The issue is further complicated by the fact that bibliographic management systems such as EndNote and Zotero do not provide a template for a data citation.

Wouldn't it be lovely if ...

- The creation of data could be recognised as a primary research output,
- The use and re-use of data were accompanied by a full data citation, including a persistent identifier,
- Data use and re-use could be tracked and recorded in the same way as print publications,
- Data citation information were used for research evaluation and reward.

The ANDS approach to data citation

An important aim of ANDS is to enable more researchers to re-use research data more often. To achieve this aim, ANDS is engaged in activities that will make it easier to share data, to recognise the importance of making data available and to make data citation a standard procedure.

- Publish My Data is a service which will enable Australian researchers and research organisations to store and publicise the existence of research collections via the internet. It will allow AAF authenticated participants to register their collection description information and obtain a persistent identifier for the collection. This information will be stored in the ANDS Collections Registry and will be discoverable through Research Data Australia.
- ANDS is joining DataCite, a group of leading research libraries and technical information providers that aims to make it easier for research datasets to be handled as independent, citable, unique scientific objects. This will be done by using Digital Object Identifiers (DOI) as permanent identifiers for datasets. ANDS is participating in the DataCite metadata standards working group. See <http://www.datacite.org/>.
- ANDS is working with both ThomsonReuters and Elsevier to investigate the feasibility of tracking and recording of data set use through DOIs, and to make that information available through *Web of Science* and *Scopus*. Both of these databases are used extensively world-wide as part of research assessment activities.
- ANDS is engaging with research funding agencies to promote data publication as a primary research output and the inclusion of data in the research assessment process.

Directions around data publication

- 'What is more, funding agencies and researchers alike must ensure that they support not only the hardware needed to store the data, but also the software that will help investigators to do this. One important facet is metadata management software: tools that streamline the tedious process of annotating data with a description of what the bits mean, which instrument collected them, which algorithms have been used to process them and so on — information that is essential if other scientists are to reuse the data effectively. Also necessary, especially in an era when data can be mixed and combined in unanticipated ways, is software that can keep track of which pieces of data came from whom. Such systems are essential if tenure and promotion committees are ever to give credit — as they should — to candidates' track-record of data contribution.' Nature editorial. 2009. 'Data's shameful neglect.' *Nature* 461 (145): 168-170. <http://www.nature.com/nature/journal/v461/n7261/full/461145a.html>
- Data only journals are now starting to appear. For example, see *Earth System Science Data*. <http://www.earth-system-science-data.net/>

¹ <http://paedpsych.jk.uni-linz.ac.at/internet/ARBEITSBLAETTERORD/LITERATURORD/ZitationISO690.html>

² Green, T. (2009). 'We Need Publishing Standards for Datasets and Data Tables.' *OECD Publishing White Paper*. <http://dx.doi.org/10.1787/603233448430>

³ Micah Altman and Gary King. 2007. 'A Proposed Standard for the Scholarly Citation of Quantitative Data.' *D-Lib Magazine*, Vol. 13, No. 3/4 (March/April), <http://www.dlib.org/dlib/march07/altman/03altman.html>



This work is licensed under a [Creative Commons Attribution-Noncommercial-Share Alike 2.5 Australia License](http://creativecommons.org/licenses/by-nc-sa/2.5/au/)