Use Case: Citing Integrated Data Sets

Matthew Mayernik and Mary Marlino

National Center for Atmospheric Research

1. Introduction

This paper focuses on challenges that will arise when the movement towards increased data integration meets the movement toward more comprehensive data citations. EarthCube seeks to develop an infrastructure that can enable the integration of data from disparate scientific projects. Such integration will create many new derivative data sets to the point where, if successful, EarthCube will generate a web of interrelated data sets. In parallel, "data citations" are increasingly seen as being critical in enabling scientific results to be traced back to their underlying data. Data citations promote the transparency of scientific work and enable scientists to be credited for producing useful data. As data integration accelerates within an EarthCube infrastructure, it will be even more difficult to create and follow data citations than it is today because of the interconnections between data sets. In this paper, we advance ideas on how to address the following question: How will data citations fit within the envisioned web of integrated EarthCube data?

2. Data Citations

Federal agencies, professional societies, and research organizations in the geo-sciences are calling for researchers to formally cite data that led to a given research result. Such "data citations" promote transparency in research by offering a direct pathway to data so that research can the validated or easily carried forward from a known starting point (Arzberger, et al., 2004; Costello, 2009; Heffernan, 2010; Science Staff, 2011). Data citations are also intended to raise the profile of data, that is, to make data as valued and rewarded in scientific settings as peer-reviewed publications. Data citations should benefit scientific communities in a number of ways, including: 1) formal citations give credit to scientists for their work in collecting and creating data, 2) formal citations allow data center managers to track the use of data sets and gain the benefits of documenting their services and creating a foundation to design better services, and 3) formal citations will help accelerate scientific progress by tightly coupling scholary publications and data, so that two-way discovery and access are common.

In order for data citations to serve these desired roles, however, there must be an alignment of information system development, scientific work practices around data use and citation, bibliometric measurements of data citations, and institutional acceptance of data citations as an indicator of scientific impact.

Many tools exist for identifying and linking to data in a web environment (Brase, 2004; Van de Sompel, et al., 2004; Bizer, 2009; Pepe, et al., 2010). Geo-scientific communities are in the beginning stages of using these tools for citing data. Data citations are intended to identify a particular resource and, in the case of internet resources, indicate where it might be acquired. Before data sets can be cited, however, they must be designated as citable objects with unique identities. The most common type of unique identifier used within our current global scholarly communication systems are Digital Object Identifiers (DOIs). DOIs are intended to sidestep the inherent unreliability of URLs by providing persistent locators for internet-based resources. DOIs are most commonly assigned to journal articles, but are seeing a growing use for data (Paskin, 2005; Cook, 2008; Parsons, Duerr, & Minster, 2010). A number of other digital identification systems provide similar functionalities as DOIs, including Archival Resource Keys (ARKs), Persistent URLs (PURLs), and handles, but DOIs are generally recommended for use in citing scholarly materials because of their familiarity and acceptance among scientific communities and scholarly publishers (Duerr, et al., 2011).

Assigning DOIs to data sets is not a straightforward process. In contrast with journal articles, which do not change over time and have one definitive published form, data sets often have indistinct identities (Wynholds, 2011). Data sets might consist of many individual units (such as files or database tables) or might themselves be subsets of larger data collections. In addition, many data sets change on a daily or weekly basis, as, for example, when new measurements are continuously added to an existing climate data set.

Nascent initiatives within the geo-sciences to develop data citation recommendations attempt to deal with this problem in various ways. The Federation of Earth Science Information Partners (ESIP), for example, has released an initial set of data citation guidelines for data archives (ESIP, 2011). Drawing on data citation lessons learned via the International Polar Year project (Parsons, Duerr, & Minster, 2010), the ESIP guidelines provide pragmatic recommendations for citing changing or highly granular data, such as recommending that citations include the date on which data were downloaded.

These recommendations are very useful, but they do not address how to cite integrated data sets, which is one of the key motivations for EarthCube.

3. Data Integration

The stated goal of the EarthCube initiative is:

> "…create a knowledge management system and infrastructure that integrates all geosciences data in an open, transparent and inclusive manner. …The decade-long vision for EarthCube is the convergence towards an integrated system to access, analyze and share information that is used by the entire geosciences community" (NSF, 2011, pg. 2-3).

The idea of data integration is not in-and-of-itself new, scientists have long been bringing together disparate data sets. The desire to integrate data globally has stimulated many attempts to standardize data collection methods and formats, as exemplified by the SEED format for seismic data and the NetCDF format for atmospheric data. Many data integration projects have been ongoing for decades, such as the continuing releases of the International Comprehensive Ocean-Atmosphere Data Set (ICOADS) which first began in the mid-1980s and most recently took place in 2011 (Woodruff, et al., 2011). The vision of the EarthCube initiative, however, is to enable data integration to be a standard scientific activity, not the time and human-effort intensive activities that they are today.

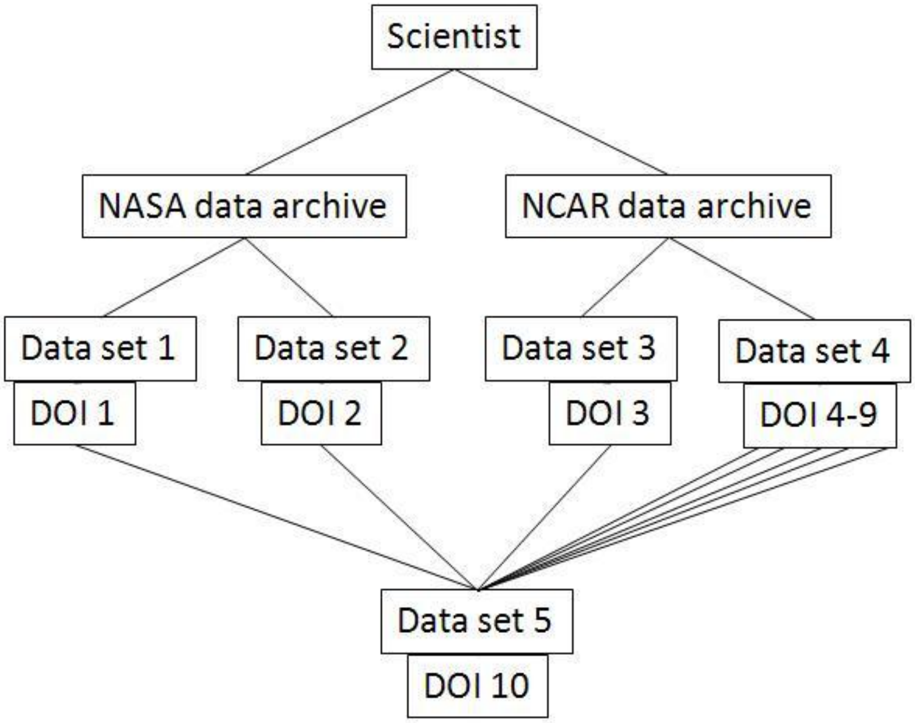
4. Citing Integrated Data

If EarthCube is successful, scientists will find it much easier to combine numerous data sets in innovative ways. As a result, scientists will also create many new derivative data sets. These derivative data sets have strong implications for data citation practices. For example, Woodruff, et al. (2011) list over 150 data sets in extensive tables in their article's appendix in order to indicate the data sets that were integrated into ICOADS. While ICOADS is more comprehensive than most current data integration projects, it is indicative of the scale that such projects can take.

Enabling researchers to compile formal citations to numerous integrated data sets will be a necessary functionality for EarthCube. If done properly, these citations will be able to make visible the relations between data sets and their derivative research products. Increasingly research organizations assign DOIs to data sets. These DOIs provide unique web-accessible identifiers through which linkages between data sets can be traced and declared.

Figure 1 shows a use case of what such data linkages can look like when integrating data from multiple sources. A scientist studying atmospheric temperature variation downloads four data sets: two remote sensing data sets from NASA's data archives and two drop-sonde data sets from NCAR's data archives. Data set 1 and 2 have been assigned DOIs, Data set 3 has been assigned a DOI, and Data set 4 has been assigned five DOIs, one for each of its principle data components. When the scientist integrates all of these data sets, a new data set is created. The scientist then archives the new integrated data set at NASA, NCAR, or her/his home institution. The new data set (Data set 5) is assigned a new DOI.

Figure 1.



How should the scientist cite these five data sets? Should all five be cited, just the originating four, or just the final integrated data set? It is easy to see how the complexity of this use case will increase dramatically if the number of scientists, data sets, and DOIs increase. Each scientist who uses the same data sets might combine them in different ways. Creating citations and tracing the linkages between the data sets will become very complex if the EarthCube system is not designed to capture these data set relationships.

DOI registration agencies are designing data and metadata systems that allow these linkages to be declared, but declaring such linkages is currently a manual process. One example of this is the DataCite metadata schema. The DataCite organization (http://datacite.org/), an international federation of libraries and research organizations, was created to promote the assignment of DOIs to data. Among the services DataCite is developing are a metadata schema and metadata store, specifically for assigning DOIs to data sets. DataCite collects metadata for each data set that is assigned a DOI through their services. The DataCite metadata schema includes a "RelatedIdentifier" field specifically to capture linkages between identifiers (Starr & Gastl, 2011). The "RelatedIdentifier" field has a "relation type" attribute that allows the precise nature of the linkages between data sets to be specified, such as that a data set "IsPartOf" another data set or that a data set "Compiles" a number of other data sets (DataCite, 2011).

The EarthCube infrastructure will need effective ways of identifying and tracing data relationships. Data citations are intended to facilitate that goal, but will become increasingly complex as data integration efforts proceed. At minimum, the EarthCube infrastructure will need to have a simple interface through which scientists or data managers can declare relationships between data sets in ways that enable citations to be compiled easily. Ideally, the EarthCube infrastructure would automatically identify and declare linkages between data sets by creating and parsing appropriate metadata, such as the "RelatedIdentifier" fields in the DataCite metadata schema, when scientists use the EarthCube systems to created integrated data sets.

5. Understanding the Impact of Data and Data Citations

To evaluate the impact of data citations, it will be necessary to count citations to data over time. Assigning DOIs to data simplifies this task, as DOIs provide a unique character string that can be searched for in databases and on the internet, but a prerequisite to any citation counting will be keeping an up-to-date index of DOIs registered to data sets.

Counting citations is an inherently uncertain process. Different citation indexes for journal articles, such as the Web of Science, Scopus, and Google Scholar, will give different citation counts for the same article. Counting citations to data sets is even more difficult because currently there is no citation index for data citations. DataCite is working with Thompson-Reuters to get data DOIs indexed in the Web of Science, but this service, if developed, is likely still a few years off. Citations can be compiled manually by searching through databases and internet search engines for DOIs or other data set identification information, such as titles. This process is very time consuming, but is the default citation chasing method for data sets because no other good method exists. Thus, developing methods for counting and tracking citations to data sets is an open research area.

6. Conclusion: Outreach to the Geo-Science Community

Data citations will have minimal impact on the geo-sciences if such citations are not promoted and rewarded within scientific communities. Anecdotal evidence shows that while scientists do formally cite data in some cases, this is not yet a regular practice in the earth and space sciences (Parsons, Duerr, & Minster, 2010; Cook, 2011). Outreach efforts should focus on raising the profile of data citations through informing scientists of DOI registration services, being proactive in providing scientists with recommended citations, and designing data systems that enable scientists to compile citations to complex integrated data sets. Additionally, as methods for assessing the impact of data citations develop, these impact assessments can promote increased rewards for scientists who produce data and receive data citations.

7. References

Arzberger, P, et al. (2004). Promoting Access to Public Research Data for Scientific, Economic, and Social Development. *Data Science Journal* 3: 135-152. http://www.jstage.jst.go.jp/article/dsj/3/0/135/_pdf

Bizer, C. (2009). The Emerging Web of Linked Data. *IEEE intelligent systems*, *24*(5), 87-92. http://dx.doi.org/10.1109/MIS.2009.102

Brase, J. (2004). Using Digital Library Techniques – Registration of Scientific Primary Data. *Research and Advanced Technology for Digital Libraries, Lecture Notes in Computer Science* (Vol. 3232, pp. 488-494). Springer Berlin / Heidelberg. http://dx.doi.org/10.1007/978-3-540-30230-8_44

Cook, R. (2008). Editorial: Citations to Published Data Sets. *FluxLetter: the Newsletter of FluxNet*, 1(4): 4-5. http://bwc.berkeley.edu/FluxLetter/FluxLetter-Vol1-No4.pdf

Cook, R. (2011). Archiving Earth Science Data: Experiences of the ORNL Distributed Active Archive Center. Presentation at *DataCite 2011 Summer Meeting*, Berkeley, CA, August 25, 2011. http://datacite.org/slides/DataCite2011/DataCite0502-Cook.ppt

Costello, M. J. (2009). Motivating Online Publication of Data. *BioScience*, 59(5), 418-427. http://www.jstor.org/stable/10.1525/bio.2009.59.5.9

DataCite. (2011). *DataCite Metadata Schema for the Publication and Citation of Research Data*. Version 2.1. http://dx.doi.org/10.5438/0003

Duerr, R., et al. (2011). On the utility of identification schemes for digital earth science data: an assessment and recommendations. *Earth Science Informatics*, 1-22. http://dx.doi.org/10.1007/s12145-011-0083-6

Federation of Earth Science Information Partners (ESIP). (2011). Interagency Data Stewardship/Citations/provider guidelines. http://wiki.esipfed.org/index.php/Interagency_Data_Stewardship/Citations/provider_guidelines

Heffernan, O. (2010). Saluting scrutiny. *Nature Reports Climate Change*, 2 March 2010. http://dx.doi.org/10.1038/climate.2010.20

National Science Foundation (NSF). (2011). Earth Cube Guidance for the Community. http://api.ning.com/files/Bc3KSNzGutpyO0DGOxcDlL1yUhSrb-JFg7uZ-J3WBsmPsMSh7syVdjxru5DmHhk561GC5RhUQb9SY-8ro6uOHs7rIW8TBveV/nsf11085.pdf

Parsons, M. A., Duerr, R., & Minster, J.-B. (2010). Data Citation and Peer Review. *Eos Transactions, AGU*, 91(34). http://dx.doi.org/10.1029/2010EO340001

Paskin, N. (2005). Digital Object Identifiers for scientific data. *Data Science Journal*, 4(0), 12-20. http://www.doi.org/topics/050210CODATAarticleDSJ.pdf

Pepe, A., Mayernik, M.S., Borgman, C.L., & Van de Sompel, H. (2010). From Artifacts to Aggregations: Modeling Scientific Life Cycles on the Semantic Web. *Journal of the American Society for Information Science and Technology*, 61(3): 567-582. http://dx.doi.org/10.1002/asi.21263

Science Staff. (2011). Challenges and Opportunities. *Science*, 331(6018): 692-693. http://dx.doi.org/10.1126/science.331.6018.692

Starr, J. & Gastl, A. (2011). isCitedBy: A Metadata Scheme for DataCite. *D-Lib Magazine*, 17(1/2). http://www.dlib.org/dlib/january11/starr/01starr.html

Van de Sompel, H., Payette, S., Erickson, J., Lagoze, C., & Warner, S. (2004). Rethinking scholarly communication: Building the system that scholars deserve. *D-Lib Magazine* 10(9). http://www.dlib.org/dlib/september04/vandesompel/09vandesompel.html

Woodruff, S. D., et al. (2011). ICOADS Release 2.5: extensions and enhancements to the surface marine meteorological archive. *International Journal of Climatology*, *31*(7), 951-967. http://dx.doi.org/10.1002/joc.2103

Wynholds, L. (2011). Linking to Scientific Data: Identity Problems of Unruly and Poorly Bounded Digital Objects. *International Journal of Digital Curation*, 6(1). http://www.ijdc.net/index.php/ijdc/article/viewFile/174/242