GROUP ON EARTH OBSERVATIONS

GEOSS Citation Standard

DRAFT for Discussion by ST-09-02

Date: March 20, 2011

Preamble

Data citation is an open issue currently discussed by a number of international organizations, including the International Polar Year (IPY) and the Federation of Earth Science Information Partners (ESIP). Data citation is far more complicated than citation of scientific publication since Data sets differ in many aspects from standard scientific publications, which makes turns data citation far more complex than citation of scientific publications. For example, data sets generally are not locatable and attributable in the same way as scientific publications. Data sets often are not static, and they are mostly not peer-reviewed. All these issues need to be addressed before a comprehensive GEOSS Citation Standard can be developed. However, the IPY and ESIP have developed simple guidelines for data citation that can serve as a temporary GEOSS Citation Standard until a more comprehensive standard is available. The document provided here describes this first version of the GEOSS Citation Standard. The proposed standard focuses on data citation since citation of scientific publication are not considered an open issue. This initial standard includes guidelines for data citation that agree to a very large extent with the IPY and ESIP guidelines, with modifications only made where the specific application to GEOSS required this. This draft is passed on to the STC for further handling.

The development of a more comprehensive GEOSS Citation Standard is the next step. This process will have to address questions such as data identification including version tracking, quality control, and attribution. This process needs to be coordinated with the GEO working groups in charge of developing the GEOSS Common Infrastructure (e.g., ADC, SIF, DSTF, GCI-CT), which needs to be able to support the requirements of a future GEOSS Citation Standard.

Note: *The draft has been extracted from a draft deliverable of the EGIDA Project prepared by Ian McCallum, and minor modifications and additions were made by ST-09-02 Task team members and discussed during ST-09-02 Telecons.*

INTRODUCTION

> "Currently a person who publishes really good data is given less credit than someone who publishes a bad paper" (AGU Town Hall Meeting, 2009)

At present, GEOSS is lacking a data citation standard. This is symptomatic of the entire field of geo-spatial data, with few exceptions. A wide variety of organizations and projects support the concept of data citation (e.g., IPY, PANGAEA, NASA DAACS, USGS, NOAA National Data Centers), however no standard has emerged. Users accessing the GEO-Portal are not obliged to cite data from the GEO-Portal – if at all, citation requirements come from the individual data providers. This naturally leads to an at-best non-standard form of data citation or in the worst case, no data citation at all. Poor or no data citation is bad for both the scientist who has produced the data and receives no credit, and for the scientist who wishes to obtain data (i.e., from a scientific publication) but can not locate them owing to a lack of information.

In the scientific community, recognition and renown are important currencies. In order to increase the attractiveness of GEO and GEOSS for scientists, their contributions must be visibly acknowledged when others use it to their benefit. A key element for acknowledgment is a proper GEOSS citation standard, which not only covers published documents but also ensures that data used to create new products is acknowledged and appropriately credited. Therefore, a GEOSS data citation standard needs to be developed and implemented. To be effective, the GEOSS citation standard must aim for compliance/compatibility with international specifications concerning data citation.

GEOSS has a unique opportunity/responsibility to the scientific community that relies on spatial data to help establish such a data citation standard. By encouraging proper citation of data sets, data providers and publishers receive appropriate credit for their efforts, the perception of data management as a discipline improves, and it is easier to track the use and impact of the data (International Polar Year). The scientific community should recognize the professional value of such activities by endorsing the concept of publication of data, to be credited and cited like the products of any other scientific activity, and encouraging peer-review of such publications (AGU Position Statement).

The goal of this document is to provide a data citation standard for all GEOSS related datasets consistent with recommendation of relevant international organizations. The guidelines presented herein were adapted from the International Polar Year (IPY) Data Information Service, which was a synthesis of the different approaches agreed to by many international data centers. These organizations acknowledge that the current guidelines still leave a number of issues uncovered (e.g., unique data object identifiers, attribution, quality information). It is recommended that GEO addresses these issues together with the Federation of Earth Science Information Partners (ESIP) and other relevant organizations in a next version of the GEOSS Citation Standard. The development of this more comprehensive Standard will require coordination with the ADC to ensure that the GEOSS Common Infrastructure (GCI) can enable the implementation of the future Standard.

DATA CITATION

To recognize the valuable role of data providers (and scientists who collect or prepare data) and to facilitate repeatability of IPY experiments in keeping with the scientific method, users of IPY data must formally acknowledge data authors (contributors) and sources. Where possible, this acknowledgment should take the form of a formal citation, such as when citing a book or journal article. Journals should require the formal citation of data used in articles they publish. Where formal citation is not possible, such as with some medical and social science data, ethical policies for data collection and data use are encouraged, building upon existing models such as Article 8(j) of the 1992 Convention on Biological Diversity (IPY Data Policy)

The following guidelines were developed by the IPY. These guidelines were adapted from internal guidelines used by the National Snow and Ice Data Center, which has encouraged formal data citation for more than a decade.

In general, data sets should be cited like books. Used here is the author-date system described in Chicago Manual of Style, 15th Edition. When users cite data, they need to use the style dictated by their publishers, but by providing an example, data publishers can give users all the important elements they should include in their citations of data sets. *So long as the mandatory fields are complete, the final publisher style is just semantics.*

An example of a citation in the author-date system is:
Algire, G. H., and F. T. Legallais. 1948. Biology of Melanomas. ed. R. W. Miner. New York: New York Academy of Sciences.

As seen in this example, the elements of the citation in order are:

1. Author(s),
2. Date,
3. Title,
4. Editor,
5. Place of Publication,
6. Publisher.

All these elements are common in data set citations, but other elements, as described below, are commonly used as well. Data publishers (e.g. data centers) have a responsibility to work with data providers and science teams to develop the actual content of the citation.

The citation should include the following elements as appropriate. Although this is shown as a literary citation, most of the elements are captured in standard metadata. The "Citation Information" section of the Federal Geographic Data Committee (FGDC) Content Standard for Digital Geospatial Metadata (CSDGM) (FGDC-STD-001-1998) is given as reference:

Author or Investigator

This is the individual(s) whose intellectual work, such as a particular field experiment or algorithm, led to the creation of the data set.

Oberbauer, S. 2000. Ecosystem carbon fluxes, Toolik Lake, Alaska 1995. Boulder, Colorado USA: National Snow and Ice Data Center. Data set accessed 2008-05-14 at http://nsidc.org/data/arcss006.html.

A particular group or organization may sometimes be the author.

Arctic Climatology Project. 2000. Environmental Working Group Arctic meteorology and climate atlas. Edited by F. Fetterer and V. Radionov. Boulder, Colorado USA: National Snow and Ice Data Center. CD-ROM.

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.

Cross, M. compiler. 1997. Greenland summit ice cores. Boulder, Colorado USA: National Snow and Ice Data Center in association with the World Data Center A for Paleoclimatology at NOAA-NGDC, and the Institute of Arctic and Alpine Research. CD-ROM.

Publication Date

For a completed data set, the publication date is simply the year of release.

Helmig, D. 2004 Vertical Boundary Layer Profiles for Ozone and Meteorological Parameters at Summit, Greenland, 2000. Boulder, Colorado USA: National Snow and Ice Data Center. Data set accessed 2008-05-14 at http://nsidc.org/data/arcss100.html.

For a data set that is updated infrequently or on an irregular basis, list the first year of publication followed by "updated" with the current update information. This is appropriate when the title or version of the data set does not change, the data are simply updated.

Osterkamp, T. 1999, updated 2001 Daily air and active layer temperatures from permafrost observatories in Alaska, 1986-2001. Boulder, Colorado USA: National Snow and Ice Data Center. Data set accessed 2008-05-14 at http://nsidc.org/data/arcss106.html.

For an ongoing data set that is updated on a regular or continual basis, list the first year of publication followed by the last update. Updates could occur annually or more frequently.

Maslanik, J. and J. Stroeve. 1999, updated quarterly. DMSP SSM/I daily polar gridded brightness temperatures, Jan. Boulder, Colorado USA: National Snow and Ice Data Center. Data set accessed 2008-05-14 at http://nsidc.org/data/nsidc-0001.html.

Hall, Dorothy K., George A. Riggs, and Vincent V. Salomonson. 2007, updated daily. MODIS/Aqua

Sea Ice Extent 5-Min L2 swath 1km V005, Oct. 2007–Apr. 2008. Boulder, Colorado USA: National Snow and Ice Data Center. Data set accessed 2008-05-14 at http://nsidc.org/data/myd29v5.html.

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 publication date. The title should indicate the new version.

Hall, Dorothy K., George A. Riggs, and Vincent V. Salomonson. 2006, updated daily. MODIS/Terra snow cover Extent 5-Min L2 swath 1km V005, Oct. 2007–Apr. 2008. Boulder, Colorado USA: National Snow and Ice Data Center. Data set accessed 2008-05-14 at http://nsidc.org/data/myd29v5.html.

Hall, D. K., G. A. Riggs, and V. V. Salomonson. 2000, updated daily. MODIS/Terra snow cover 5-Min L2 swath 500m V004, Oct. 2007–Apr. 2008. Boulder, Colorado USA: National Snow and Ice Data Center. Data set accessed 2008-05-14 at http://nsidc.org/data/myd29v5.html.

If a particular version of a time series is discontinued, it is appropriate to indicate when the final update occurred.

Hall, D. K., G. A. Riggs, and V. V. Salomonson. 2000, updated 2002. MODIS/Terra snow cover 5-Min L2 swath 500m V003, Jan. 2001–Apr. 2001. Boulder, Colorado USA: National Snow and Ice Data Center. Data set accessed 2008-05-14 at http://nsidc.org/data/myd29v5.html.

Title

This is the formal title of the data set. It may also include version or edition information.

Liu, H., K. Jezek, B. Li, and Z. Zhao. 2001. Radarsat Antarctic Mapping Project digital elevation model version 2. Boulder, CO: National Snow and Ice Data Center. Data set accessed 2008-05-14 at http://nsidc.org/data/nsidc-0082.html.

Dates Used

For time series, especially continually updated time series, indicate which dates of data were used. Note this is distinct from the publication date.

Hall, Dorothy K., George A. Riggs, and Vincent V. Salomonson. 2006, updated daily. MODIS/Terra snow cover Extent 5-Min L2 swath 1km V005, Oct. 2007–Apr. 2008. Boulder, Colorado USA: National Snow and Ice Data Center. Data set accessed 2008-05-14 at http://nsidc.org/data/myd29v5.html.

Editor or Compiler

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 an individual author's data set. Occasionally, there may be both a compiler and editor. Some products will have neither.

Armstrong, R., J. Francis, J. Key, J. Maslanik, T. Scambos, and A. Schweiger. 1998. Polar Pathfinder sampler: Combined AVHRR, SMMR-SSM/I, and TOVS time series and full-resolution samples. Compiled by S. Khalsa. Boulder, CO, USA: National Snow and Ice Data Center. CD-ROM.

Cline, D., R. Armstrong, R. Davis, K. Elder, and G. Liston. 2002, Updated July 2004. CLPX-Ground: ISA snow pit measurements. Edited by M. 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.

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.

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

Publication Place

This is the city, state (when necessary), and country of the publisher.

Cavalieri, D., C. Parkinson, P. Gloersen, and H. J. Zwally. 1996, updated 2006. Sea ice concentrations from Nimbus-7 SMMR and DMSP SSM/I passive microwave data, March 2002–Sept. 2003. Boulder, Colorado USA: National Snow and Ice Data Center. Data set accessed 2008-05-14 at http://nsidc.org/data/nsidc-0051.html.

Publisher

The publisher is whoever published the data set. A publisher often has an implied responsibility for stewardship of the data set. This is usually a data center and is written immediately after the place.

Cavalieri, D., C. Parkinson, P. Gloersen, and H. J. Zwally. 1996, updated 2006. Sea ice concentrations from Nimbus-7 SMMR and DMSP SSM/I passive microwave data, March 2002–Sept. 2003. Boulder, Colorado USA: National Snow and Ice Data Center. Data set accessed 2008-05-14 at http://nsidc.org/data/nsidc-0051.html.

Distributor or Associate Publisher

This field should be used only when it differs from the publisher, i.e. rarely. Its listing should be written in the same manner as that of publisher. Sometimes NSIDC acts as a simple distributor; sometimes we are an associate publisher; sometimes others are associate publishers.

Environmental Working Group. 2000. Environmental Working Group: Joint U.S.-Russian Arctic sea ice atlas. Ann Arbor, MI: Environmental Research Institute of Michigan; distributed by the National Snow and Ice Data Center. CD-ROM.

Cross, M. compiler. 1997. Greenland summit ice cores. Boulder, CO: National Snow and Ice Data Center in association with the World Data Center A for Paleoclimatology at NOAA-NGDC, and the Institute of Arctic and Alpine Research. CD-ROM.

Distribution Medium and Location

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

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.

If data are available over the internet or through multiple digital media options it is best to include a reference to the location of the data. Often this is through a standard URL.

Cavalieri, D., C. Parkinson, P. Gloersen, and H. J. Zwally. 1996, updated 2006. Sea ice concentrations from Nimbus-7 SMMR and DMSP SSM/I passive microwave data, March 2002–Sept. 2003. Boulder, Colorado USA: National Snow and Ice Data Center. Data set accessed 2008-05-14 at http://nsidc.org/data/nsidc-0051.html.

Ideally, a persistent identifier such as a Digital Object Identifier should be used.

König-Langlo, Gert and Hatwig Gernandt. 2006. Compilation of radiosonde data from the Antarctic Georg-Forster station of the German Democratic Republic from 1985 to 1992. Bremerhaven, Germany: Alfred Wegener Institute for Polar and Marine Research Data set accessed 2008-05-22. doi:10.1594/PANGAEA.547983

Access Date

Because data can be dynamic and changeable in ways that are not always reflected in publication dates and versions, it is important to indicate when on-line data were accessed. It is not necessary to indicate an access date for a fixed medium like a DVD.

Cavalieri, D., C. Parkinson, P. Gloersen, and H. J. Zwally. 1996, updated 2006. Sea ice concentrations from Nimbus-7 SMMR and DMSP SSM/I passive microwave data, March 2002–Sept. 2003. Boulder, Colorado USA: National Snow and Ice Data Center. Data set accessed 2008-05-14 at http://nsidc.org/data/nsidc-0051.html.

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 doi:10.1594/PANGAEA.476007, 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. doi:10.1029/2006GL026776.

IMPLEMENTATION

It is recommended that the implementation of the GEOSS Citation Standard should proceed along the following lines:

1. Use IPY data citation Guidelines to implement a first version of the GEOSS Citation Standard.
2. Address issues such as the unique identification of digital objects, versions of datasets, data stability, attribution, etc. in a next version of the GEOSS Citation Standard.
3. Implement the concepts in the GEOSS Common Infrastructure.
4. Lobby publishers to require data citation.

1. Owing to the fact that the basic IPY Guidelines generally reflect the current consensus on data citation and are relatively simple in their structure, it is proposed that GEOSS applies these guidelines in their current form (as described in the previous section). It is understood that these guidelines are dynamic and would be modified or improved over time to suite the specific needs of GEOSS.

In addition, it would be beneficial if GEOSS could automatically generate a standard data citation for each GEOSS dataset as per literature citation databases - i.e. export the citation in Text, RefWorks, RIS format (e.g. EndNote), BibTeX, comma separated. This would likely facilitate improved referencing. It is worth noting that e.g. EndNote does not contain a default reference type for data – thus the recommendation to cite data using the format of a book avoids this problem.

2. Options for data identification include Digital Object Identifiers (DOI), which are used to provide current information, including where data (or information about data) can be found on the Internet. Information about a digital object may change over time, including where to find it, but its DOI name will not change. Unfortunately, at this point no one identifier scheme has emerged to meet all the needs of scientific data publications (Parsons et al. 2010). This needs to be explored further. A UUID (Universal Unique Identifier) is a 128-bit number used to uniquely identify some object or entity on the Internet. A UUID addresses the need to uniquely and unambiguously identify a particular data set or

subset no matter which copy a user has (Parsons et al. 2010). This seems promising in terms of uniquely identifying data, but should also be explored further.

3. The GCI will have to be designed to provide the attributes required in a future GEOSS Citation Standard. This will require a coordination between the development of this standard and the implementation of, e.g., object identifiers, UUID, and other relevant attributes in the GCI.

4. A long term goal of GEOSS should be to actively lobby publishers to require data citation. Ultimately journal editors and reviewers need to be more rigorous in demanding that authors accurately cite the data they use in there research (Parsons et al. 2010). Along these lines of thinking, one journal already exists which is dedicated to publishing data: Earth System Science Data. This journal aims to establish a new subject of publication: to publish data according to the conventional fashion of publishing articles, applying the established principles of quality assessment through peer-review to datasets.

In summary, it is proposed here that GEOSS proceed as a first step with establishing a data citation standard along the lines of the IPY citation standard (Parsons 2010). This is in part possible as the standards are assumed to be dynamic, allowing for improvements over time. The majority of open issues surrounding data citation, e.g, persistent identifiers, should be dealt with in the near future.

RELEVANT LINKS

http://libraries.mit.edu/guides/subjects/data/access/citing.html

http://www.dlib.org/dlib/march07/altman/03altman.html

http://ands.org.au/guides/data-citation-awareness.html

http://wiki.esipfed.org/index.php/Interagency_Data_Stewardship/Citations

REFERENCES

AGU Town Hall meeting, 2009. See http://wiki.esipfed.org/index.php/Interagency_Data_Stewardship/2009AGUTownHall.

Parsons, M.A., 2010. Presentation at ST-09-02 Task Team Meeting, Rome, Italy, October 30, 2010. See http://www.geo-tasks.org/st0902/meetings and http://www.geo-tasks.org/st0902/meetings/20100930_Rome/parsons_citation_geoss.pdf.

Parsons, M.A., Duerr, R., & Minster, J.B., 2010. Data citation and peer review. *Eos , Trans. Am. Geophys. Union*, ***91***, 297-298