

# DATA CITATION STANDARD: A MEANS TO SUPPORT DATA SHARING, ATTRIBUTION AND TRACEABILITY

I. McCallum<sup>1</sup>, H.-P. Plag<sup>2</sup>, and S. Fritz<sup>1</sup>

<sup>1</sup> International Institute for Applied Systems Analysis, ESM Program, Laxenburg, Austria

<sup>2</sup> Nevada Bureau of Mines and Geology and Seismological Laboratory, University of Nevada, Reno, Mail Stop 178, Reno, NV 89557, USA

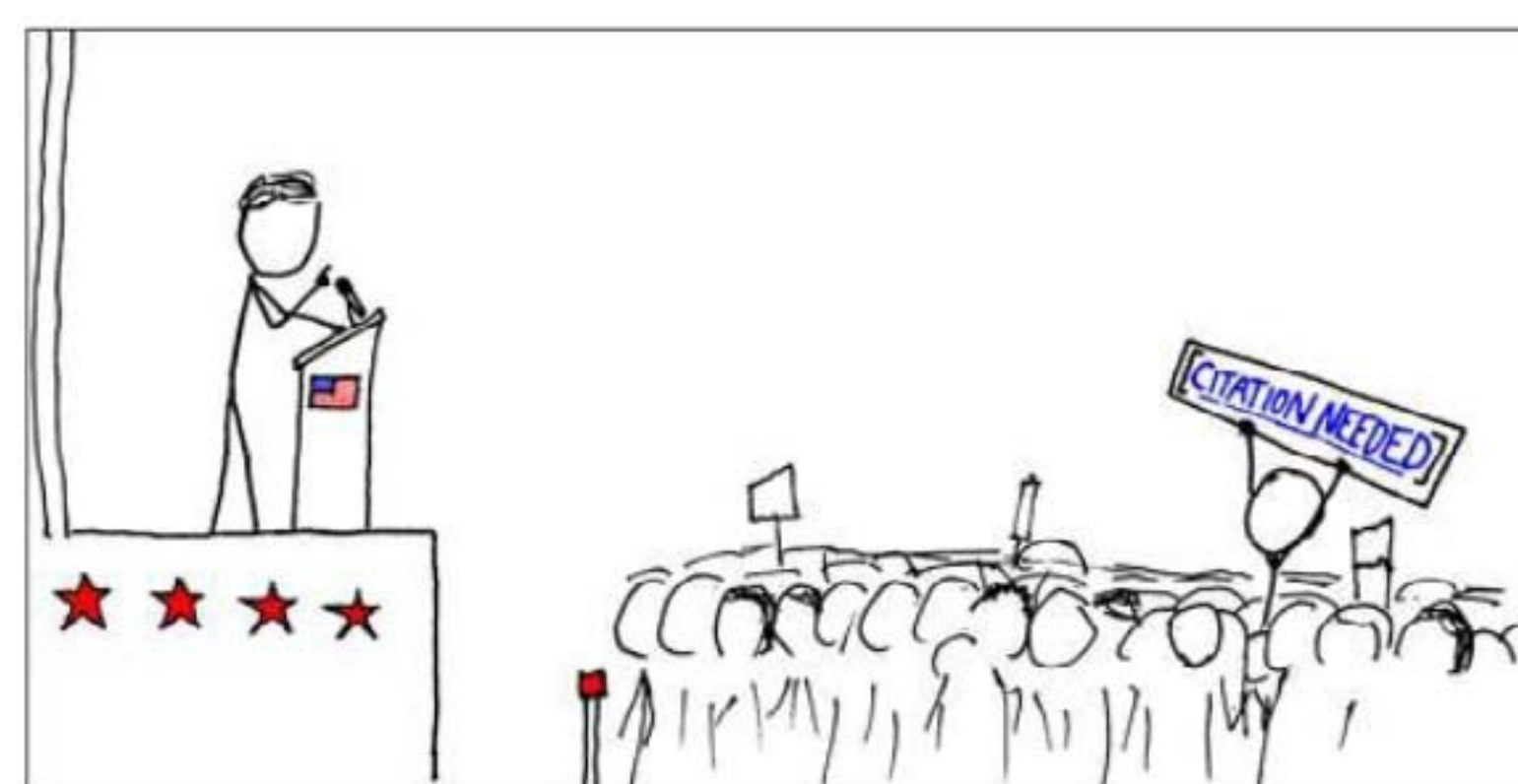


Fig. 1: There is a need for citation of data sources, not only for scientific papers. From John Wilbanks, Creative Commons, DataCite Annual Meeting 2011, Berkeley, August 25, 2011, see <http://www.slideshare.net/wilbanks/datacite-wilbanks>

## 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 rules are well developed for the use of papers published by others, very little rules 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 (Fig. 1). 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 for further research.

## TOWARDS DATA CITATION

The need for data citation rules is increasingly acknowledged and addressed by leading scientific organizations. A number of organizations and projects have started to address the concept of data citation (Tab. 1). Data repositories are also expanding, offering long term storage and access to archived data (e.g. 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. Data citation is far more complicated than citation of scientific publications as additional factors must be considered (i.e. 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 rules are implemented and put to a test. Elements of a data citation are described in Tab. 2.

## GEOSS DATA CITATION STANDARD

The Global Earth Observation System of Systems (GEOSS) developed by the Group on Earth Observations (GEO) aims to provide comprehensible Earth observations (EOs) in support of decision making in a wide range of societal benefit areas. The nine interdependent Societal Benefit Areas (SBAs) addressed by GEO require an interdisciplinary scientific approach, and scientific interpretation of the EOs provided by GEOSS is necessary in order to derive actionable information. A strong engagement of science and technology (S&T) communities in both the development and use of GEOSS is necessary to address the complex issues of the global integrated Earth system; improve interoperability between global observing, modeling, and information systems; facilitate data sharing, archiving, dissemination, and reanalysis; optimize the recording of observations, assimilation of data into models, and generation of data products; enhance the value of observations from individual observing systems through their integration in the SBAs; and harmonize well-calibrated, highly accurate, stable, sustained in-situ and satellite observations of the same variable recorded by different sensors and different agencies. The former Science and Technology Committee (STC) of GEO with support of the EGIDA Project has developed a draft standard. This draft is based on guidelines developed by international groups. This draft standard is under constant improvement as new aspects are addressed.

“Citation” is about acknowledging the use of another's ideas.  
“Data citation” is about acknowledging the use of another's data.

Tab. 1: Selected organizations contributing to the development of data citation rules.

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 <a href="http://www.codata.org/taskgroups/TGdatacitation/index.html">http://www.codata.org/taskgroups/TGdatacitation/index.html</a>
DATA-PASS	The Data Preservation Alliance for Social Sciences addresses data citation in the context of social sciences. See <a href="http://www.data-pass.org/">http://www.data-pass.org/</a>
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 <a href="http://www.datacite.org/">http://www.datacite.org/</a>
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 <a href="http://sites.nationalacademies.org/PGA/brdi/PGA_063656">http://sites.nationalacademies.org/PGA/brdi/PGA_063656</a>
EGIDA	The project provided important input to the GEOSS Citation Standard V1.0. See <a href="http://www.egida-project.eu/">http://www.egida-project.eu/</a>

## A TESTBED FOR DATA CITATION: GEOSS

The GEO Plenary supports a testbed implementation of the draft GEOSS Data Citation Standard. 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 draft GEOSS Data Citation Standard 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 standard. The implementation of the standard in the GEOSS Common Infrastructure (GCI) is illustrated in Fig. 2. The process of implementing the draft and iteratively improving it 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.

Tab. 2. Elements of a data citation: Ball, A. & Duke, M. 2011:

[www.dcc.ac.uk/resources/how-guides](http://www.dcc.ac.uk/resources/how-guides)

Field	Description
<b>Author</b>	The creator of the dataset
<b>Publication Date</b>	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
<b>Title</b>	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
<b>Edition</b>	The level or stage of processing of the data, indicating how raw or refined the dataset is
<b>Version</b>	A number increased when the data changes, as the result of adding more data points or re-running a derivation process, for example
<b>Feature name and URI</b>	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.
<b>Resource type</b>	Examples: 'database', 'dataset'.
<b>Publisher</b>	The organisation either hosting the data or performing quality assurance.
<b>Unique numeric fingerprint</b>	A cryptographic hash of the data, used to ensure no changes have occurred since the citation
<b>Identifier</b>	An identifier for the data, according to a persistent scheme
<b>Location</b>	A persistent URL from which the dataset is available. Some identifier schemes provide these via an identifier resolver service

Items in bold reflect minimum requirements

GEOSS Infrastructure interactions VERSION GCI2-4B

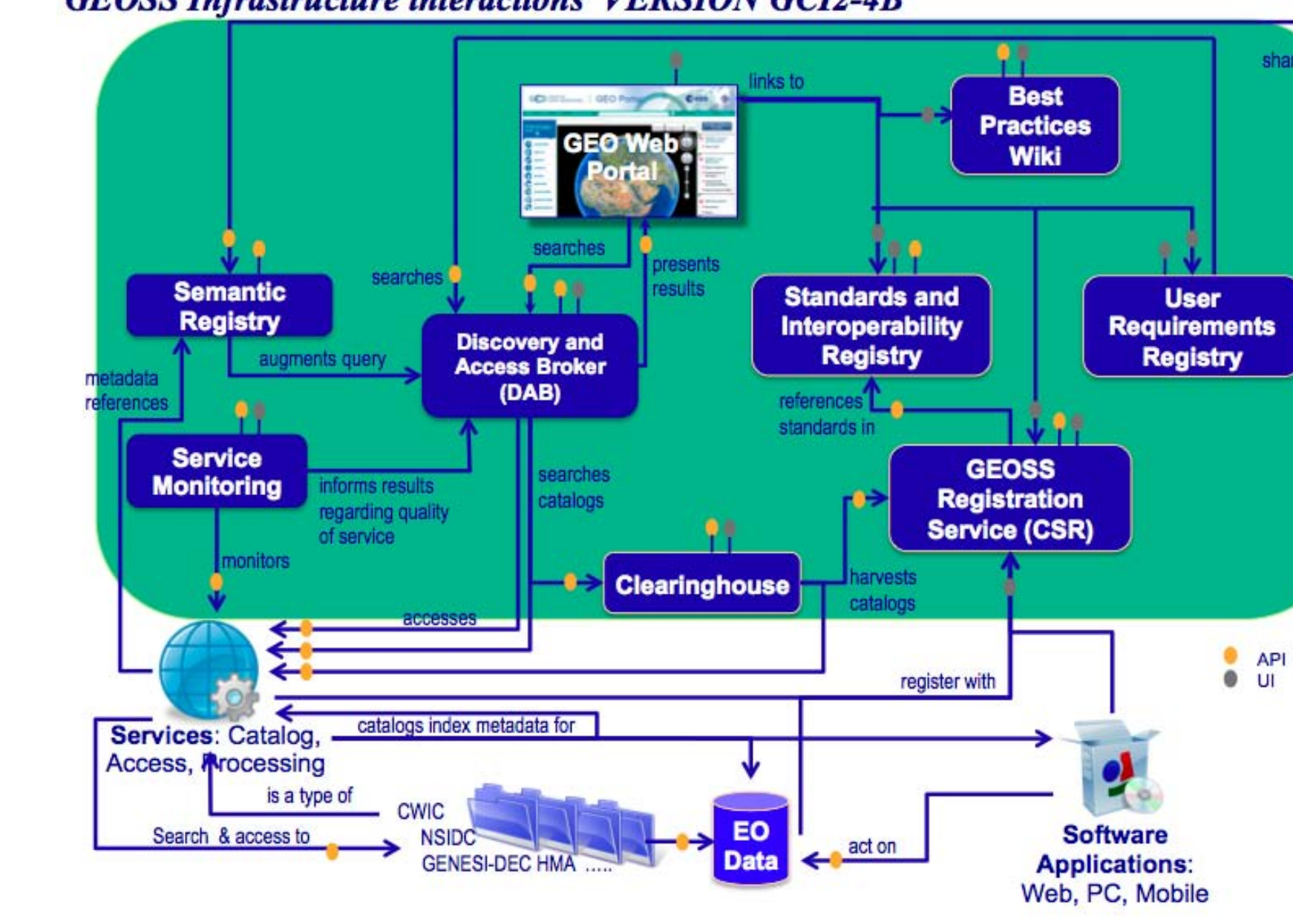


Fig. 2: Overview of the structural components of the GEOSS Common Infrastructure. Implementation of a data citation standard and request will impact several components: the standard will be registered in the Standards and Interoperability Registry; a Best Practice in data citation will be developed and included in the Best Practices Wiki; a request for data citation will be included in the GEO Portal, and the Data Discovery and Access Broker will transmit information required for proper data citation.

Can we solve data citation before we understand how to store and preserve, how to forget, how to name, how to find or how to distinguish data?

## ISSUES TO BE ADDRESSED

ICSU CODATA provides a comprehensive overview of issues considered in developing data citation rules <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 will consider most of the above issues that are currently not taken into account by V1.0 of the GEOSS data citation standard in V2.0, currently being developed by the GEO Task ID-03.