EOS, TRANSACTIONS, AMERICAN GEOPHYSICAL UNION

VOLUME 91 NUMBER 34
24 AUGUST 2010
PAGES 297–304

Data Citation and Peer Review

PAGES 297–298

A scientific publication is fundamentally an argument consisting of a set of ideas and expectations supported by observations and calculations that serve as evidence of its veracity. An argument without evidence is only a set of assertions. Consider the difference between the statement "The hairy woodpecker population is declining in the northwest region of the United States" and the statement "Hairy woodpecker populations in the northwest region of the United States have declined by 11% between 1992 and 2003, according to data from the Institute for Bird Populations (http://www.birdpop.org/)." Both or neither of these statements could be true, but only the second one can be verified. Scientific papers do, of course, present specific data points as evidence for their arguments, but how well do papers guide readers to the body of those data, where the the data's integrity can be further examined? In practice, a chasm may lie across the path of a reviewer seeking the source data of a scientific argument.

The collective text that describes scientific knowledge, consisting of peer-reviewed publications connected by citations, is strained by the vast amounts of data in the digital age. Rules and practices are well established for text but less so for data. Yet data are as vital to scientific knowledge as publications are. In a position statement that was revised and reaffirmed in May 2009, (http://www.agu.org/sci_pol/positions/geodata.shtml), the AGU Council asserts that the scientific community should recognize the value of data collection, preparation, and description and that data "publications" should "be credited and cited like the products of any other scientific activity." It further encourages peer review of such publications. These are important assertions with significant ramifications. Currently, authors rarely cite data formally in journal articles (see Figure 1), and they often lack guidance on how data should be cited. Data can be much more dynamic than traditional publications, yet it is often scientifically critical to indicate exactly which version of a data set was used to generate a particular result.

Data peer review is even more complex. Data centers have few established practices for peer review of data. Indeed there is no clear definition of what peer review of data really means. Is it a review of data accuracy or validity, or is it a review of data documentation to ensure complete description of uncertainty and context? Despite these challenges, scientists and data managers have a professional and ethical responsibility to do their best to meet the data publication goals asserted by AGU.

The Earth and space science data community has been discussing data publication issues for decades. In recent years, the Federation of Earth Science Information Partners and AGU's Earth and Space Science Informatics Focus Group have sponsored data publication conference sessions, working groups, and discussion fora including a town hall meeting at the 2009 AGU Fall Meeting (see http://wiki.esipfed.org/index.php/Interagency_Data_Stewardship/2009AGUTownHall). As a result, some best practices and critical research needs are beginning to emerge, and scientists are collectively calling for greater attention to these practices and needs.

Lack of a Consistent Method for Data Citation

The 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.

Data citation has been described in the literature [e.g., *Klump et al.,* 2006; *Costello,* 2009], and many geophysical data centers,

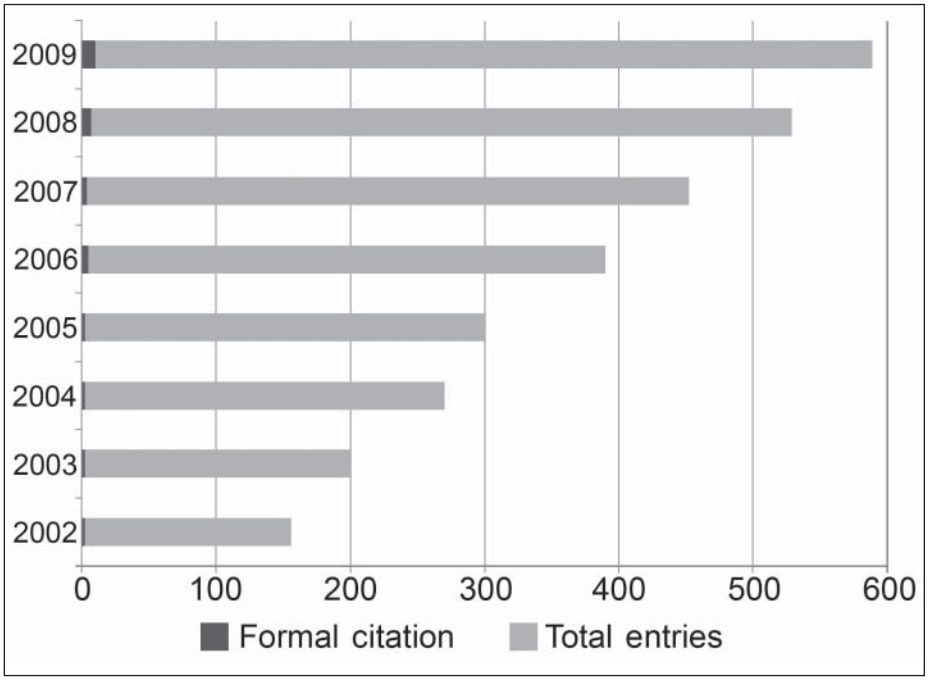

Fig 1. The National Snow and Ice Data Center distributes a variety of different snow cover products derived from the Moderate Resolution Imaging Spectroradiometer (MODIS). The results of a quick analysis of how many scientific papers mention use of "MODIS snow cover data" (according to Google Scholar™) and how often the data sets themselves are formally cited show a huge disparity, illustrating the infrequency of proper data citation in practice. Moreover, the lack of data citation standards introduces the possibility that informal references to data do not point to the data set actually used.

By M. A. Parsons, R. Duerr, and J.-B. Minster

including most NASA centers, recommend specific ways to cite their data. However, their approaches vary. Some data centers, including the U.S. National Oceanic and Atmospheric Administration's (NOAA) National Data Centers, do not request formal citation; they simply request that data be acknowledged in the text. Some data centers, including some U.S. Geological Survey centers, take different approaches for different products. For example, citation may be requested for digital maps, while only acknowledgment may be requested for tabular data.

Occasionally a data publisher may request that data users cite a journal article or other document describing the data. Ironically, these types of citations seem to be broadly used despite the fact that the citation does not directly refer to the actual data used. In some cases, the data may actually be a supplement to the article (e.g., http://dx.doi.org/10.1594/PANGAEA.727522); more often, though, the data extend well beyond a specific article.

For example, the recommended citations for the widely cited (and controversial) global temperature data sets from the Climatic Research Unit (CRU) of the University of East Anglia are a variety of papers published in journals such as the *International Journal of Climatology* and *Journal of Geophysical Research-Atmospheres*. While anyone using the CRU data should read and probably cite the suggested articles, they are static publications and do not contain the actual gridded data. Furthermore, the data continue to evolve and change in ways not documented in the original articles. It is noteworthy that one issue in the recent controversy over the e-mails stolen from CRU was the availability of certain data and the techniques applied to those data—information not available in the referenced journal articles.

Toward Standardized Data Citation: The International Polar Year Model

The International Polar Year (IPY), a huge, interdisciplinary initiative of the International Council for Science and the World Meteorological Organization, explicitly recommends data citation in the IPY Data Policy (see http://classic.ipy.org/Subcommittees/final_ipy_data_policy.pdf) and has developed guidelines for how data should be cited. These guidelines, like any, are imperfect, but they harmonize different approaches and have been adopted by many data centers around the world. They can be used now and serve as the basis for evolving approaches to formally citing data.

The IPY guidelines recommend an approach much like citing a book. Elements of citation include author, editor, publication date and version, data set title, publisher, access date, and unambiguous data location or medium. More details can be found in the full guidelines (see http://ipydis.org/data/citations.html). An example of a citation in this format is as follows:

> Cline, D., R. Armstrong, R. Davis, K. Elder, and G. Liston. 2002, Updated July 2004. *CLPX-Ground: ISA snow pit measurements*. Edited by M. A. Parsons and M. J. Brodzik. Boulder, CO: National Snow and Ice Data Center. Data set accessed 2008-05-14 at http://nsidc.org/data/nsidc-0176.html.

Authors are those who put the intellectual effort into collecting and preparing the data. They may be data collectors, team leaders, algorithm developers, etc. In this example, the authors, Cline et al., designed the experiment and data collection protocol and oversaw the data collection process conducted by dozens of people. Many other people may be involved in creating a quality data set, by creating metadata, processing data, running quality control procedures, etc. In the above example, Parsons and Brodzik are credited as editors. They oversaw the transfer of data from field notebooks to digital files, established and conducted manual and automatic quality control processes, and determined the final data formats.

Most of the other elements are straightforward. The publication date is the date the data were made available (as opposed to the period of data coverage). This date may change as new data set versions are created. The data publisher is typically a data center, but it could be a university or other institution. The International Council for Science World Data System, currently in development, is defining the data publisher role more formally (see http://www.icsu-wds.org/wds-members/join-icsu-wds/wds-components). Data access date is also important because changes in data or calibrations are not always captured as new versions (e.g., ongoing time series). The last element of the citation, the location or medium, is perhaps the most difficult. In the above example it is simply a URL, but sometimes data may be published on media such as a DVD. Digital Object Identifiers (DOIs) are increasingly being used as a way to precisely indicate which particular data set was used and to enable traceability to the original source.

The use of DOIs and other unique and persistent identifiers is increasing. With ever-growing data volumes and the reprocessing of lower level data into new products, it is critical to determine precisely which data were used to generate a result and to be able to access those exact data. Unfortunately, no one identifier scheme has emerged to meet all the needs of scientific data publication. R. Duerr et al. (On the utility of identification schemes for Earth science data: An assessment and recommendations, manuscript in preparation, 2010) review the various identifiers and how they serve different purposes. These purposes include the need to uniquely and unambiguously identify a particular data set or subset no matter which copy a user has (e.g., universally unique identifiers); the need to locate data no matter where they are currently held (e.g., handles, persistent uniform resource locators, object identifiers); and the need to determine if two files contain the same data (i.e., are scientifically identical) even if the formats are different. Related to identifiers is the need to establish conventions on what constitutes a data publication. What is the citable unit with a DOI? A file? A collection of files? How many? Further, it is important to note that data products can be purged from an archive; such deleted information still needs to be able to be referenced. Even if the products themselves are not preserved, the raw data must be preserved along with detailed documentation describing how the product was created, and that documentation must be citable.

Ultimately, more is needed to develop completely unambiguous ways to cite data precisely, but it is reasonable to work now within existing norms of publication to cite data as clearly as possible. The IPY guidelines and similar approaches may indeed work well for relatively complete data collections, especially when they are well described. But for this method of data citation to be fully effective, journal editors and reviewers would need to be more rigorous in demanding that authors accurately cite the data they use in their research.

However, data citation goes well beyond journals and penetrates deeply into the overall culture of science. A data citation not only identifies data used in a study but also is a way to recognize and hold accountable the authors of data. Data publication should be tracked and assessed just like article publication in funding, promotion, and tenure decisions. Currently, someone who publishes really good data receives less credit than someone who publishes a minor paper in a journal. This culture of rewarding only papers and not data will not change until the scientific community collectively works to change academia's centuries-old approach to faculty assessment, promotion, and award recognition.

Data Peer Review

A fundamental issue for acknowledging and rewarding data collection is determining what constitutes "really good" data. Even the minor paper in a minor journal undergoes a formal peer review; such a process has not been established for data.

In many ways, good data have always undergone some level of peer review, and many NASA and NOAA data centers vet the data they handle, but there is no formally recognized or established process. Developing that process is a greater challenge than data citation, but it is no less vital to modern, data-driven science. The first step is defining what is meant by data peer review. One participant at the AGU town hall meeting suggested that presenting non-peer-reviewed data is like presenting a paper at a conference—interesting but incomplete. A quality, peer-reviewed data

set should have sensible, understandable documentation and all the contextual information needed to truly understand and use the data (e.g., calibration information).

Some take this to its logical extreme and have established specialized journals, such as *Earth System Science Data* (http://earth-system-science-data.net/), as a means for publishing high-quality data and all their relevant documentation in a classically peer-reviewed journal. This may work well for certain, high-quality, benchmark data sets, but it seems unlikely and impractical that this can be scaled to cover all types of data or new versions of data. Indeed, an important question in any peer-review scheme will be how to handle different versions of the same data set. When does a new version require additional review? Does the earlier review still apply? An alternate approach suggested by a town hall participant is that peer review is more like auditing. It answers the question of whether scientists are following established professional principles in producing their data. In this model, the data center and data "editors" play a critical role in establishing those professional principles.

Others suggest various open review processes used by some journals and more informal approaches, such as simply capturing online comments much like Amazon and other commercial sites. Historically, informal processes like these have identified problems with data that have later been corrected. For example, a user accessing a long, satellite-derived sea ice time series noticed some bad scans that were not adjusted in routine processing. These errors would have biased analyses of the data if left uncorrected, but once notified, the data center was able to correct the error and notify users of the changes. This highlights the need to track data versions and to ensure transparency in how data are used and assessed.

Indeed, data use in its own right provides a form of review. If data are broadly used and this use is recorded through citation, it indicates a certain level of confidence in the data. Of course, this is not an objective assessment, and broad citation can also be pointing out failings in the data.

It is also important to note that data quality is not simply a function of the data but, rather, a function of the data application. Data that are appropriate for one use are often totally inappropriate for another. A quality, reviewed data set would include documentation describing appropriate use. This illustrates a difference between quality assurance and peer review. Quality assurance might best be done under the auspices of the producers of the data, while peer review should be done by independent groups or individuals and may be more the responsibility of the data publisher or data center. An independent peer review also enhances the credibility of the data publication. Just like an author of a peer-reviewed paper receives greater recognition than an author of a report in the gray literature, so should a peer-reviewed data author receive greater credit than the author of a more casual data publication.

Ultimately, data publishers will likely have a central role in establishing appropriate peer-review processes. Best practices for the scientific community should be developed that address basic data management issues such as standard formats and data validation, and more complex community issues such as scaling—what level of assurance is necessary to apply at large scales when millions of data files may be produced? For example, is academic review of processing algorithms, such as those documented in NASA's Algorithm Theoretical Basis Documents, sufficient? It is rigorous, but does it receive the same recognition as peer review? How does this contrast with review processes for research data collections produced by individual investigators or small projects that rarely produce the level of documentation or undergo the levels of review of the large programs?

These, along with data citation, are the sorts of issues the data management community is beginning to address in collaboration with scientific researchers. AGU's reaffirmed position statement on data can guide these future endeavors.

References

Costello, M. J. (2009), Motivating online publication of data, *BioScience, 59*(5), 418–427, doi:10.1525/bio.2009.59.5.9.

Klump, J., R. Bertelmann, J. Brase, M. Diepenbroek, H. Grobe, H. Höck, M. Lautenschlager, U. Schindler, I. Sens, and J. Wächter (2006), Data publication in the open access initiative, *Data Sci. J., 5,* 79–83, doi:10.2481/dsj.5.79.

Author Information

Mark A. Parsons and Ruth Duerr, National Snow and Ice Data Center, Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder; E-mail: parsonsm@nsidc.org; and Jean-Bernard Minster, Institute of Geophysics and Planetary Physics, Scripps Institution of Oceanography, University of California, San Diego, La Jolla

NEWS

Survey Highlights Search for Habitable Extrasolar Planets

PAGE 299

The search for nearby, habitable planets outside of our solar system is one of three priority science objectives identified by the U.S. decadal survey of astronomy and astrophysics for 2012–2021, released on 13 August. The other top objectives outlined in the U.S. National Research Council (NRC) report, *New Worlds, New Horizons in Astronomy and Astrophysics*, are searching for the first stars, galaxies, and black holes, as well as advancing the understanding of the fundamental physics of the universe, including determining the properties of dark energy.

"The search for exoplanets is one of the most exciting subjects in all of astronomy, and one of the most dynamic, with major new results emerging even as this report was being written," the survey notes, adding that since the first exoplanets were found in the early 1990s, discovery techniques have improved and the number of planets discovered has increased to about 500 currently known.

"This survey is recommending a program to explore the diversity and properties of planetary systems around other stars, and to prepare for the long-term goal of discovering and investigating nearby, habitable planets," the report states. "Generating a census of Earth-like or terrestrial planets is the essential first step toward determining whether our own home world is a commonplace or rare outcome of planet formation."

The report's top priority recommendation for a space mission is the $1.6 billion Wide Field Infrared Survey Telescope (WFIRST), a 1.5-meter wide-field-of-view near-infrared-imaging and low-resolution-spectroscopy telescope that could help find exoplanets as well as help determine the effect of dark energy on the evolution of the universe. The top-ranked medium-sized space-based project is the New Worlds Technology Development Program, which the report recommends that NASA initially fund at $4 million annually, "to lay the technical and scientific foundations for a future space imaging and spectroscopy mission."

Several astronomers on the committee emphasized that the search for habitable planets is an important and timely topic.

"The search for habitable planets, and the implicit expectation that that exercise will include the search for life, is something that is a profound activity conducted by human beings. And the fact that we might be able to accomplish this in this next decade—it shouldn't be a surprise that such a goal was raised high on that list," Neil deGrasse Tyson, member of the NRC Committee for a Decadal Survey