

Data citation

Who should read this?

This guide is intended for eResearch infrastructure support providers and researchers. It includes some suggestions about citing data as well as discussing issues around data citation and activities underway to develop a culture of data citation.

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 seldom cited in the same way as a journal article or other publication might be. This culture is, however, gradually changing. If datasets were cited, they could achieve a validity and significance within the scholarly communications cycle. Citation of data could enable recognition of scholarly effort in disciplines and organisations that want to acknowledge and reward data outputs.

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 other research publications, and
- Data citation information was incorporated into practices 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.

- ANDS has joined DataCite (www.datacite.org), 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 is 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 will offer, from mid-2011, a DOI minting service, Cite My Data, to provide datasets with a unique and traceable identifier. This can then be used by researchers when they cite their own data in publications. Data can then be registered through ANDS online services, Register My Data (<http://ands.org.au/guides/register-my-data-awareness.html>) or Publish My Data (<http://ands.org.au/guides/publish-my-data.html>) and be discoverable through Research Data Australia (<http://services.ands.org.au>). The DOI assigned to dataset can be used by other researchers when the data is re-used and cited. Initially this will be a machine to machine service only.
- ANDS is working with both ThomsonReuters (<http://thomsonreuters.com/>) and Elsevier (<http://www.elsevier.com>) to investigate the feasibility of tracking and recording of data use through DOIs, and making that information available through *Web of Science* (http://thomsonreuters.com/products_services/science/science_products/a-z/web_of_science/) and *Scopus* (<http://www.scopus.com/>). 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.



How do you cite data?

Data citation standards vary across disciplines. However, DataCite has undertaken some work in this area and recommends the following format.

- Creator (Publication Year): Title. Publisher. Identifier

They recognise that it may also be desirable to include two optional properties, Version and ResourceType (as appropriate). If so, the recommended form is as follows:

- Creator (PublicationYear): Title. Version. Publisher. ResourceType. Identifier

For citation purposes, the Identifier may optionally appear both in its original format and in a linkable, http format. Here are some example data citations:

- Irino, T; Tada, R (2009): Chemical and mineral compositions of sediments from ODP Site 127-797. Geological Institute, University of Tokyo. doi:10.1594/PANGAEA.726855.
<http://dx.doi.org/10.1594/PANGAEA.726855>
- Geofon operator (2009): GEFON event gfz2009kciu (NW Balkan Region). GeoForschungsZentrum Potsdam (GFZ). doi:10.1594/GFG.GEOFON.gfz2009kciu.
<http://dx.doi.org/10.1594/GFG.GEOFON.gfz2009kciu>

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 connection between data and publication is increasingly recognised. The following citation comes from PANGAEA, the Publishing Network for Geoscientific & Environmental Data in Germany. This applies to a data set, and the subsequent citation is to the article based on analysis of the data.

- Kuhlmann, H et al. (2009): Age models, iron intensity, magnetic susceptibility records and dry bulk density of sediment cores from around the Canary Islands. doi:10.1594/PANGAEA.727522.
<http://dx.doi.org/10.1594/PANGAEA.727522> Supplement to:
Kuhlmann, Holger; Freudenthal, Tim; Helmke, Peer; Meggers, Helge (2004): Reconstruction of paleoceanography off NW Africa during the last 40,000 years: influence of local and regional factors on sediment accumulation. *Marine Geology*, 207(1-4), 209-224. doi:10.1016/j.margeo.2004.03.017
<http://dx.doi.org/10.1016/j.margeo.2004.03.017>

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>

