1-23-2011

Data Citation: Supporting scholarly activity, encouraging data sharing

Suzie Allard
University of Tennessee - Knoxville, sallard@utk.edu

Data Citation: Supporting scholarly activity, encouraging data sharing

Providing guidelines for consistent data citation has four primary outcomes:

1. It enables the reader to find the data used in the research and to repeat the researcher's analysis. This is an important component of the scientific method.
2. It enhances the visibility of the data producer's research, thus providing a means for data producers to garner scholarly recognition for their research work in a manner similar to the recognition received for traditional scholarly article.
3. It increases the visibility of the data repository, which may increase use of the data holdings.
4. It leads to increased willingness to share data, which supports scientific discovery and the scientific enterprise.

What is data citation?

Just as we cite papers, so too should we cite data sets. Data citation provides a bibliographic means to reference a data set so it may be properly attributed in a scholarly work.

Why we need data citation

Data citation provides the reader of a paper with a path for finding the data used in the research and for repeating the research analysis. The scientific method is founded on inquiry that is based in collecting observable and measurable evidence. The complete scientific process includes identifying a question, doing background research, constructing a hypothesis, testing the hypothesis with data, and analyzing and reporting the results. The steps must be repeatable and should allow for predicting future results. Full disclosure is also vital to the scientific method, since it facilitates the sharing of the data and methodologies used allowing other scientists to verify the research results . Data citation plays an important role in this verification process, as citation provides scientists with a link to the data used in the testing and analysis steps, which allows scientists to repeat the analysis, and verify the results. Data citation fosters "openness, fairness and economy in the pursuit of scientific knowledge" (Sieber & Trumbo, 1995). Seeber (2008) also notes that it is best for the citation to be in the formal references list, rather than in a table or supplementary information. This practice simplifies the process of tracking citations of a given source.

Data citation allows data producers to receive attribution for their work. Just as article citation allows authors to receive attribution for their scholarly articles, data citation allows data producers to receive acknowledgement for the scholarly contribution of careful collection and curation of data sets that similarly benefit the scientific community. The lack of data citation best practices is a hindrance to the scientific community, since it limits the ability for credit attribution (Altman & King, 2007).

The need for data citation is becoming widely recognized in the scientific community. For example, the American Geophysical Union (AGU) (2009) states that "The scientific community should recognize the professional value of such [data production] activities by endorsing the concept of publication of data, to be credited and cited like the products of any other scientific activity, and encouraging peer-review of such publications."

Many other groups have similarly affirmed the need for data citation. Such statements include, but are not limited to, the Distributed Active Archive Centers (DAAC), International Polar Year (IPY), Global Biodiversity Information Facility (GBIF), and the U.S. Long Term Ecological Research Network (LTeR). (A more complete list of organizations and their statements is included at the end of this document.) Even publishers are becoming active partners in encouraging a data archiving policy that includes preserving data sets and linking them to articles. Partners who have signed the Joint Data Archiving Policy created by the Dryad project include publishers of journals such as *The American Naturalist, Evolution, Heredity* and *Journal of Evolutionary Biology.*

Data citation has an important role in the scientific scholarly process, as it provides an incentive to publish, share and preserve data sets. Since data citation supports attribution in scholarly literature and allows for replication of scientific analyses, it increases the incentive to publish a data set. A published data set facilitates sharing of the data. It also supports curation and preservation activities. The outcome is that data is more likely to be accessed and re-used, a condition which leads to increased scholarly productivity from an initial data product.

Data citation has an important role in intellectual property issues. Data citation gives credit to data producers and data publishers. The practice also provides a link from the traditional literature to the data, further enhancing the intellectual property link. Data citation furthermore gives intellectual legitimacy to the creation of data, since through the citation process data is treated it as an important scientific product worthy of being stored and cited.

The current state of data citation

As noted above, the need for data citation is recognized by scientists, scientific societies and scientific publishers. Many learned societies (e.g. , research organizations (,i.e., DAAC) and standards organizations .i.e. NISO) are discussing the issue of data citation, but only a few (i.e. AGU) have put forth recommendations. . In addition, only a few journals and funders specify data citation standards, and only YYY of XXX repositories surveyed offer a data citation recommendation (Enriquez et al., 2010)

The DataCite (established 2009) initiative encourages data publishing via global data citation support. The initiative includes the adoption of standards and persistent reference to data sets in regional archives. There are also a range of standards being developed in a number of communities, including ORNL DAAC, Pangaea, GCMD, ESIP, GBIF, TDWG, OECD, NISO/NFAIS, IPYDIS, and Dataverse.

At this time there is no standard for data citation (Parsons & Duerr, 2010). This has led to a great diversity in citation practices (Seeber, 2008 Enriquez et al., 2010), a reality which makes tracking data very difficult (Enriquez et al., 2010). Tracking data reuse can lead to a better understanding of the data lifecycle, as well as of what topics can be well studied by reusing data (Chavan, 2010).

The lack of a standard has resulted in data citations that are highly variable. For example, a citation may not include an identifier, and it may be unclear if the citation is for the full data set or a data subset. There is also ambiguity regarding whether the dates represent the date of collection or the date of publication.

There is growing recognition that we need data citation standards (Brase et al., 2009; Green, 2009; Kelly, 2008). Data set citations provide:

1. precise identification of a data set, including at the level of version, file, and table.
2. a means for the reader to find and understand the data.
3. credit to the data producers and data publishers.
4. a means to gain credit for data sharing and archiving, which provides an incentive for publishing the data.
5. a link from the traditional literature to the data.
6. the means to give intellectual legitimacy to the creation of data sets.
7. the means to collect research metrics for data sets.

An ideal data citation format will:

1. describe any data set, database, or data file.
2. provide for all levels of granularity (table, row, cell).
3. be used for any snapshot (version, e.g., in time)
4. be formatted for any view: XML, HTML, CSV, etc.
5. have (or not have) annotations
6. link to older, newer, and latest versions
7. provide actionability (“Click-through”)
8. provide persistence (validity into the future)
9. be machine readable, thus allowing for automatic parsing

References & items for further reading

Altman, M., & King, G. (2007). A proposed standard for the scholarly citation of quantitative data, *13*(3/4). Retrieved from http://gking.harvard.edu/files/abs/cite-abs.shtml

Brase, et al. (2009). Approach for a joint global registration agency for research data. *Information Services & Use, 29*(1), 13-27. (i.e, DataCite)

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

Cook, R. (2008) Citations to published data sets. *FLUXNET Newsletter*. Retrieved fromhttp://daac.ornl.gov/ornl_daac_citations_200812.pdf

Dryad. (2010, November 16). *Dryad: About us*. Retrieved from http://datadryad.org/about.

Dryad. (2010, July 7). *Dryad: Using data in Dryad*. Retrieved from http://datadryad.org/using.

Enriquez V., Judson S.W., Weber N.M., Allard, S., Cook, R.B., Piwowar, H.A., Sandusky, R.J., Vision, T.J., & Wilson, B. (2010). Data citation in the wild. Chicago, IL: IDCC. Retrieved fromhttp://dataonedatacitations.wordpress.com/2010/09/13/dcc-poster-submission-data-citation-in-the-wild/

Green, T. (2009). *We need publishing standards for datasets and data tables*. OECD Publishing White Paper, OECD Publishing.

:Kelly, M.C. (2008). *NISO thought leader meeting on research data*. Retrieved from http://www.niso.org/topics/tl/NISOTLDataReportDraft.pdf.

ORNL DAAC. (2010, June 16). *About ORNL DAAC*. Retrieved from http://daac.ornl.gov/about_us.shtml

ORNL DAAC. (2010, May 27). *Data product citation policy*. Retrieved from http://daac.ornl.gov/citation_policy.html

Page, R.D.M. (2008). Biodiversity informatics: The challenge of linking data and the role of shared identifiers. *Briefings in Bioinformatics*,*9*(5), 345-54. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/18445641

Parsons, M.A., Duerr, R., Minster, J.-B. (2010). Data citation and peer review. *Eos, Transactions American Geophysical Union, 91*(34), 297-298. Retrieved fro http://public.deltares.nl/download/attachments/16876020/DataCitation_EOS_2010EO340001.pdf

Seeber, F. (2008). Citations in supplementary information are invisible. *Nature, 451*(7181), 887. Retrieved from http://www.nature.com/nature/journal/v451/n7181/full/451887d.html

Sieber, J. E., & Trumbo, B. E. (1995). (Not) giving credit where credit is due: Citation of data sets. *Science and Engineering Ethics, 1*(1), 11-20. Retrieved from http://www.springerlink.com/index/10.1007/BF02628694

Vision, T.J. (2010). Open data and the social contract of scientific publishing. *American Institute of Biological Sciences, 60*(5), 330-331.. Retrieved from http://caliber.ucpress.net/doi/abs/10.1525/bio.2010.60.5.2

EXEMPLARS

At number of scientific research programs have implemented policies for data citation. Listed below are two examples of such projects and their citation standards.

ORNL DAAC

ORNL DAAC (Oak Ridge National Laboratory Distributed Active Center) for Biogeochemical Dynamics, operated by the ORNL Environmental Sciences Division, oversees “data archival, product development and distribution, and user support for biogeochemical and ecological data and models” It is among a number of the NASA Earth Observing System Data and Information System (EOSDIS) data centers managed by the Earth Science Data and Information System (ESDIS) Project, which provides access to data from NASA (About ORNL DAAC, available at http://daac.ornl.gov/about_us.shtml).

ORNL DAAC has created a Data Product Citation Policy intended to accredit publications to their respective authors and to facilitate access to these works. This policy indicates that citations should include as much of the following information as possible: contributing investigators/authors, publication year, product title, medium (for non-print items), online location/URL, publisher, publisher’s location, date accessed, and the digital object identifier (DOI) (About ORNL DAAC). ORNL DAAC also asks that authors who include ORNL DAAC’s data products or services in their publication email electronic reprints of their publications to the organization. This practice enables ORNL DAAC to better understand how its products and services are used and to keep current its product-related references.

Below are ORNL DAAC citation examples (available at http://daac.ornl.gov/citation_policy.html)

- On-Line Data Set

  Turner, D.P., W.D.Ritts, and M. Gregory. 2006. BigFoot NPP Surfaces for North and South American Sites, 2002-2004. Data set. Available on-line [http://daac.ornl.gov] from Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, U.S.A.doi:10.3334/ORNLDAAC/750.

- MODIS Subset

  Oak Ridge National Laboratory Distributed Active Archive Center (ORNL DAAC). 2009. MODIS subsetted land products, Collection 5. Available on-line [http://daac.ornl.gov/MODIS/modis.html] from ORNL DAAC, Oak Ridge, Tennessee, U.S.A. Accessed November 20, 2009.

Online Map

Oak Ridge National Laboratory Distributed Active Archive Center (ORNL DAAC). 2009. FLUXNET Network Map. Available online [http://www.fluxnet.ornl.gov/fluxnet/Maps/Political_fluxnet_networks_cropped_small_april 2009.png] from ORNL DAAC, Oak Ridge, Tennessee, U.S.A.

CDROM Chapman, B., A. Rosenqvist, and A. Wong. 2001. JERS-1 SAR Global Rain Forest Mapping Project. Vol. AM-1, South America, 1995-1996. CD-ROM. National Space Development Agency of Japan, Earth Observation Research Center; National Aeronautics and Space Administration, Jet Propulsion Laboratory; European Commission Joint Research Centre; Earth Remote Sensing Data Analysis Center of Japan; Remote Sensing Technology Center of Japan; and Alaska SAR Facility. Available from Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, U.S.A.

Dryad

Dryad is an international repository for data from peer-reviewed articles in the basic and applied biosciences. According to the project's website, it "enables scientists to validate published findings, explore new analysis methodologies, repurpose data for research purposes unanticipated by the original authors, and perform synthetic studies" (Dryad: Using Data in Dryad).Dryad is a project of the National Evolutionary Synthesis Center and the University of North Carolina Metadata Research Center, which coordinate with numerous journal and societies in evolutionary biology and ecology (Dryad: About Us; available at http://datadryad.org/about).

Dryad asks that, when using data taken from its repository, authors cite the original article, as well as the Dryad data package. This package should include: author(s), publication date, data file name when applicable, data title package ("Data from: [Article name]", the name "Dryad Digital Repository," and the data identifier. The project also suggests that, when using a large number of data sources, the author may wish to provide a list of referenced data packages, rather than citing sources individually (Dryad: Using Data in Dryad; available at http://datadryad.org/using). Dryad provides the following citation example:

> Sidlauskas, B. 2007. Data from: Testing for unequal rates of morphological diversification in the absence of a detailed phylogeny: a case study from characiform fishes. Dryad Digital Repository. doi:10.5061/dryad.20

(Dryad: Using Data in Dryad)

Data citation related comments from other organizations

1. Polar Information Commons (PIC): Appropriate behavior when contributing and using PIC data (http://www.polarcommons.org/ethics-and-norms-of-data-sharing.php)

   The project asks that users agree to a number of conditions. Among these some are related to citation:

   - formal scientific publication citation be used and that PIC users acknowledge authorship and co-authorship of PIC materials.
   - PIC users "give appropriate recognition" to PIC as a digital community resource

- PIC users make efforts to notify PIC contributors about use of specific digital materials and about suspected errors or other problems found while using PIC materials
-

2. International Polar Year (2008, April 1). International Polar Year 2007-2008 data policy. Draft Version 1.2, Retrieved from http://classic.ipy.org/Subcommittees/final_ipy_data_policy.pdf

   "The International Polar Year (IPY) 2007-2008, an intense, interdisciplinary, and internationally coordinated campaign of research and observations, will deepen understanding of polar processes and their global linkages." (p.1)

   "To recognize the valuable role of data providers (and scientists who collect or prepare data) and to facilitate repeatability of IPY experiments in keeping with the scientific method, users of IPY data must formally acknowledge data authors (contributors) and sources. Where possible, this acknowledgment should take the form of a formal citation, such as when citing a book or journal article. Journals should require the formal citation of data used in articles they publish. Where formal citation is not possible, such as with some medical and social science data, ethical policies for data collection and data use are encouraged, building upon existing models such as Article 8(j) of the 1992 Convention on Biological Diversity." (p.4)

3. Parson, M.A. (2006, December 18). Geographic data sharing: Continuing political and cultural barriers amid a trend toward increased openness. Retrieved from http://wiki.esipfed.org/images/b/b4/Parsons_datasharing.pdf

   "The FGDC [Federal Geographic Data Committee] Content Standard for Metadata is the most broadly adopted metadata standard in geography and includes specific fields for a data citation, but they are often not used. More importantly, journals (and reviewers) do not require authors to cite their use of data and in some cases may actively discourage formal data citation. Part of the resistance of journals to accepting data citations is rooted in a similar resistance to allowing formal citation of web sites in that they are seen as mutable and impermanent. The American Geophysical Union has addressed this by requiring that data cited in their journals by archived in a formal data center (htttp://www.agu.org/pubs/data_policy.html). A farther reaching example is that the German National Library of Science and technology includes data "publications" in their catalog, provided the data citation includes a Digital Object Identifier (DOI) as a way to ensure the data can be permanently found and referenced (Klump et al. 2006). This is an important development because these data references can be identified and tracked in formal citation databases allowing for better measurement of citation impact." (p.12)