GEOSS Data Citation Guidelines: Version 2.0

October 2, 2012

Edited by I. McCallum[1], H.-P. Plag[2], S. Fritz[1]

[1] International Institute for Applied Systems Analysis, ESM Program, Laxenburg, Austria
[2] Nevada Bureau of Mines and Geology and Seismological Laboratory, University of Nevada, Reno, Mail Stop 178, Reno, NV 89557, USA

Preamble

An initial Draft Global Earth Observation System of Systems (GEOSS) Citation Standard Version 1.0 (EGIDA, 2011) was compiled in the frame of the EGIDA Project (http://www.egida-project.eu/).

The Group on Earth Observations (GEO) endorsed at the Plenary in November 2011 the following Position Statement on Data Citation, where GEO:

- recognizes the importance of citing the source of data in publications, derived products and services;
- recognizes that data citation is a critical acknowledgement for data providers and creates an important incentive for the publication, documentation, registering and provision of data; and
- particularly encourages users of the GEOSS Common Infrastructure to cite data found or accessed through the GEOSS Common Infrastructure. A recommendation was made to the STC to implement the Draft GEOSS Citation Standard V1.0 as a testbed for data citation.

The former Science & Technology Committee has asked ID-03 to develop GEOSS Data Citation Guidelines - Version 2, that address many of the currently open issues; are coordinated with the development of the GEOSS Common Infrastructure (GCI); and are consistent with the development of guidelines for data citation by other relevant organizations. The ID-03 Task Team has developed this Version 2 in close cooperation with the EGIDA project.

The development of the Guidelines has been coordinated with the GEO working groups in charge of developing the GEOSS Common Infrastructure (e.g., ADC, SIF, DSTF, GCI-CT), which has to take into account and support the requirements of the GEOSS Data Citation Guidelines in terms of meta data, citation recommendations, and citation indexing. Further coordination was necessary with similar activities initiated by other groups within GEO, such as the Data Sharing Task Force, in order to avoid duplication of work and ensure consistency across GEOSS.

It will be important to continue the discussion of the Data Citation Guidelines with the global community engaged in the international discussion led by among others, ESIP, DataCite, and ICSU's CODATA. Future revisions of the GEOSS Data Citation Guidelines will be informed by this international discussion.

This document outlines initial steps to implement the GEOSS Citation Guidelines within the GCI. This approach is referred to as a testbed and will allow GEOSS to develop and dynamically test data citation that will provide feedback which is valuable not only to GEOSS, but to the wider S&T community.

Table of Contents

1. Giving Credit to Data Producers

An important incentive for scientists and researchers is the recognition and renown given to them in citations of their work. While citation guidelines are well developed for the use of scientific papers published by others, very few guidelines are available for the citation of data made available by others. Increasingly, citation of the source of data is also requested in the context of socially relevant topics, such as climate change and its potential impacts. Providing means for data citation would be a strong incentive for data sharing. Geo-referenced data are crucial for addressing many of the burning societal problems and to support related interdisciplinary research. The lack of a widely accepted method for giving credit to those who make their data freely available and for tracking the use of data throughout their life-cycle hampers data sharing. Furthermore, only clear and transparent data citation allows other scientists to obtain the identical data to replicate findings or to perform further research.

2. Towards Data Citation

The need for data citation guidelines is increasingly acknowledged and addressed by leading scientific organizations. A number of organizations and projects have started to address the concept of data citation (Table 1). Data repositories are also expanding, offering longterm storage and access to archived data (e.g., ICSU's World Data System, PANGAEA, Dryad, Dataverse). Furthermore, dedicated data journals are appearing, which aim to help researchers specifically publish data (e.g., Earth System Science Data). Several proposals for guidelines have emerged and a better understanding of the many issues at hand is evolving, but to date, no standard has been accepted. One very positive development was the recent launch by Thomson Reuters of a Data Citation IndexTM for discovering global datasets (http://www.reuters.com/article/2012/06/22/idUS109861+22-Jun-2012+HUG20120622). An attempt will be made to align the GEOSS Data Citation Guidelines with the requirements of the Data Citation IndexTM.

Table 1: Selected organizations contributing to the development of data citation guidelines.

Organization	Contribution
International Polar Year (IPY)	The IPY developed a set of rules for data citation, which have been used both within and outside of the IPY context
ICSU/CODATA	Acknowledges the need for robust data citation and identified key issues to be addressed by citation rules. See http://www.codata.org/taskgroups/TGdatacitation/index.html
DATA-PASS	The Data Preservation Alliance for Social Sciences addresses data citation in the context of social sciences. See http://www.data-pass.org/
ESIP Federation	Adapted the IPY rules and modified them into data citation rules for the ESIP Federation
DATA-CITE	Aims to support data access and re-usability through citation rules. See http://www.datacite.org/
U.S. National Academies	The Board on Research Data and Information of the U.S. National Academies is preparing a report on data citation. See http://sites.nationalacademies.org/PGA/brdi/PGA_063656
EGIDA	The project provided important input to the GEOSS Citation Standard V1.0. See http://www.egida-project.eu/

Data citation is far more complicated than citation of scientific publications as additional factors must be considered (e.g., versions, etc.). However data citation can adopt many of the well established rules for

scientific publications. Nonetheless, there is consensus that some of the issues will only emerge when initial data citation guidelines are implemented and put to a test. Elements of a data citation are described in Table 2.

Table 2. Elements of a data citation: Ball, A. & Duke, M. (2012): www.dcc.ac.uk/resources/how-guides

Field	Description
Author*	The creator of the dataset
Publication Date*	Whichever is the later of: the date the dataset was made available, the date all quality assurance procedures were completed, and the date the embargo period (if applicable) expired
Title*	As well as the name of the cited resource itself, this may also include the name of a facility and the titles of the top collection and main parent sub-collection (if any) of which the dataset is a part
Edition	The level or stage of processing of the data, indicating how raw or refined the dataset is
Version	A number increased when the data changes, as the result of adding more data points or re-running a derivation process, for example
Feature name and Uniform Resource Identifier (URI)	The name of an ISO 19101:2002 'feature' (e.g. GridSeries, ProfileSeries) and the URI identifying its standard definition, used to pick out a subset of the data.
Resource type	Examples: 'database', 'dataset'.
Publisher	The organisation either hosting the data or performing quality assurance
Unique numeric fingerprint	A cryptographic hash of the data, used to ensure no changes have occurred since the citation
Identifier	An identifier for the data, according to a persistent scheme
Location*	A persistent Uniform Resource Locator (URL) from which the dataset is available. Some identifier schemes provide these via an identifier resolver service

* Minimum Requirements according to Ball & Duke 2011

3. A Testbed for Data Citation: GEOSS

The GEOSS Data Citation Guidelines will be implemented in the CGI. As a first step, a testbed for data citation will be created. Currently, users of the GEO-Portal are not obliged or encouraged to cite data accessed through GEOSS – if at all, citation requirements come from the individual data providers. The testbed implementation of the GEOSS Data Citation Guidelines will rectify this situation; increase the attractiveness of GEO and GEOSS for scientists by making their contributions visibly acknowledged; and help to identify issues not covered by the Guidelines.

The implementation of the Guidelines in the GCI is illustrated in Fig. 2. The process of implementing the Guidelines and iteratively improving them is led by the GEO Work Plan Task ID-03 "Science and Technology in GEOSS", and coordinated with the GEO working groups in charge of developing the GEOSS Common Infrastructure; other relevant GEO components (e.g., the GEO Data Sharing Task Force); and is aligned with the emerging international specifications concerning data citation. The experience with the testbed will be infused into the international discussion on data citation. An important open issue is the question of how citations can be tracked, so that citation statistics can be made available to data authors. Most likely, the new Data Citation IndexTM introduced by Thomson Reuters will rectify this problem.

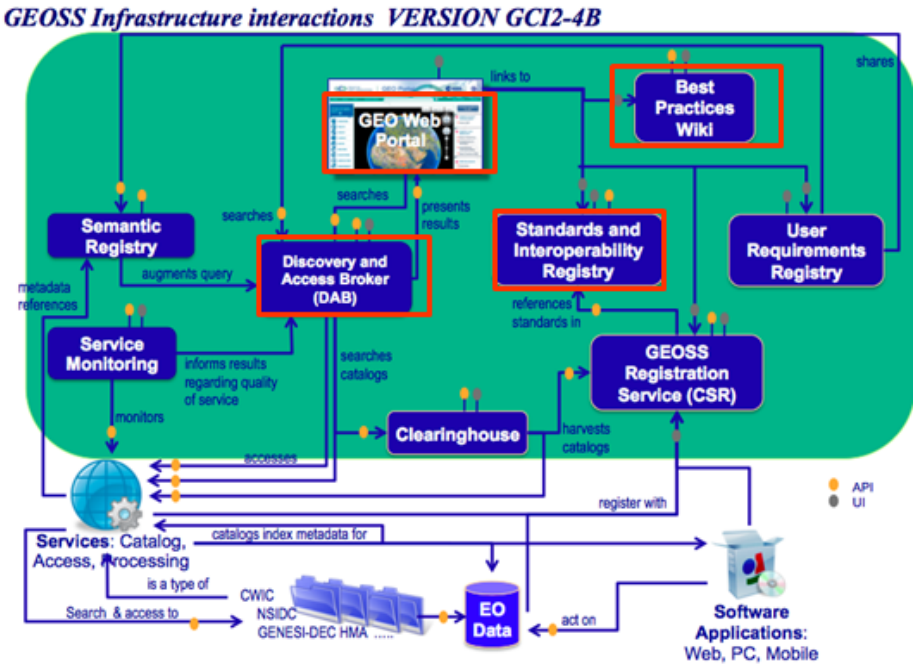

Fig. 2: Overview of the structural components of the GEOSS Common Infrastructure. Implementation of the Data Citation Guidelines will impact several components: the Guidelines will be registered in the Standards and Interoperability Registry; the Guidelines will be included in the Best Practices Wiki for data citation (http://wiki.ieee-earth.org/Best_Practices/GEOSS_Transverse_Areas/Data_and_Architecture/How_to_Cite_a_Data_Set); a request for data citation will be included in the GEO Portal, and the Discovery and Access Broker (DAB) will transmit information required for proper data citation.

3.1 Data Citation Guidelines

The following Data Citation Guidelines have been adapted for implementation in the data citation testbed in the GCI. They are based upon the Draft GEOSS Data Citation Guidelines Version 1.0 (EGIDA, 2011). They have been simplified to aid in the initial creation of the testbed taking into account the current capabilities and limitations of the GCI, such as the metadata currently available within the GEO Portal. Table 3 shows the five basic requirements of the data citation, along with the matching field contained in the GEO Portal metadata database and a short description about the field.

Using the GEO Portal Equivalent Fields listed in Table 3, it should be possible to automatically harvest the required information to create a GEOSS data citation recommendation directly from the metadata

coming from the data provider. This is assuming that the chosen fields are not blank. The GEOSS data citation recommendation would then be provided to anyone who accessed the metadata or the data itself. This could initially take the form of a simple text string which could be copied directly, but would ideally be presented in various standard bibliography formats for easy ingestion into bibliography software.

A scan of the metadata available via the Geo Portal for the described Fields in Table 3, would suggest that in many instances the minimum elements of a data citation are present. If however, any of the first three fields are blank (i.e. Author, Date, and Title), no citation should be created. In this case, a message could appear to anyone accessing either the metadata or the data itself, that owing to a lack of metadata, no citation can be provided, and that the user is requested to contact the data provider directly.

Table 3. Minimum elements of a data citation and their corresponding GEO Portal metadata database equivalent field (modified from Ball, A. & Duke, M. (2012)). An example of a data citation based upon the guidelines in Table 3, appears below. The fields should appear in the order they are listed in the table, each separated with a period.

Field	GEO Portal Metadata Database Equivalent Field	Description
Author or Investigator	*Individual Name*	The creator of the dataset. The individual(s) whose intellectual work, such as a particular field experiment or algorithm, led to the creation of the data set. A particular group or organization may sometimes be the author.
Publication Date	*Date stamp*	Whichever is the later of: the date the dataset was made available, the date all quality assurance procedures were completed, and the date the embargo period (if applicable) expired.
Title	*Title*	This is the formal title of the data set. It may also include version or edition information. As well as the name of the cited resource itself, this may also include the name of a facility and the titles of the top collection and main parent sub-collection (if any) of which the dataset is a part.
Publisher	*Organization*	The organization either hosting the data or performing quality assurance. A publisher often has an implied responsibility for stewardship of the data set. This is usually a data center.
Location	*Linkage*	A persistent URL from which the dataset is available. Some identifier schemes provide these via an

	identifier resolver service.

Hence, a data citation as per the GEO Portal Metadata equivalent fields should look as such:

Individual Name. Date stamp. Title. Organization. Linkage.

An example of a data citation based upon the guidelines in Table 3 is as follows:

Fritz, S., et al. 2011. Percent cropland coverage for sub-Saharan Africa. PANGAEA. doi:10.1594/PANGAEA.761139

4. Ongoing Issues to Be Addressed

ICSU CODATA provides a comprehensive overview of issues that need to be considered in developing data citation guidelines (http://www.codata.org/taskgroups/TGdatacitation/index.html):

A. Technical

- Interoperability and Facilitation of Re-use
- Citation Formats
- Metadata
- Database Versioning

B. Scientific

- Different disciplines may have disparate needs for granularity;
- Differences among disciplines that need to be addressed distinctly?

C. Institutional

- What are the roles of the respective stakeholders?
- What are the implications for these stakeholders?
- Does this vary by discipline?

D. Financial

- Lot of granularity can be cost-prohibitive.
- Must be accessible and its costs affordable by all necessary user communities.

E. Sustainability

F. Persistent Identifiers

- e.g., use of the DOI (Digital Object Identifier) System.
- Use of DOI names for datasets is promoted by the not-for-profit DataCite consortium, which has registered over 600,000 datasets;
- However, significant differences between data and documents, that may make some aspects of the DOI system less attractive.

G. Legal Issues/Intellectual Property Rights

- Any registry system must accommodate emerging intellectual property rights mechanisms, e.g. Creative Commons and Science Commons licensing, as well as traditional copyright law.

H. Socio-cultural and Community Norms

- Develop a common basis and community of practice for recognizing and rewarding data work;

I. Other Issues will arise …

GEO Task ID-03 in coordination with groups inside and outside of GEO will continue to address these issues in revisions of the GEOSS Data Citation Guidelines.

5. References

Ball, A. & Duke, M. (2012). How to Cite Datasets and Link to Publications'. DCC How-to Guides. Edinburgh: Digital Curation Centre. Available online: http://www.dcc.ac.uk/resources/how-guides

EGIDA, 2011. D.3.1 Draft GEOSS Citation Standard. http://www.egida-project.eu/images/documents/proposalforageosscitation.pdf