How to cite an Earth science data set

National Snow and Ice Data Center
University of Colorado Boulder

Mark Parsons and the Federation of Earth Science Information Partners Preservation and Stewardship Cluster

parsonsm@nsidc.org

Author

The name of the individual(s) or organization(s) whose intellectual work, such as a particular field experiment or algorithm, led to the creation of the data set. We prefer the term author over data creator because of its implied intellectual effort. The archive, in close collaboration with data providers, needs to determine who deserves to receive credit and accept responsibility for the data set and to define the appropriate level of aggregation for the data set. In some cases, the data authors may have also published a paper describing the data in detail. These sort of data papers should be encouraged, and both the paper and the data set should be cited when the data are used.

Title

The formal title of the data set not the project or a related publication. It is important for the data set to have an identity and title of its own.

Version

Careful versioning and documentation of version changes are essential for accurate citation. Data stewards need to track and clearly indicate precise versions as part of the citation. It may be appropriate to track major and minor versions.

Archive or Distributor

The organization that maintains and manages the release or distribution of the data set. There is often an implied responsibility for stewardship of the data set. This role is often considered that of a data "publisher," but we avoid that term because it may imply proprietary restrictions or unintended assertions of quality or peer-review.

Release Date

For a completed data set, the release date is simply the year of release. A more precise date can be used if needed. Note also if there is an update.
For an ongoing data set that is updated on a regular or continual basis, list the first year of release followed by the last update.

Cline, D., R. Armstrong, R. Davis, K. Elder, and G. Liston. 2002, Updated 2003. *CLPX-Ground: ISA snow depth transects and related measurements ver. 2.0*. Shapefiles from 2002. Edited by M. Parsons and M. J. Brodzik. National Snow and Ice Data Center. Data set accessed 2008-05-14 at http://dx.doi.org/10.5060/D4MW2F23z

Editor and Other Roles

In addition to the data author, there may be "editors" or other roles that could be included in the citation. Other minor roles could be credited elsewhere in data documentation.

Locator or Identifier

When data are available over the internet, it is necessary to include a persistent reference to the location of the data. Often this is through a standard URL, but the lack of persistence of URLs is a known problem. Assigning a unique and persistent locator offers a more consistent approach for managing location information. Any reasonably persistent location service such as DOIs, ARKs, Handles, or PURLs is acceptable. Scientific publishers, however, are most familiar with the DOI. Furthermore, the Web of Science, is building a new index of data sets, including DOIs or ARKs. This suggests that DOIs, and possibly ARKs, are more likely to be accepted by publishers.

Access Date and Time

Because data can be dynamic and changeable in ways that are not always reflected in release dates and versions, it is important to indicate when on-line data were accessed. This is in keeping with common citation practice for online documents and other resources. Depending on how frequently the data change, it may be necessary to include time as well as date of access.

Subset

It is necessary to enable "micro-citation" to refer to the specific data used—the exact files, granules, records, etc. (the page number in a literary citation). Ideally, an identifier or repeatable query ID would be assigned to a particular data subset, but that is not always available. Nevertheless, there is often a consistent structural form to how a data set is organized that can help users cite a specific subset. Data stewards should suggest how to reference subsets of their data. Subsets can often be identified by referring to a temporal and spatial range or possibly a file type.

The primary purpose of data citation is to aid scientific reproducibility through direct, unambiguous reference to the precise data used in a particular study. This may not always be possible, but this approach coupled with good version tracking, comprehensive documentation, and due diligence on the part of data stewards, can provide a useful and precise citation for the great majority of Earth science data most of the time.

Scientific honesty requires fair and precise data citation: http://bit.ly/data_citation.

Some notes on DOIs and other persistent identifiers

Locators vs. Identifiers

Identity and location are often confused or conflated. While one can often use an item's location to identify it or an item's identity to locate it, the concepts are distinct. This is easily conceived when we consider a human example. A name such as "John James Doe" (Office Manager at the FOO Data Center) is an identifier. An address such as "123 Main St. #201, Peoria, IL, 12345-1234, USA" is a locator.

The locator might work as an identifier, because you might find John in his office, but he may also have retired and there is a new Office Manager who plays the same role but is not the same person. Similarly, you may be able to locate John based on his name and title, but what happens if he is telecommuting this week and is in Poughkeepsie not Peoria? It is similar with digital objects. One might be able to identify a data set by its URL, for example, but there is no guarantee that what is at that URL today is the same as what was there yesterday,

Confusingly, a Digital Object Identifier (DOI) is a locator. It is a Handle based scheme whereby the steward of the digital object registers a location (typically a URL) for the object. There is no guarantee that the object at the registered location will remain unchanged.

While it is desirable to uniquely identify the cited object, it has proven extremely challenging to identify whether two data sets or data files are scientifically identical. Furthermore, Earth science data sets can be highly mutable. For know, we must rely on location information combined with other information such as author, title, and version to uniquely identify data used in a study.

Versioning and Locators

The key to using registered locators, such as DOIs, to unambiguously identify and locate data sets is through careful tracking and documentation of versions. Individual stewards and data centers will need to develop and follow their own practices, but here are some suggestions on how to handle different data set versions relative to an assigned locator.

- Track major_version.minor_version.
- Individual stewards need to determine which are major vs. minor versions and describe the nature and range of every version. Typically, something that affects the whole data set like a reprocessing would be considered a major version.
- Assign unique locators (DOIs) to major versions.
- Old locators for retired versions should be maintained and point to some web site that explains what happened to the old data.
- A new major version leads to the creation of a new collection-level metadata record that is distributed to appropriate registries. The older metadata record should remain with a pointer to the new version and with explanation of the status of the older version data.
- Major and minor version should be listed in the recommended citation.
- Minor versions should be explained in documentation, ideally in file-level metadata.
- Ongoing additions to an existing time series need not constitute a new version. This is one reason for capturing the date accessed when citing the data.
- Applying UUIDs, or other locators, to individual files upon ingest aids in tracking minor versions and historical citations.

9 December 2011

CIRES University of Colorado Boulder