The Importance of Data Set Provenance for Science

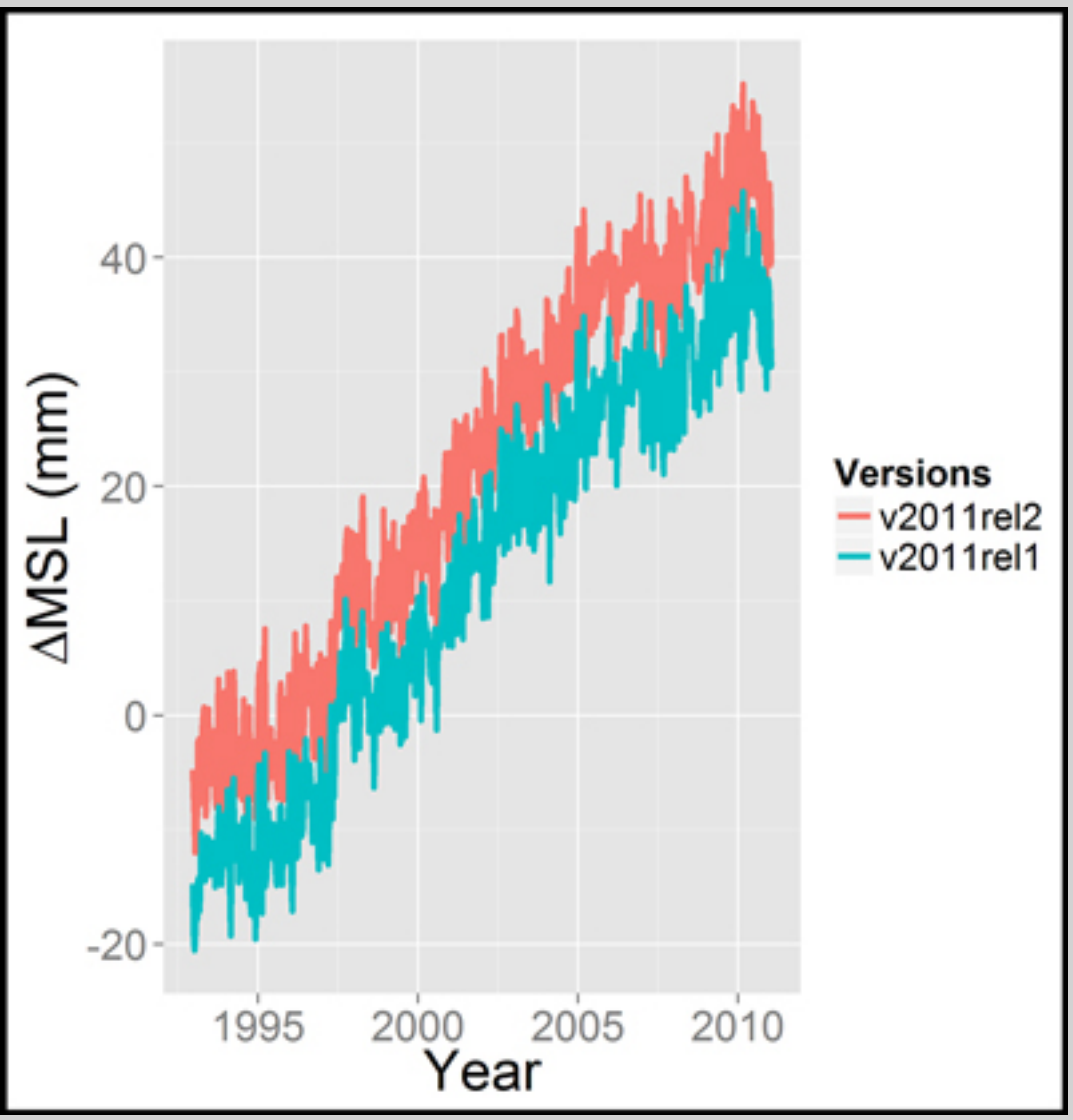
Data do not exist in a vacuum. To be useful, data must be accompanied by context on how they are captured, processed, analyzed, and validated and other information that enables interpretation and use.

Scientific research must rest on a solid foundation of data preservation, provenance, and contextual information. New technical and normative approaches are required to identify, capture, and track all details necessary to demonstrate data validity and to better ensure scientific reproducibility. Credit: Geralt, CC0 Public Domain

By Denise J. Hills, Robert R. Downs, Ruth Duerr, Justin C. Goldstein, Mark A. Parsons, and Hampapuram K. Ramapriyan 4 December 2015

Recently, an undocumented change was found in the long-term Antarctic sea ice record that

seemed to reverse an established trend [*Eisenman et al.*, 2014]. It turned out that Antarctic sea ice extent was not growing nearly as fast as thought. In fact, “much of this [past] expansion may be a spurious artifact of an error in the processing of the satellite observation” (p. 1289). This misunderstanding could have been avoided if the history—the provenance—of the data had been more clearly documented. Science requires transparency and verifiability, and scientists must always ask, Are these data trustworthy? From where did they originate? How were they generated and processed? What other data were used to calibrate, validate, and process these data? In other words, what are the provenance and context of the data?



(https://eos.org/opinions/the-importance-of-data-set-provenance-for-science/attachment/15-0241_hills_f01_web)

Fig. 1. This plot, constructed using data from *Nerem et al.* [2010], shows two releases of “the global mean sea level time series (season signals removed)” and could raise questions about the validity of the measurements if the provenance and context of the two releases were not documented.

Another example where provenance and context are crucial to scientific integrity is shown in Figure 1. The recommended citation for the data set shown is a static research article [*Nerem et al.*, 2010] that does not recognize the continual updating of the data values. Although the general

trend remains unaffected by revisions, the values are quite different. Sound research requires investigators to indicate exactly which version of a data set was used in a study, yet the scientific literature is rife with examples of these types of imprecise references and loose tracking of provenance.

Tackling the Provenance Problem

Data lie at the heart of these issues of transparency in published scientific research. Traditionally, it has been assumed that these issues were addressed through peer review and the self-correcting nature of science, but these methods are not proving to be fully effective in their current form. The literature is replete with cases of citing publications in lieu of data, not citing data set versions, and data set versions not being updated by data producers. Recent efforts to motivate authors to provide accurate data citations include the American Geophysical Union's (AGU) own data policy [e.g., *Hanson and van der Hilst*, 2014; *Hanson et al.*, 2015]. Data citation alone, however, does not solve the transparency issue. Full documentation of data set provenance and context is necessary.

Data citation alone does not solve the transparency issue. Full documentation of data set provenance and context is necessary.

Governments and other funders recognize this need for data preservation, provenance, and contextual information and are increasingly provide guidelines for ensuring the credibility of information. In 1998, the U.S. Global Change Research Program (USGCRP (http://www.globalchange.gov)), a confederation of 13 federal agencies active in global change research, convened a workshop to establish requirements for long-term preservation of Earth observation data.

Participants investigated positive and negative experiences involving the reuse and reanalysis of a variety of historical data products. This workshop produced a high-level list of information needed to "ensure [a data set's] usefulness to the scientist, and subsequent generations of scientists, with no prior knowledge of the data set or product" [*USGCRP*, 1999].

Also in 1998, NASA founded the Federation of Earth Science Information Partners (ESIP (http://esipfed.org/)) to engage a broader community of stakeholders in improving methods to make Earth science data easy to preserve, locate, access, and use. ESIP has begun to improve practices for enabling reuse of data by addressing issues of provenance and context in Earth science [e.g., *Duerr et al.*, 2011; *ESIP Stewardship Committee*, 2012; *Downs et al.*, 2015; *Mayernik et al.*, 2015] (see additional information on ESIP's role below).

Revitalizing the Effort

Despite the USGCRP workshop's early recognition that data and accompanying information merited preservation [*USGCRP*, 1999], for nearly a decade very little progress was made in implementing this understanding within relevant agencies or promoting it to the broader scientific community. ESIP now seeks to accelerate the actual implementation of the USGCRP workshop findings.

ESIP advocates new technical and normative approaches to identify, capture, and track all details necessary to demonstrate data validity and to better ensure scientific reproducibility, through the Provenance and Context Content Standard (PCCS) matrix, summarized in Table 1, developed by ESIP's Data Stewardship Committee. The PCCS matrix details the content required to describe provenance and context and identifies the major categories of data, metadata, and documentation that need to be preserved to increase trust in and understanding of research results.

Table 1. The Provenance and Context Content Standard (PCCS) Matrix[a]

Category	Content
Preflight/preoperations calibration	Instrument description; calibration information
Data set products	Raw data set; level 1 data set (e.g., unprocessed sensor data); level 2 data set (e.g., derived geophysical variables); level 3 data set (e.g., variables mapped on uniform scales); level 4 data set (e.g., model outputs); discovery metadata
Data set product documentation	Team members; product requirements; product development; processing history; product algorithms; quality assessment; references; user feedback
Data set calibration	Calibration method; in situ environment; platform history; calibration data; calibration software
Data set product software	Source code; output data set description; programming considerations; exceptions; test data sets; test plans; test results
Data set product algorithm inputs	Algorithm input documentation; algorithm input data sets
Data set product validation	Validation record; validation data sets
Data set software tools	Software readers and display tools

[a]The PCCS matrix details the content required to describe provenance and context and identifies the major categories of data, metadata, and documentation that must be preserved to increase trust in and understanding of research results (Federation of Earth Science Information Partners, Provenance Context Content 2011-06-08 (http://wiki.esipfed.org/index.php/Provenance_and_Context_Content_Standard)).

ESIP's PCCS has been adopted by NASA within the Earth Science Data Preservation Content Specification (https://earthdata.nasa.gov/standards/preservation-content-spec) (PCS) as a requirement for new Earth science missions. New missions using NASA's PCS include the Soil Moisture Active Passive (SMAP (http://smap.jpl.nasa.gov/)) and the Ice, Cloud, and land Elevation Satellite 2 (ICESAT2 (http://icesat.gsfc.nasa.gov/icesat2/)) missions. In the case of the older missions (those currently in operation or that have ended), NASA's PCS had not been included as a requirement but is used instead as a checklist for ensuring that as much of the relevant content as feasible is preserved.

Although NASA is not a formally designated "preservation agency" for Earth science data, it is essential for NASA to preserve all the data and associated content beyond the lives of NASA's missions to enable continued access to data and services for active scientific research. Furthermore, NASA must ensure that the data and associated content are preserved for transition to permanent archival agencies. In fact, all agencies should do the same.

The Future of Provenance Documentation

It is becoming more common for journals to require the availability of data supporting an article (e.g., AGU (http://publications.agu.org/author-resource-center/publication-policies/data-policy/) publications [*Hanson et al.*, 2015] and *Nature* (http://www.nature.com/authors/policies/availability.html) [*Nature Publishing Group*, 2013]). Availability of data may be broadly defined, however, and sometimes does not include the necessary documentation (provenance and context) that are necessary for reuse. This then raises the question of whether these journal policies truly ensure long-term data reuse.

Journals must require that data not only be available but also be reusable.

Journals must require that data not only be available but also be reusable. If, for example, the provenance and context of the two releases of the data sets plotted in Figure 1 were not well documented, many would question the validity of the measurements. With information captured through the PCCS (specifically, data set calibration), the reason for the shift in values would become clearer.

The Coalition for Publishing Data in the Earth and Space Sciences (COPDESS (http://www.copdess.org/)) released a "Statement of Commitment (http://www.copdess.org/statement-of-commitment/) from Earth and Space Science Publishers and Data Facilities" in early 2015 that in part states, "The major data repositories provide leading practices that should help guide the types of samples, data, metadata, and data processing descriptions that should be maintained, including information about derivations, processing, and uncertainty." This aligns with the goals of ESIP's PCCS. Signatories to COPDESS's statement of commitment (including AGU) are taking appropriate steps in this direction [e.g., *Hanson et al.*, 2015], but questions remain.

How do we truly solve the provenance problem? What needs to be documented? We'd like to do it consistently, but how do we communicate this information across disciplines? Can a single documentation standard work for all disciplines and data types, or do we need multiple domain-specific documentation standards? The community must consider the capture and preservation of provenance more seriously and begin employing appropriate practices routinely.

We believe the PCCS effort initiated by ESIP is an important first step toward solving the provenance problem. Although its current emphasis is on remote sensing data, the issue goes well beyond any one discipline or method. Contributions from other disciplines and initiatives can extend and improve such practices and generally increase trust in and understanding of research results.

ESIP continues to improve the PCCS by involving other communities within Earth and space sciences and by soliciting contributions from other national and international groups and the Research Data Alliance (https://rd-alliance.org/). We welcome broader collaboration and seek to work with others documenting data provenance and context to see if this PCCS can be extended more broadly.

Returning to our example from Antarctica, *Eisenman et al.* [2014, p. 1293] conclude, "These results illustrate the need for thorough documentation and version control in observational data sets. Ideally all observational data sets, especially those used widely and included in IPCC assessment reports, would have sufficient documentation of algorithms and algorithm changes for previous and current versions of the data to be independently replicated from the raw sensor data." ESIP's PCCS defines the information necessary to do that.

References

Downs, R. R., R. E. Duerr, D. J. Hills, and H. K. Ramapriyan (2015), Data stewardship in the Earth sciences,

D-Lib, *21*(7/8), doi:10.1045/july2015-downs.

Duerr, R. E., R. R. Downs, C. Tilmes, B. Barkstrom, W. C. Lenhardt, J. Glassy, L. E. Bermudez, and P. Slaughter (2011), On the utility of identification schemes for digital Earth science data: An assessment and recommendation, *Earth Sci. Inf.*, *4*, 139–160, doi:10.1007/s12145-0.

Eisenman, I., W. N. Meier, and J. R. Norris (2014), A spurious jump in the satellite record: Has Antarctic sea ice expansion been overestimated?, *Cryosphere*, *8*, 1289–1296, doi:10.5194/tc-8-1289-2014.

ESIP Stewardship Committee (2012), Data citation guidelines for data providers and archives, edited by M. A. Parsons et al., Fed. of Earth Sci. Inf. Partners. [Available at http://dx.doi.org/10.7269/P34F1NNJ (http://dx.doi.org/10.7269/P34F1NNJ).]

Hanson, B., and R. van der Hilst (2014), AGU's data policy: History and context, *Eos Trans. AGU, 95*(37), 337, doi:10.1002/2014EO370008.

Hanson, B., K. Lehnert, and J. Cutcher-Gershenfeld (2015), Committing to publishing data in the Earth and space sciences, *Eos, 96*, doi:10.1029/2015EO022207.

Mayernik, M. S., S. Callaghan, R. Leigh, J. Tedds, and S. Worley (2015), Peer review of datasets: When, why, and how, *Bull. Am. Meteorol. Soc.*, *96*, 191–201, doi:10.1175/BAMS-D-13-00083.1.

Nature Publishing Group (2013), Reducing our irreproducibility, *Nature, 496,* 398, doi:10.1038/496398a.

Nerem, R. S., D. P. Chambers, C. Choe, and G. T. Mitchum (2010), Estimating mean sea level change from the TOPEX and Jason altimeter missions, *Mar. Geod., 33*, 435–446, doi:10.1080/01490419.2010.491031.

U.S. Global Change Research Program (USGCRP) (1999), Global change science requirements for long-term archiving, workshop report, Washington, D.C., doi:10.7930/J0CZ353N.

—Denise J. Hills, Energy Investigations, Geological Survey of Alabama, Tuscaloosa; email: dhills@gsa.state.al.us (mailto:dhills@gsa.state.al.us); Robert R. Downs, Center for International Earth Science Information Network, Columbia University, Palisades, N.Y.; Ruth Duerr, National Snow and Ice Data Center, University of Colorado, Boulder; now at Ronin Institute for Independent Scholarship, Boulder, Colo.; Justin C. Goldstein, U.S. Global Change Research Program, Washington, D.C., and University Corporation for Atmospheric Research, Boulder, Colo.; Mark A. Parsons, Institute for Data Exploration and Applications, Rensselaer Polytechnic Institute, Troy, N.Y.; and Hampapuram K. Ramapriyan, NASA Goddard Space Flight Center, Greenbelt, Md., and Science Systems and Applications, Inc., Lanham, Md.

Citation: Hills, D. J., R. R. Downs, R. Duerr, J. C. Goldstein, M. A. Parsons, and H. K. Ramapriyan (2015), The importance of data set provenance for science, *Eos, 96,* doi:10.1029/2015EO040557. Published on 4

December 2015.

This article does not represent the opinion of AGU, ***Eos,*** or any of its affiliates. It is solely the opinion of the author.
