Data Citation Guidelines for Data Providers and Archives **[1]**

Submitted by superadmin on Thu, 2012-03-01 11:36 **Event:** Winter Meeting 2012 [2]
Collaboration Area: Data Preservation [3]
DOI /EZid: doi:10.7269/P34F1NNJ
Technical Reports:
These guidelines are currently being revised and updated by the Data Stewardship Committee, who welcome feedback and participation in the effort. To get involved visit the Data Stewardship wiki (http://wiki.esipfed.org/index.php/Preservation_and_Stewardship [4])

Document Status

This document was approved by the ESIP Assembly 5 January 2012. The Data Stewardship Committee was charged with maintaining the Guidelines to ensure they remain functional and relevant.

The document was put out for review by all ESIP members 17 August 2011. As of 31 December 2011 some minor revisions have been made in response to feedback from the ESIP community and continually emerging guidance from the broader information science community.

Introduction and Summary

Data citation is an evolving but increasingly important scientific practice. We see several important purposes of data citation:

- To aid scientific reproducibility through direct, unambiguous reference to the precise data used in a particular study. (This is the paramount purpose and also the hardest to achieve).
- To provide fair credit for data creators or authors, data stewards, and other critical people in the data production and curation process.
- To ensure scientific transparency and reasonable accountability for authors and stewards.
- To aid in tracking the impact of data set and the associated data center through reference in scientific literature.
- To help data authors verify how their data are being used.
- To help future data users identify how others have used the data.

Nevertheless, data are rarely cited formally in practice. There are many reasons for this discrepancy, but part of the problem is that there is a lack of consistent recommendations on how to cite and reference a data set and on how to actually construct a proper data citation. We need consistent guidelines on how to create data citations for Earth science data. Then data stewards need to work closely with data providers and science teams to define citation content and provide clear instructions to users on how data sets should be cited.

Current recommendations by ESIP members for citing their data range from casual acknowledgement within the text of a paper to formal and specific citations within the references section of the paper with unique and persistent digital identifiers. At the same time, various international organizations have been working to develop formal guidelines for data citation. Examples of these include:

- International Polar Year How to Cite a Data Set [5]
- DataCite [6]—a consortium of libraries and related organizations working to define a citation approach around DOIs. They have defined a DataCite Metadata Scheme for the Publication and Citation of Research Data, Version 2.2, July 2011 [7] describing myriad elements that could potentially be included in a citation.

- A new CODATA Task Group [8] on Data Citation Standards and Practices in collaboration with the International Council for Scientific and Technical Information and the National Academy of Sciences.
- DataVerse Network Project [9]—an approach from the social science community using a Handle locator and “Universal Numerical Fingerprint” as a unique identifier.
- GEOSS Science and Technology Committee is also working on guidelines by building from the IPY guidelines. See their draft [10]
- A "Working Level Guide" from the Digital Curation Center. How to Cite Datasets and Link to Publications [11]. This is the most useful guide and was only released in October 2011, after the ESIP Guidelines were written.

The ESIP Preservation and Stewardship cluster has examined these and other current approaches and has found that they are generally compatible and useful, but they do not entirely meet all the purposes of Earth science data citation. Indeed, we believe it is currently impossible to fully satisfy the requirement of scientific reproducibility in all situations. We do believe, however, that applying a reasonably rigorous 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.

In general, data sets should be cited like books. (Used here is the author-date style described in "Chicago Manual of Style, 15th Edition" [12].) When users cite data, they need to use the style dictated by their publishers, but by providing an example, data stewards can give users all the important elements that should be included in their citations of data sets. Data stewards need to work closely with data providers and science teams to develop the actual content of the citation to ensure that each data citation unambiguously refers to the data that was utilized for a particular work.

The core required elements of a citation are

- Author(s)--the people or organizations responsible for the intellectual work to develop the data set. The data creators.
- Release Date--when the particular version of the data set was first made available for use (and potential citation) by others.
- Title--the formal title of the data set
- Version--the precise version of the data used. Careful version tracking is critical to accurate citation.
- Archive and/or Distributor--the organization distributing or caring for the data, ideally over the long term.
- Locator/Identifier--this could be a URL but ideally it should be a persistent service, such as a DOI, Handle or ARK, that resolves to the current location of the data in question.
- 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.

Additional fields can be added as necessary to credit other people and institutions, etc. Additionally, it is important to provide a scheme for users to indicate the precise subset of data that were used. This could be the temporal and spatial range of the data, the types of files used, a specific query id, or other ways of describing how the data were subsetted.

An example citation:

> 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. Edited by M. Parsons and M. J. Brodzik. National Snow and Ice Data Center. Data set accessed 2008-05-14 athttp [13] ://dx.doi.org/10.5060/D4MW2F23z [13]

.

Detailed Citation Content

These citation guidelines help data stewards define and maintain precise, persistent citations for data they manage. The guidelines build from the IPY Guidelines and are compatible with the DataCite Metadata Scheme for the Publication and Citation of Research Data, Version 2.2, July 2011, and How to Cite Datasets and Link to Publications. DCC How-to Guides. The Cluster will continue to work with the CODATA Task Group, the GEOSS Science and Technology Committee, DataCite, and others to ensure that our guidelines remain compatible with broader community practice. The approach defined here could be viewed as an extension or detailed profile of existing approaches.

The citation should include the following elements as appropriate. Mappings to the "Citation Information" section of the Federal Geographic Data Committee (FGDC) Content Standard for Digital Geospatial Metadata (CSDGM) ("FGDC-STD-001-1998" [14]) and the DataCite Metadata Scheme Ver. 2.2 (July 2011) [7] are included. Each of the elements are described in more detail below.

Citation Element	FGDC CSDGM field	DataCite Metadata Scheme ID and Property
Author or Creator*	idinfo > citation > citeinfo > "origin"	2 Creator*
Release Date*	idinfo > citation > citeinfo > "pubdate" and sometimes "othercit"	5 PublicationYear*
Title*	idinfo > citation > citeinfo > "title" and possibly "edition"	3 Title*
Version*		15 Version
Archive and/or Distributor*	idinfo > citation > citeinfo > "publish"	4 Publisher*
Locator, Identifier, or Distribution Medium*	idinfo > citation > citeinfo > "othercit" or "onlink"	1 Identifier*
Access Date and Time*	not applicable	8 Date
Subset Used	not applicable	12 RelatedIdentifier DataCite recommends obtaining an identifier for any subset that needs to be cited as well as an identifier for the larger whole. See further discussion below.
Editor or Other Important Role	idinfo > citation > citeinfo > "origin"	7 Contributor
Publication Place	idinfo > citation > citeinfo > "pubplace"	17 Description
Distributor, Associate Archive, or other Institutional Role	idinfo > citation > citeinfo > "othercit"	7 Contributor or possibly 4 Publisher
Data Within a Larger Work	idinfo > citation > citeinfo > "othercit" or "lworkcit"	12 RelatedIdentifier

*Mandatory (if applicable)

Mandatory Content

Author

This is 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. This is sometimes called the data creator. We prefer the term author because of its implied intellectual effort and its conventional use in citing traditional works. The archive, in close collaboration with data providers, needs to determine who deserves to receive credit and accept responsibility for the data set. Similarly, the archive needs to work closely with the provider to define the appropriate level of aggregation for the data set.

In some cases, the data set authors may have also published a paper describing the data in great 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.

In addition to the data author, there may be "editors" or other roles that could be included in the citation if they made significant intellectual contributions. Other roles should be credited elsewhere

in data documentation or metadata.

> Doe, J. and R. Roe. 2001. The FOO Data Set. The FOO Data Center.http://dx.doi.org/10.xxxx/notfoo.547983 [15]. Accessed 1 May 2011.

A particular group or small organization may sometimes be the author, but one should be as specific as possible in accountability and crediting intellectual contribution. Naming an entire funding agency as an author is usually too vague.

> The FOO Working Group. 2001. The FOO Data Set. The FOO Data Center.http://dx.doi.org/10.xxxx/notfoo.547983 [15]. Accessed 1 May 2011.

If the data set is a collection of several smaller, independent data sets, the individual data sets would have their own specific citations with author, but the whole collection would not have an author. The collection would likely have an editor or compiler, though. See "Data Within a Larger Work"

> Doe, J. (compiler) 2001. The FOO Collection. The FOO Data Center.http://dx.doi.org/10.xxxx/notfoo.547983 [15]. Accessed 1 May 2011.

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 to indicate when exactly the data became available and citable.

> Doe, J. and R. Roe. 2001. The FOO Data Set. The FOO Data Center.http://dx.doi.org/10.xxxx/notfoo.547983 [15]. Accessed 1 May 2011.

If detailed versioning information is lacking for a data set it may be appropriate to try and capture when updates occurred. For a data set that is updated infrequently or on an irregular basis, list the first year of released followed by "updated" with the current update information.

> Doe, J. and R. Roe. 2001, updated 2005. The FOO Occasionally Updated Data Set. The FOO Data Center. http://dx.doi.org/10.xxxx/notfoo.547983 [15]. Accessed 1 May 2011.

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. Updates could occur annually or more frequently.

> Doe, J. and R. Roe. 2001, updated daily. The FOO Time Series Data Set. The FOO Data Center.http://dx.doi.org/10.xxxx/notfoo.547983 [15]. Accessed 1 May 2011.

A note on updates vs. new versions:

Ongoing updates to a time series do change the content of the data set, but they do not typically constitute a new version or edition of a data set. New versions typically reflect changes in sampling protocols, algorithms, quality control processes, etc. Both a new version and an update may be reflected in the release date. The version number should also be included. See also the "Note on Versioning and Locators."

> Doe, J. and R. Roe. 2001, updated daily. The FOO Time Series Data Set. Version 3.2. The FOO Data Center. http://dx.doi.org/10.xxxx/notfoo.547983 [15]. Accessed 1 May 2011.

Title

This is the formal title of the data set. It may also include version or edition information, but should be carefully controlled. A better alternative is to track version information independent of the title. Note this is the 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.

> Doe, J. and R. Roe. 2001. The FOO Data Set. The FOO Data Center.http://dx.doi.org/10.xxxx/notfoo.547983 [15]. Accessed 1 May 2011.

Archive and/or Distributor

This is 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. These issues are beyond the intent of data citation. DataCite describes this role well: "The entity that holds, archives, publishes, prints, distributes, releases, issues, or produces the resource. This property will be used to formulate the citation, so consider the prominence of the role." This may be an appropriate place to recognize a major sponsor of the data.

> Doe, J. and R. Roe. 2001. The FOO Data Set. The FOO Funding Agency Data Center.http://dx.doi.org/10.xxxx/notfoo.547983 [15]. Accessed 1 May 2011.

Version

Careful versioning and documentation of version changes are central to enabling accurate citation. Data stewards need to track and clearly indicate precise versions as part of the citation for any version greater than 1. It may be appropriate to track major and minor versions. See the note below on versions and locators.

> Doe, J. and R. Roe. 2001. The FOO Data Set. Version 2.3. The FOO Data Center.http://dx.doi.org/10.xxxx/notfoo.547983 [15]. Accessed 1 May 2011.

Locator, Identifier, or Distribution Medium

If there is one fixed medium, indicate it. For example, CD-ROM, DVD.

> Doe, J. and R. Roe. 2001. The FOO Data Set. The FOO Data Center. CD-ROM.

More typically, data are available over the internet or through multiple digital media options. Then it is necessary to include a persistent reference to the location of the data. Often this is through a standard URL, but the mutability and lack of persistence of URLs is a known problem. The assignment of a unique and persistent locator offers a more consistent approach for managing current location information. Any reasonably persistent location service such as DOIs, ARKs, Handles, PURLs etc. is acceptable. Scientific publishers, however, are most familiar with the DOI. Furthermore, Thomson Reuters, who manages the Web of Science, is building a new index of data sets, yet to be

named. They plan to register data sets in this index including DOIs or ARKs. Also, data sets that are cited by articles in the Web of Science will also show up in Web of Science, so there is an incentive for authors to cite data sets. All this suggests that while any persistent locator can be used in a citation, DOIs and possibly ARKs are more likely to be accepted by publishers.

Note DOIs, ARKs, and Handles are useful to locate full data sets or collections by pointing to a data landing page that describes and provides access to the data. Other locators and identifiers may be more appropriate for locating individual records or files. Best practice is that the suffix of the identifier does not include a reference to the archive in case the data are moved from the original location where the persistent identifier was assigned initially. Also to aid human usability the locator should include an easy access protocol, i.e. an http address. For DOIs this takes the formhttp [16]://dx.doi.org/ [16].

> Doe, J. and R. Roe. 2001. The FOO Data Set. Version 2.3. The FOO Data Center.http://dx.doi.org/10.xxxx/notfoo.547983 [15]. Accessed 1 May 2011.

See additional discussion in the "Subset Used" section as well as the "Note on Locators and Versions" and "Note on Locators versus Identifiers".

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.

> Doe, J. and R. Roe. 2001. The FOO Data Set. Version 2.3. The FOO Data Center.http://dx.doi.org/10.xxxx/notfoo.547983 [15]. Accessed 1 May 2011.

Suggested Content as Needed

Subset Used

This may be the most challenging aspect of data citation. It is necessary to enable "micro-citation" or the ability to refer to the specific data used--the exact files, granules, records, etc. An example in a traditional context would be quoting a certain passage in a book, where one then references a specific page number in the citation. Alternatively, one might make reference to the "structural index" of a canonical text (e.g. book, chapter, and verse in the King James Bible). Unfortunately data sets typically lack page numbers or canonical versions. 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. With Earth science data, subsets can often be identified by referring to a temporal and spatial range.

> Doe, J. and R. Roe. 2001, updated daily. The FOO Gridded Time Series Data Set. Version 3.2. Oct. 2007- Sep. 2008, 84°N, 75°W; 44°N, 10°W. The FOO Data Center.http://dx.doi.org/10.xxxx/notfoo.547983 [15]. Accessed 1 May 2011.

Sometimes, the data may be packaged in different sub-collections or representations or "Archive Information Units (`AIUs)," which can be referenced.

> Doe, J. and R. Roe. 2001. The FOO Data Set. Version 2.0 shapefiles. The FOO Data Center.http://dx.doi.org/10.xxxx/notfoo.547983 [15]. Accessed 1 May 2011.

Ideally, we would not need to rely on the structure of the data set. Specific identifiers or locators could be used for individual AIUs, or if the archive tracks provenance well, it may be possible to assign identifiers to a particular query.

Editor, Compiler, or other important role

Occasionally, there are other people besides the authors who played an important role in the creation or development of a data set. Often these people can be characterized as editors or compilers, but other roles might also be identified. An editor is the person or team who is responsible for creating a value-added and possibly quality-controlled product from the data. In cases where there is minimal scientific or technical input, yet still substantial effort in compiling the product, the person may be more correctly cited as a compiler. Editors and compilers may often be responsible for a larger work that includes multiple data sets from different authors. Occasionally, there may be both a compiler and editor as well as other roles. Myriad other roles should be credited elsewhere in data documentation or metadata.

> Doe, J. 2001. The FOO Data Set. Version 2.0 R. Roe (ed.) The FOO Data Center.http://dx.doi.org/10.xxxx/notfoo.547983 [15]. Accessed 1 May 2011.

When there is an editor but no author, the editor is listed first.

Archive or Distributor Place

This is the city, state (when necessary), and country of the archive or distributor. It is primarily to be consistent with historical literary citation, but it can be useful to provide some context and to help access data delivered on media.

> Doe, J. and R. Roe. 2001. The FOO Data Set. Peoria, IL, USA: The FOO Data Center. CD-ROM.

Distributor, Associate Archive, or other Institutional Role

Just like there may be multiple human roles associated with a data set, there may also be multiple institutional roles. For example, sometimes there are different distributors or associate archives. This could also be a place to recognize the sponsor of the data collection.

> Doe, J. 2001. The FOO Data Set. Version 2.0 The FOO Data Center. Distributed by the FOO Distribution Center. http://dx.doi.org/10.xxxx/notfoo.547983 [15]. Accessed 1 May 2011.

> Doe, J. 2001. The FOO Data Set. Version 2.0 The FOO Data Center in association with the FUU Data Center. http://dx.doi.org/10.xxxx/notfoo.547983 [15]. Accessed 1 May 2011.

Data Within a Larger Work

A particular data set may be part of a compilation, in which case it is appropriate to cite the data set somewhat like a chapter in an edited volume.

> Bockheim, J. 2003. "University of Wisconsin Antarctic Soils Database". In International Permafrost Association Standing Committee on Data Information and Communication (comp.). 2003. Circumpolar Active-Layer Permafrost System, Version 2.0. Edited by M.

> Parsons and T. Zhang. Boulder, CO: National Snow and Ice Data Center/World Data Center for Glaciology. CD-ROM.

Increasingly, publishers are allowing data supplements to be published along with peer-reviewed research papers. When using the data supplement one need only cite the parent reference. For example, when using the data at"http://dx.doi.org/10.1594/PANGAEA.476007" [17], the following reference is appropriate.

> Stein, Ruediger, Bettina Boucsein, and Hanno Meyer. 2006. "Anoxia and high primary production in the Paleogene central Arctic Ocean: first detailed records from Lomonosov Ridge." Geophysical Research Letters, 33:L18606. http://dx.doi.org/10.1029/2006GL026776 [18].

Note on 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 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. Consider a continually updated data time series, for example. Indeed, the advantage is to separate location information from other information about the resource. The location is not part of the resource description so that location can be managed independently, thereby enabling URL changes, migrations to other archives, etc.

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. At this point, we must rely on location information combined with other information such as author, title, and version to uniquely identify data used in a study.

The locator should not point directly to the data but rather to a "landing page" with a description of the data, versioning info, and mechanisms to access the data. This is not the same as a "data paper". The landing page is a living document that can be updated as data are moved or changed. This page is also a good place to provide more attribution for sponsors and other people important in the creation and stewardship of the data. Ideally, the landing page would be both human and machine readable, but the infrastructure for machine readable landing pages is still evolving. The DCC Guidelines [11] discuss issue of manual and automatic use of citations in more detail.

See also the "Note on Versioning and Locators" and the section on "Subset Used."

Note on Versioning and Locators

The key to making registered locators, such as DOIs, ARKS, or Handles, work unambiguously to 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 there are some

suggestions on how to handle different data set versions relative to an assigned locator. The DCC Guidelines [11] suggest that DOIs be assigned to different data snapshots taken at regular intervals or as needed. This would work well for infrequently changed data sets. DCC also suggests a "time slice" approach where "the citable entity becomes the set of updates made to a dataset during a particular time period rather than the full dataset itself (e.g. the 2008 data from a series running since 1950)." This may be workable in some situations, but we find it unwieldy for the regularly updated time series common in Earth science.

The National Snow and Ice Data Center did an initial study exploring many different kinds of data sets and production patterns and suggest the following more nuanced approach.

- Track major_version.minor_version.[archive_version]. Archive version is only used internally to track any change in the archive. Major and minor version are exposed publicly and reflect actual scientific change in the data.
- Individual stewards need to determine which are major vs. minor versions and describe the nature and file/record range of every version. Typically, something that affects the whole data set like a reprocessing would be considered a major version.
- Assign unique locators to major versions.
- Old locators for retired versions should be maintained and point to some appropriate web site that explains what happened to the old data if they were not archived.
- 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 identifiers, to individual files upon ingest aids in tracking minor versions and historical citations.

Source URL: http://commons.esipfed.org/node/308

Links:
[1] http://commons.esipfed.org/node/308
[2] http://commons.esipfed.org/event/winter-meeting-2012
[3] http://commons.esipfed.org/collaboration-area/data-preservation
[4] http://wiki.esipfed.org/index.php/Preservation_and_Stewardship
[5] http://ipydis.org/data/citations.html
[6] http://datacite.org/
[7] http://schema.datacite.org/meta/kernel-2.2/doc/DataCite-MetadataKernel_v2.2.pdf
[8] http://www.codata.org/taskgroups/TGdatacitation/
[9] http://thedata.org/citation
[10] http://www.geo-tasks.org/st0902/a21/geoss_citation_standard_ST0902_draft_v2.0.doc
[11] http://www.dcc.ac.uk/resources/how-guides/cite-datasets
[12] http://www.chicagomanualofstyle.org/home.html
[13] http://dx.doi.org/10.5060/D4MW2F23z
[14] http://www.fgdc.gov/metadata/csdgm/
[15] http://dx.doi.org/10.xxxx/notfoo.547983
[16] http://dx.doi.org/
[17] http://doi.pangaea.de/10.1594/PANGAEA.476007
[18] http://dx.doi.org/10.1029/2006GL026776