8-13-2009

Data Citation

UA/SC WG

DATA CITATION

Tenopir reported for Cook, who is the lead on preparing a citation white paper that began at the Knoxville DataONE meeting with a breakout group of Bill Michener, Todd Vision, Suzie Allard, Mary Beth Manoff, Carol Tenopir, and Bob Cook. Trisha Cruse is also on the white paper team.

INTRODUCTION

Publication of scientific findings in journals has traditionally been the way that the accomplishments of investigators and the benefits of sponsored research have been measured. Similarly, publication of data sets in a long-term archive demonstrates the products of research. A data set archived at a data center should have a formal data citation.

"Data Citation" refers to the information and format that is used to describe a dataset. It is used by the author of a paper, derived data set, or other publication. It is informed by metadata attached to the cited data set, but is not exactly the same as metadata.

The data citation practice targets several audiences. Of primary importance are the data providers (authors of the data set). Data providers should receive credit for generating and publishing a data set, much as they currently get credit for publishing a paper. Citation Indices (e.g., ISI or Web of Science) could be used to show how often the data set has been cited. Citations to these published data sets should enable others (e.g., student or a researcher) to obtain the actual data files themselves. Making the data available for others is good scholarly practice—it provides the information necessary to support published results. Furthermore, a citation to a data product in a data archive allows researchers to find, access, and use published data to conduct further analyses (reuse in ways not intended by the data providers).

The citation would also benefit the publisher (data center that houses the data products). The data center can demonstrate to its sponsors that data products at the center have been reused (cited) in published studies. These citations are an indication that the data center is fulfilling its mission by providing data for use by others and furthermore that the center has an impact on science. In addition, the citation enables a data center to show other users how the data product has been reused in other publications. Government agencies are under increasing pressure to show that the data generated by sponsored research are published and available for reuse (GAO 2007). The GAO report (2007) argues that agencies need to demonstrated the benefits of the research they sponsor, both in terms of scientific findings—published papers— as well as data products.

A recommended data product citation would benefit Societies, Publishers, style manuals, and indexers like ISI or the Web of Science. Citing published data sets is a new practice. Those organizations would understand the rationale for data citations as well as the elements of a data citation.

A citation to a data product should contain the following elements:

- contributing investigators/authors (need to assist providers by defining what effort & contributions constitute authorship)
- year of publication
- data product title (descriptive)
- Data center or publisher, including address URL
- DOI/GUID (machine readable), or, if not available, URL of the data product. Each new version of a data product should have a new DOI/GUID.
- Version number or, if that is not available, date accessed medium (for items other than printed text, such as CD-ROM or data set)

What should be cited?

- Need to cite portions of data products
  - Citable granules, versions, and views as defined and advertised by the data provider
- Need persistent link to the version used in a paper / publication, as well as to newer versions
- Need to comply with provider and data centers Policy
  - Environmental Sciences: notify provider before use

The National Information Standards Organization (NISO), which is a subset of ANSI, which is a member of ISO, has a Working Group on Data Citations. We need to work with them, perhaps have membership or liaison with this Working Group (but do not delay our white paper to wait for their decisions.) The NISO WG has the following goals:

1. Survey and summarize successful data management and citation conventions for existing data repositories.
2. Develop a thesaurus of terms relevant to data sharing
3. Develop guidelines for data citation
4. Work collaboratively with other organizations that are addressing these issues

DataONE will acquire, archive and distribute its own holdings and establish a data citation process. In addition, DataONE will form partnerships with other data center/organizations, each having its own recommended citation so that the data product is associated with that center. DataONE will not claim "ownership" of these data products (as a collector of collections) through a new data citation. The group envisioned that over time if an archiving organization disappears a data product may be "reprinted" by DataONE.

Data Set Citation examples:

Hervé Sinoquet, Sylvain Pincebourde, Boris Adam, Nicolas Donès, Jessada Phattaralerphong, Didier Combes, Stéphane Ploquin, Krissada Sangsing, Poonpipope Kasemsap, Sornprach Thanisawanyangkura, Géraldine Groussier-Bout, and Jérôme Casas. 2009. 3-D maps of tree canopy geometries at leaf scale. Ecological Archives E090-019. http://www.esapubs.org/archive/ecol/E090/019/default.htm

Bowyer, P., N. M. Trodd, and F. M. Danson. 2005. SAFARI 2000 Canopy Structural Measurements, Kalahari Transect, Wet Season 2001. 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/768

Model for Data Citation:

Gary King; Langche Zeng, 2006, "Replication Data Set for 'When Can History be Our Guide? The Pitfalls of Counterfactual Inference'" hdl:1902.1/DXRXCFAWPK UNF:3:DaYlT6QSX9r0D50ye+tXpA== Murray Research Archive [distributor]

Related topics include:

- Dataset needs to be a meaningful collection, not broken down too finely in order to create too many citations.
- Each new version includes a new DOI/GUID.
- The documentation for a data set needs to include a suggested data citation, acknowledgements, link to the data file, description of the data set, parameters reported, where the data have been used, etc.
- This citation practice is intended for data that is submitted to a data center and is finalized.

The White Paper will include links to additional information such as:

1. GBIF How to Cite: http://www.gbif.org/DataProviders/Cite/howToCite

2. Dataverse and their work regarding data citation and identifiers: http://thedata.org/citation/standard

3. NISO Thought Leader Meeting on standards for citing data sets and data collections: http://www.niso.org/topics/tl/NISOTLDataReportDraft.pdf

Research blog on the topic (http://niso-researchdata.blogspot.com/), which includes some commentary from Lucy Nowell on the issue.

4. Ecological Society of America and TDWG data citation activities in our community: http://wiki.tdwg.org/twiki/bin/view/GUID/WebHome

5. TDWG standard on data citation: http://www.tdwg.org/standards/150/

6. ORNL DAAC citation policy http://daac.ornl.gov/citation_policy.html

7. GAO report on Agency actions to make data available: http://www.gao.gov/products/GAO-07-1172

8. Implementing persistent identifiers: Overview of concepts, guidelines and recommendations. London: Consortium of European Research Libraries, 2006. http://bibpurl.oclc.org/web/16923

9. Persistent Identification of Digital Resources Environmental Scan. Library and Archives Canada,

http://www.carl-abrc.ca/projects/nmrdi/Alouette-PersistentID_Scan-e.pdf