AWARENESS LEVEL

A Digital Curation Centre Briefing Paper

19th July 2011

Data Citation and Linking

By Alex Ball and Monica Duke, UKOLN, University of Bath

Introduction

On the surface, citing datasets is a trivially easy thing to do. Style manuals such as the *Publication Manual of the American Psychological Association* and the *Oxford Manual of Style* have provided sample citations for datasets since at least the early 2000s. The process of making datasets citable, however, is rather more difficult. In consequence of this and other factors, a culture of citing datasets has been slow to develop. Nevertheless, it is vital that researchers cite the datasets they use, if datasets are to be regarded as legitimate academic outputs in their own right.

Short-term Benefits and Long-term Value

There are several short-term benefits to making datasets citable, citing them in practice, and linking datasets to papers that make use of the data.

- If the authors of a scientific publication properly cite the data that underlies it, it is much easier for the reader to locate that data. This in turn makes it easier for the reader to validate and build on the publication's findings.
- Data citations ensure that data contributors receive proper credit when their work is reused by other researchers.
- If a dataset links back to the paper that describes its collection, a reader coming to the dataset direct can use that link to put it in context and understand the methodology used.
- If a dataset links to other papers that make use of it, these links can be used by the contributors and data publishers to demonstrate the impact of the data. Potential reusers might use these links to discover critiques of the data or to provide inspiration for how to use them.

Once a culture of data citation has been established, several other benefits are likely to become apparent.

- The publishing infrastructure that makes the data citable will also help to ensure they are available for reference and reuse long into the future.
- There will be less danger of rival researchers 'stealing' results from those who publish their data openly, as failure to give due credit would amount to plagiarism and thus be punishable.
- Services built around data citation will make it easier for researchers to discover relevant datasets.
- Data citations could be used to measure the impact of both individual datasets and their contributors.
- Researchers could gain professional recognition and rewards for published data in the same way as for more traditional publications.

Taking these points together, there would likely be an increase in the quantity and quality of data published, with all the benefits this implies for the transparency and rate of scientific research.

Perspectives on Data Citation

"Adequate citation of data sets is crucial to the encouragement of data sharing, to the integrity and cost-effectiveness of science and to easy access to the work of others."
— Sieber & Trumbo (1995)

"Without an effective data citation mechanism the implementation of the 'Data Publishing Framework' would remain incomplete. Thus, universal standards for citing datasets are essential. [...] currently we lack consistency in data citations, which is sure to provide much needed high visibility to data."
— Chavan & Ingwersen (2009)

Roles and Responsibilities

Maintaining data in a citable state is primarily the responsibility of a data publisher or distributor: typically an institutional data repository or disciplinary data archive. Data publishers have the following responsibilities.

- They should ensure that the data they publish, along with any explanatory metadata, remain static over time, so that readers can look up precisely the resources used by the author. This implies a formal digital preservation and versioning strategy.
- They should assign identifiers to the data to make them easier to find. These identifiers should be persistent and unique,[1] and remain associated with the correct version of the data. Ideally it should be possible to locate data by passing the identifier to a resolver service, as with Digital Object Identifiers (DOIs) and Handles.
- They should ensure that published data remain accessible, even if that access is mediated or restricted in the case of sensitive, commercial or obsolete data.[2]
- They should provide depositing authors with a citation they can include in their associated paper. Once notified of that paper's publication, they should add a link to it from the dataset catalogue record.

Journal publishers also have a role to play in enforcing data citation standards.

- They can suggest or mandate specific repositories in which underlying data should be deposited. Any mandates should be sensitive to the wishes of funding agencies.
- They should provide clear guidance on how and where datasets should be cited in their papers.
- When publishing a paper that cites a dataset, they could alert the data publisher to that fact.

It remains the responsibility of authors to

- use common standards when generating, recording, coding, packaging, and so on, the data underlying their research;
- work up these data to publication standard and deposit them with an appropriate data publisher;
- obtain a citation from the data publisher and include it in the associated paper, using the format required by the editor or otherwise one of the standard formats;
- also include citations for any prior datasets used in the course of research; and
- notify the data publisher about the associated paper, whether at the data deposition stage, the paper publication stage or both.

Issues to be Considered

- At what granularity should data be made citable? If single datasets are given identifiers, what about collections of datasets, files within datasets or individual data?
- It would harm the integrity of publications if the data they cited changed over time. To avoid this, different citations/identifiers should be used for different versions of the same dataset. This is true for versions formed by adding data, modifying data, or even migrating it to a new format. For datasets that are frequently updated, should a new identifier be assigned after each update, or should these updates be 'saved up' and published at regular intervals, or on demand? Should time-series data be published as complete snapshots or in instalments? Must all published versions of datasets be stored, or can previously published versions be generated reliably when needed?

- It is accepted practice that a data citation should point to a catalogue page (or other form of landing page) for data, rather than to the data directly. Such pages typically provide descriptive metadata, a sample citation, a link to an accompanying paper, instructions on how to access the data, and the licence under which the data are released. Is it possible to provide this page in such a way that both human readers and automated scripts are able to navigate to the data? What other sorts of information should this page provide?

- Where within a research paper should a data citation be given? While the most logical place is within the bibliography or references section, there are arguments for placing it elsewhere.[3] For example, a journal publisher might limit the total number of citations allowed, or one might want to distinguish between data reported by the paper and data reused by the paper.

- One of the ways citations are used is to monitor the impact of authors in their field. Measuring this accurately is next to impossible, though, if one has to rely on names alone to identify authors. Author identifiers such as ORCID[4] or ISNI[5] would make the task vastly simpler, but how should they be included in citations?

Related Research

- **ACRID*** (Advanced Climate Research Infrastructure for Data) developed a linked-data approach to citing and publishing climate research data along with full provenance information, including the workflows and what software was used.

- **Data Linking with Knowledge-Blogging*** extended existing blogging tools to create a lightweight, semantically linked publication environment. The environment supports peer review and bidirectional links between data and narrative publications.

- **DataCite** is an international initiative that provides the infrastructure for assigning Digital Object Identifiers (DOIs) to datasets.

- **Dryad UK*** established a UK mirror of the US-based Dryad data repository, extended its support to new publishers and disciplines, and developed a sustainability plan and performance metrics.

- **FISH.Link*** produced tools for converting and mapping freshwater biology data to linked data, while supporting semantic markup, attribution and provenance.

- **SageCite*** produced a demonstrator citation service for network models, workflows and associated data in the Sage Commons, using a linked data approach.

- **SPQR*** (Supporting Productive Queries for Research) trialled the use of linked data to express and integrate datasets related to classical antiquity, as a way of overcoming the challenges raised by the interpretive and uncertain nature of the material.

- **Webtracks*** extended previous work by the **CLADDIER** and **StoreLink** projects in order to produce a secure method for communicating semantic links between data repositories, publication repositories, open science notebooks and publishers.

- The **XYZ Project*** developed tools and an exemplar workflow for co-ordinating the deposition of data in archive with the review and publication of an associated paper.

** indicates a project from the JISC Managing Research Data Programme.*

Additional Resources

Altman, M. & King, G. (2007). A proposed standard for the scholarly citation of quantitative data. *D-Lib Magazine*, 13(3/4). doi:10.1045/march2007-altman

Chavan, V. & Ingwersen, P. (2009). Towards a data publishing framework for primary biodiversity data: Challenges and potentials for the biodiversity informatics community. *BMC Bioinformatics*, 10(Suppl 14), S2. doi:10.1186/1471-2105-10-S14-S2

Data citation. (2011, May 3). [Awareness Level Guide]. Retrieved 6 June 2011, from the Australian National Data Service: http://www.ands.org.au/guides/data-citation-awareness.html

Data Seal of Approval: Quality guidelines for digital research data.(2010, May 3). Version 2.0. Data Seal of Approval Board. Retrieved 29 June 2011, from http://assessment.datasealofapproval.org/documentation/

Davidson, J. (2006, October 17). Introduction to curation: Persistent identifiers. Retrieved 8 June 2011, from the Digital Curation Centre: http://www.dcc.ac.uk/resources/briefing-papers/introduction-curation/persistent-identifiers

Green, T. (2010, February). *We need publishing standards for datasets and data tables.* OECD Publishing. doi:10.1787/787355886123

ISO/DIS 27729. (n.d.). *Information and documentation – International standard name identifier* (ISNI). Draft International Standard. International Organization for Standardization.

Lawrence, B. N., Jones, C. M., Matthews, B. M. & Pepler, S. J. (2008, February 1). *Data publication* (Claddier Project Report No. 3). BADC. Retrieved 11 May 2011, from http://purl.org/oai/oai:epubs.cclrc.ac.uk:work/43641

Piwowar, H. (2011, May 5). Links from the data collection article: Inline or in the bibliography? [blog post]. Retrieved 3 June 2011, from the Research Remix blog: http://researchremix.wordpress.com/2011/05/05/inline-or-biblio/

Sieber, J. & Trumbo, B. (1995). (Not) giving credit where credit is due: Citation of data sets. *Science and Engineering Ethics*, 1(1), 11–20. doi:10.1007/BF02628694

Starr, J. & Gastl, A. (2011). isCitedBy: A metadata scheme for DataCite. *D-Lib Magazine*, 17(1/2). doi:10.1045/january2011-starr

Please cite this briefing paper as:

Ball, A., Duke, M. (2011). 'Data Citation and Linking'. *DCC Briefing Papers*. Edinburgh: Digital Curation Centre. Available online: http://www.dcc.ac.uk/resources/briefing-papers/

Acknowledgements

The authors would like to thank David Shotton, Gudmundur Thorisson and Angus Whyte for their helpful comments.

Notes

1. Davidson (2006).
2. Data Seal of Approval (2010).
3. Piwowar (2011).
4. See the Open Researcher and Contributor Identifier (ORCID) Initiative Website.
5. ISO/DIS 27729 (n.d.).