Data Citation Initiatives and Issues

by Matthew S. Mayernik

Research Data Access & Preservation (RDAP)

EDITOR'S SUMMARY

The importance of formally citing scientific research data has been recognized for decades but is only recently gaining momentum. Several federal government agencies urge data citation by researchers, DataCite and its digital object identifier registration services promote the practice of citing data, international citation guidelines are in development and a panel at the 2012 ASIS&T Research Data Access and Preservation Summit focused on data citation. Despite strong reasons to support data citation, the lack of individual user incentives and a pervasive cultural inertia in research communities slow progress toward broad acceptance. But the growing demand for data transparency and linked data along with pressure from a variety of stakeholders combine to fuel effective data citation. Efforts promoting data citation must come from recognized institutions, appreciate the special characteristics of data sets and initially emphasize simplicity and manageability.

KEYWORDS

research data sets
bibliographic citations
data set management
access to resources
information infrastructure
motivation
cultural aspects

Matthew Mayernik is a research data services specialist in the library of the National Center for Atmospheric Research (NCAR)/University Corporation for Atmospheric Research (UCAR). His work within the NCAR/UCAR library is focused on developing research data services, including developing a policy and implementation plan for data citations. His research interests also include metadata practices and standards, cyber-infrastructure development and social aspects of research data. He can be reached at mayernik<at>ucar.edu.

Many institutions across academic disciplines are investigating and promoting *data citations*. A data citation, as the term suggests, is a citation included in the reference list of a published article that formally cites data that led to a given research result. Formal data citations are becoming more common, but are still more the exception than the rule across disciplines. Authors of research papers typically refer to their source data in other ways, such as free-text mentions in the methods section of a paper or a mention in the acknowledgements section of a paper. Data citations, on the other hand, appeal to many data producers, repositories and research funders because of their potential to become an accountable part of the scholarly communication process. If researchers begin to cite their source data through the formal citation mechanism, data sets can potentially be better assessed for their impact on the scientific community through the compilation of data citation indices.

Broad Interest and Activity

Interest in data citations is emerging in many sectors. Multiple federal agencies have held workshops related to data citations within the past year. The U.S. Committee on Data for Science and Technology (CODATA) and the National Academies Board on Research Data and Information collaborated with the CODATA-International Council for Scientific and Technical Information Task Group on Data Citation Standards and Practices to sponsor a workshop in August 2011 on data citation and attribution [1]. That workshop is part of a larger set of CODATA activities related to data citation and will eventually result in a consensus report on the topic.

The NSF Directorate for Geosciences has also been actively studying and promoting data citations. The NSF funded a workshop in March 2011 titled "Geo-Data Informatics: Exploring the Life Cycle, Citation and Integration

of Geo-Data." The report that resulted from this workshop provides a range of recommendations related to data citation, including recommendations relating to journals, professional societies, standards organizations, data repositories, educators and research funders [2]. In addition, the NSF Directorate for Geosciences released a "Dear Colleague Letter" on March 29, 2012, specifically addressing data citations [3]. The goal of the letter is to stimulate geosciences communities to "establish data citation within the geosciences as the rule rather than the exception."

Other organizations are also contributing to the growing data citation movement. The most prominent is the DataCite organization [4]. DataCite is an international federation of libraries and data centers that is promoting the practice of citing data. From a technical point of view, DataCite provides digital object identifier (DOI) registration services specifically designed for datasets. Founded in 2009, DataCite has grown rapidly and has members on four continents.

Additional activity has focused on developing guidelines and recommendations for how to cite datasets. The Federation for Earth Science Information Partners (ESIP) in the United States and the Digital Curation Centre (DCC) in the United Kingdom both released data citation guidelines in 2011. The ESIP guidelines are targeted to data repositories in the geosciences [5], while the DCC guidelines are more discipline-agnostic and are targeted toward multiple stakeholders, including authors, repositories and publishers [6].

In another example of the interest in data citations at local scales, the University Corporation of Atmospheric Research in Boulder, Colorado, held a workshop April 5-6, 2012, titled "Bridging Data Lifecycles: Tracking Data Use via Data Citations," that drew over 75 attendees from university libraries, federal agencies, research organizations and others [7]. Finally, the 2012 ASIS&T Research Data Access and Preservation Summit (RDAP) featured a panel on data citations [8].

Data Citation Motivations

The main motivators for this interest in data citation are sundry: 1) the desire to connect research publications to their underlying data, 2) the desire to understand the use and impact of data, 3) to promote professional credit and rewards for creating and managing data collections and 4) to promote scientific and data transparency. These motivations, however, provide little direct incentive for data users to actually cite data in a formal way. As Parsons, Duerr and Minster noted in a 2010 article, "[t]he scientific method and the credibility of science rely on full transparency and explicit references to both methods and data. These require that science data be open and available without undue and proprietary restriction. However, a consistent, rigorous approach to data citation is lacking." [9, p. 297] Both of the aforementioned CODATA and NSF workshops noted the strong cultural inertia against data citations within many research communities. Researchers receive professional credit for publishing papers and receiving citations to those papers, not for producing datasets and receiving citations to those datasets. In addition, data users may not know how to cite data or that they are being asked to cite data at all.

While this cultural inertia on the part of data users is widely noted today, it is not a newly discovered phenomenon. Sieber and Trumbo, in a 1995 article, noted that social scientists (primarily sociologists, political scientists and psychologists) largely did not provide proper citations to data, stating "[our studies] show that researchers' behavior, attitudes and knowledge concerning the citation of datasets fall short of the ideal that would foster openness, fairness and economy in the pursuit of scientific knowledge" [10, p. 18]. Looking even farther back, Howard D. White noted this need for formal citation of datasets in 1982: "An argument by no means new is that social scientists who work with machine-readable data files (MRDF) should cite them in their writings, with formal references set apart from main text, just as they now do books, papers and reports" [11, p. 467].

White's comment is quite striking given the growing visibility of data citations today. Why is "an argument by no means new" in 1982 – that data use should be formally cited – now the focus of multiple workshops and international initiatives?

This increasing visibility can be most strongly attributed to the interrelations of two factors: 1) calls from both the scientific and public communities for greater transparency and openness of scientific research [12, 13] and 2) the availability of tools for identifying and linking to data in a web environment [14, 15, 16].

Neither of these two factors in isolation, broad desire for transparency or web-based linking tools, has been enough to produce the broad interest in data citation that we see currently. Data centers have long sought recognition and attribution for their services and products, and, as the above Sieber and Trumbo and White quotes illustrate, the lack of effective data citation practices is not a new phenomenon. In parallel, web-based linking technologies have been continuously evolving since the mid-1990s. Even DOIs, the persistent linking mechanism most widely recommended for data citations, were developed in 1998 in the early days of the web boom [17].

In combination, however, these two factors build on each other. Within the geosciences in particular, the so-called "Climategate" controversy brought considerably more public scrutiny onto scientific and data management processes. The many independent audits of the scientists whose stolen emails were at the center of the controversy found that no impropriety actually took place, though the audits noted that the scientists were in some situations reluctant to provide data to people who had requested it [18]. With web-based linking tools, particularly DOIs, becoming more familiar within scientific communities, calls for transparency build on and supplement calls for using these linking mechanisms to connect data to scholarship [19]. In fact, the initial mainstream media coverage of the "Climategate" stolen emails began in early December of 2009, within days of the official founding of the DataCite organization [20]. While the timing was coincidental (preparations for the founding of DataCite had been ongoing for months by December 2009), the confluence of interest in data transparency and dataset linking mechanisms during that time has stimulated many of the activities noted above.

In addition, these two factors affect – and leverage the expertise of – multiple stakeholders. Scientists and data centers are obviously impacted by any additional scrutiny on data and other scientific products and would benefit from any new tools and services that increase the availability and usefulness of data. In addition, as techniques and tools used for traditional library collections are expanded to address datasets [21], libraries have a lot to contribute to data citation initiatives. Libraries, along with scholarly publishers, bring knowledge of the academic publishing industry and experience with DOIs and other persistent linking mechanisms. Different stakeholders can contribute in different ways to data citation initiatives, increasing the broad base of participants at local, national and international levels.

Considerations for Data Citation Initiatives

Many characteristics of data make them difficult to cite. Three key considerations when developing a data citation initiative are

- citations must be promoted in the context of larger curatorial institutions;
- datasets often have indistinct identities; and
- simplicity is a must, especially in the beginning.

Citations must be promoted in the context of larger data-curation institutions. A dataset cannot be cited if it is not archived. Journal publishers are very aware of the need to have citations that are persistent over time. For example, as of this writing, the American Geophysical Union's policy for referencing data states that authors should only cite data if the data are archived in stable archives. The AGU policy states that "[d]ata sets that are available only from the author, through miscellaneous public network services, or academic, government or commercial institutions not chartered specifically for archiving data, may not be cited in AGU publications." [22] This quote is an explicit acknowledgement that data curation and preservation over time is an institutional responsibility; individuals cannot be trusted with this task.

A recent editorial in *Science*, however, shows that the majority of researchers do not formally archive their data [23]. Most datasets are stored on personal and lab servers and not submitted to data repositories. Thus, in addition to promoting the act of citing data, data citation initiatives must promote the logical prior act of formally archiving data.

Datasets often have indistinct identities. Defining a *dataset* is a difficult and highly situated task. Digital objects can be considered to be part of a dataset if they fall in a particular grouping, contain representations of particular content, are considered to be related or can be used for a particular purpose [24]. Datasets are often combined into composite datasets or pulled apart into sub-sets. In addition, many datasets, such as climate observations, stock prices and social media usage metrics, are highly dynamic, changing on a daily or weekly basis.

Wynholds describes how this dynamism leads to difficulties in establishing distinct identities for datasets [25]. To have a distinct identity, according to Wynholds, a dataset must have at least four characteristics: a) it must be a semantically concrete object; b) it must have its identity embedded and/or inseparable from the data objects themselves; c) it must have a stable notion of authorship; and d) it must be translatable into a mechanism for retrieval and citation.

Each of these characteristics is problematic in relation to citing datasets. Datasets are often highly dynamic, as noted above, which means that semantic concreteness is an elusive characteristic. Duerr, et al., categorize this problem as the need to establish the "scientific uniqueness" of a dataset and note that it is a problem that no existing web-based identifier scheme such as DOIs or Handles can solve [26]. Embedding identity into datasets is also difficult. Books have title pages while journal articles and conference proceedings have embedded identities in the form of standard headers on the first page that list the article title and author(s). Datasets have no standard form of embedded identity. Individual data repositories can develop standard practices for creating embedded headers, but such practices rarely cross institutional or organizational borders. Authorship itself is a problematic notion in relation to datasets. Datasets are very commonly the product of collaborations. Different collaborators on the same project may have very different notions about dataset (and metadata) authorship and responsibility [27]. Properly attributing authorship credit must be negotiated amongst collaborators and can be a fraught political issue. From a dataset retrieval point of view, dataset identity is typically instantiated through persistent web identifiers such as DOIs, persistent URLs and Handles. Assigning these kinds of actionable locators, however, to objects that are not concrete, do not have embedded identities and have varying forms of authorship is a challenge that is the subject of active research.

Simplicity is a must, especially in the beginning. With many datasets having such amorphous identities, the citation recommendations being developed by organizations such as ESIP and the DCC necessarily must be flexible. As noted in the introduction of this article, however, data citations are currently rare. The ambiguity of the data citation task is one contributor to the current paucity of data citations, as are the established cultural norms for acknowledging data use in other ways.

Cultural norms in scientific and research communities are difficult to change. There is currently strong cultural inertia against citing data (researchers get professional credit for citations to peer-reviewed publications, not datasets). In a vivid illustration, Parsons, Duerr and Minster conducted an informal study of the use of the NASA MODIS (Moderate Resolution Imaging Spectroradiometer) snow-cover dataset provided by the National Snow and Ice Data Center in Boulder, Colorado [9]. They show that the use of the MODIS dataset increased at a steady pace between 2002 and 2009, but that very few published papers during that time period formally cited MODIS data in their bibliographies. In an April 2012 presentation at the UCAR "Bridging Data Life Cycles" workshop [7], Tim Killeen, the assistant director of the NSF Geosciences Directorate, indicated that the NSF's goal is to raise the rate of data citations from the current level (estimated at 5%) to 50% in the next five to 10 years [28].

To see any change in current practice, organizations promoting data citations must be crystal clear in their recommendations and policies. Data users must know exactly what they are being asked to do. They must know what they are supposed to cite, when to cite it and how to cite it. Data repositories must interpret data-citation recommendations such as the aforementioned ESIP and DCC recommendations as appropriate for their local collections and must present suggested citations to data users with little ambiguity. This clarity will require data providers to make decisions about the dataset identity challenges noted above. In addition, any tools that make it easier for users to find and access the proper citations for datasets should be built into data repositories. For example, data repositories can provide one-click downloads of citations in RIS or BibTeX format, so that users can easily import data citations into their citation management software like EndNote, Zotero or RefWorks or document creation software like LaTeX.

Conclusion

Despite all of the challenges and issues still to be ironed out, the potential benefits and utility of a robust data citation infrastructure ensure that data

citation initiatives will continue apace at the organizational, national and international levels. Data citations are one step toward more integrated and transparent data collections and supplement parallel movements promoting linked data approaches to digital information provenance tracking [29, 30]. Data citations also promise to bring visibility to the considerable efforts of researchers and data archives to make data accessible and usable. Data citations are not yet integrated into organizational processes for hiring, promotion and tenure, but as more data users acknowledge data use in formal ways, research and funding organizations will have more ways to account for the value of "data work" to the wider research communities.

Acknowledgements

I would like to thank all of the members of the UCAR/NCAR data-citation working group, as well as the National Oceanographic and Atmospheric Administration (NOAA) and the UCAR Joint Office for Science Support for funding the data citation workshop noted in reference [7]. ■

Resources Mentioned in the Article

[1] U.S. CODATA, Board on Research Data and Information, & CODATA-ICSTI Task Group on Data Citation Standards and Practices. (August 22-23, 2011). *Developing data attribution and citation practices and standards: An international symposium and workshop.* Policy and Global Affairs Division, Board on Data Research and Information, National Academy of Sciences. Retrieved April 21, 2012, from http://sites.nationalacademies.org/PGA/brdi/PGA_064019

[2] *NSF Geo-Data Informatics: Exploring the life cycle, citation and integration of geo-data: Workshop report.* (2011). Retrieved April 21, 2012, from http://tw.rpi.edu/media/latest/WorkshopReport_GeoData2011.pdf

[3] National Science Foundation. (March 29, 2012). *Dear colleague letter - Data citation.* (NSF 12-058). Retrieved April 21, 2012, from www.nsf.gov/pubs/2012/nsf12058/nsf12058.jsp?WT.mc_id=USNSF_25&WT.mc_ev=click

[4] *DataCite:* http://datacite.org/

[5] Federation of Earth Science Information Partners (ESIP). (2011). *Interagency data stewardship/Citations/Provider guidelines.* Retrieved April 21, 2012, from http://wiki.esipfed.org/index.php/Interagency_Data_Stewardship/Citations/provider_guidelines

[6] Ball, A. & Duke, M. (2011). How to cite datasets and link to publications. *DCC How-to Guides.* Edinburgh: Digital Curation Centre. Retrieved April 21, 2012, from www.dcc.ac.uk/resources/how-guides

[7] *Bridging data lifecycles: Tracking data use via data citation [workshop].* (April 5-6, 2012). Agenda and presentations retrieved April 21, 2012, from http://library.ucar.edu/data_workshop/

[8] American Society for Information Science and Technology. (2012). *ASIS&T Third Annual Research Data Access and Preservation Summit 2012 (RDAP12): Program.* Retrieved April 21, 2012, from http://rdap12.posterous.com/pages/program

[9] Parsons, M. A., Duerr, R., & Minster, J.-B. (2010). Data citation and peer review. *Eos, Transactions, American Geophysical Union, 91*(34), 297-298. Retrieved April 21, 2012, from http://dx.doi.org/10.1029/2010EO340001

[10] Sieber, J. E. & Trumbo, B. E. (1995). (Not) giving credit where credit is due: Citation of data sets. *Science and Engineering Ethics, 1*, 11-20.

[11] White, H. D. (1982). Citation analysis of data file use. *Library Trends, 31*(3), 467-477.

[12] Costello, M .J. (May 2009). Motivating online publication of data. *BioScience, 59*(5), 418-427. Retrieved April 21, 2012, from www.jstor.org/stable/10.1525/bio.2009.59.5.9

[13] Heffernan, O. (March 2010). Saluting scrutiny. *Nature Reports Climate Change, 2*(3). Retrieved April 21, 2012, from http://dx.doi.org/10.1038/climate.2010.20

[14] Brase, J. (2004). Using digital library techniques – Registration of scientific primary data. *Research and advanced technology for digital libraries* [pp. 488-494]. *Lecture Notes in Computer Science, (3232).* Berlin: Springer. Retrieved April 21, 2012, from http://dx.doi.org/10.1007/978-3-540-30230-8_44

Resources continued on next page

Resources Mentioned in the Article, cont.

[15] 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). Retrieved April 21, 2012, from www.dlib.org/dlib/september04/vandesompel/09vandesompel.html

[16] 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. Retrieved April 21, 2012, from http://dx.doi.org/10.1002/asi.21263

[17] Paskin, N. (2005). *Digital object identifiers for scientific data. Data Science Journal, 4*, 12-20. Retrieved April 21, 2012, from www.doi.org/topics/050210CODATAarticleDSJ.pdf

[18] Hasselmann, K. (2010). The climate change game. *Nature Geoscience, 3*(8), 511–512. Retrieved April 21, 2012, from http://dx.doi.org/10.1038/ngeo919

[19] Cook, R. (December 2008). Editorial: Citations to published data sets. *FluxLetter: The Newsletter of* FluxNet, *1*(4), 4-5. Retrieved April 21, 2012 from http://bwc.berkeley.edu/FluxLetter/FluxLetter-Vol1-No4.pdf

[20] Dates for the "Climategate" and DataCite events were taken from http://en.wikipedia.org/wiki/Climategate and http://en.wikipedia.org/wiki/Datacite respectively. Both websites were accessed April 10, 2012.

[21] Buckland, M. (2011). Data management as bibliography. *Bulletin of the American Society for Information Science and Technology, 37*(6), 34-37. Retrieved April 21, 2012, from www.asist.org/bulletin /Aug-11/AugSep11_Buckland.html

[22] American Geophysical Union (AGU). (1996). *Policy on referencing data in and archiving data for AGU publications.* Retrieved April 21, 2012, from www.agu.org/pubs/authors/policies/data_policy.shtml

[23] Science Staff. (2011). Challenges and opportunities. *Science, 331*(6018), 692-693. Retrieved April 21, 2012, from http://dx.doi.org/10.1126/science.331.6018.692

[24] Renear, A., Sacchi, S., & Wickett, K. (2010). Definitions of dataset in the scientific and technical literature. *Proceedings of the Annual Meeting of the American Society for Information Science and Technology, 47*, 1-4. Retrieved April 21, 2012, from http://dx.doi.org/10.1002/meet.14504701240

[25] Wynholds, L. (2011). Linking to scientific data: Identity problems of unruly and poorly bounded digital objects. *International Journal of Digital Curation, 6*(1). Retrieved April 21, 2012, from www.ijdc.net/index.php/ijdc/article/viewFile/174/242

[26] Duerr, R. E., Downs, R. R., Tilmes, C., Barkstrom, B., Lenhardt, W. C., Glassy, J., Bermudez, L. E., et al. (2011). On the utility of identification schemes for digital earth science data: an assessment and recommendations. *Earth Science Informatics, 4*(3), 139-160. Retrieved April 21, 2012, from http://dx.doi.org/10.1007/s12145-011-0083-6

[27] Wallis J. C., & Borgman C. L. (2011). Who is responsible for data? An exploratory study of data authorship, ownership and responsibility. *Proceedings of the Annual Meeting of the American Society for Information Science and Technology, 48*, 1-10. Retrieved April 21, 2012, from http://dx.doi.org/10.1002/meet.2011.14504801188

[28] Killeen, Timothy. (2012). *Data citation motivation and policies* [slides]. Presentation at the workshop "Bridging Data Lifecycles: Tracking Data Use via Data Citations," held at University Corporation for Atmospheric Research (UCAR), Boulder, CO, April 5-6, 2012. Retrieved April 21, 2012, from http://library.ucar.edu/data_workshop/presentations/Killeen_BridgingDataLifecyclesWorkshop.pdf

[29] Bizer, C. (2009). The emerging web of linked data. *IEEE Intelligent Systems, 24*(5), 87-92. Retrieved April 21, 2012, from http://dx.doi.org/10.1109/MIS.2009.102

[30] Moreau, L., Clifford, B. Freire, J., Futrelle, J., Gil, Y., Groth, P., et al. (June 2011). The Open Provenance Model core specification (v1.1). *Future Generation Computer Systems, 27*(6), 743-756. Retrieved April 21, 2012, from http://dx.doi.org/10.1016/j.future.2010.07.005