GEOSS Data Quality Guidelines

Approved by DDQ Subgroup on 12 June 2013
Approved by DSWG on 19 June 2013

These Guidelines are offered by the GEO Data Sharing Working Group (DSWG) to foster understanding of the importance of data quality information associated with GEOSS data resources[1]. It is essential for users to understand the quality of data sets and to combine this quality information with other metadata components in order to determine the appropriateness, or fitness, of these data sets for the users' applications and/or purposes. Thorough documentation of data characteristics and data quality will support a wider scope of use than the purpose for which the data resources were originally acquired.

GEOSS Data Providers should

1. For instruments used to collect Earth observations identify, establish and exploit a "reference standard" as a means of evaluating performance or compliance for a particular activity. Ideally this should be undertaken as part of an internationally harmonized Quality Assurance procedure (1). For many data providers, who use commercial instruments, this may mean expressing the means by which the instruments are calibrated by the manufacturer, including the schedule followed for recalibrations.

2. Provide data resources lineage, also called provenance, recording the data collection and/or generation, including auxiliary information used, in detail sufficient to allow reproducibility. (2) (5)

3. Provide information about the quality of the data resources, and any quality assurance procedures followed in producing the data. (2) (5)

4. Specify what purposes the data resource was collected or created for, or is known to be useful for, and any known caveats. (2) (5)

5. Provide data quality assessments in a manner that ensures the quality information is supplied alongside the data resource itself, such as via associated metadata or documentation tightly coupled to the data.

6. Provide quality control information at product level, taking into account instrument characteristics, environmental characteristics at the time the observation is made, and any algorithmic and ancillary data characteristics.

7. Address the multiple dimensions of quality. The purpose is not to judge or rank data resources, but to describe the characteristics needed to be known in order for the user to decide whether he/she should use them. (4) (5)

Data Quality/Fit-for-Purpose Examples:
The term "fit-for-purpose" is a rich term, conveying that data quality is in the eye of the beholder. What makes a data set valuable for one purpose may be a detriment for another purpose. Data quality information should help the multitude of Earth observation users to choose the best data for their own purpose. Categories of users include Earth science researchers, applications for the GEO societal benefit areas, disaster response teams, policy makers, and so on. It should be noted that some of the characteristics defining fitness-for-purpose are explicit data quality measurements (e.g. positional accuracy or completeness), while others are standard metadata elements that are not traditionally seen as denoting quality (e.g. spatial resolution). An example of the latter is timeliness of data: for weather forecast models, data are not helpful after they are several hours old, but for a climatologist studying a time series, old data are quite valuable. Therefore, an accompanying description of data properties such as those listed below is essential to help different user communities determine which specific data set is best for their own purpose.

Core metadata and quality tags for fitness-for-purpose assessment include

[11] Data Resources are defined for purposes of these guidelines to include observation data, derived products, information, models, and research results.

* Coverage, including both spatial and temporal dimensions
* Consistency, including long-term consistency
* Uncertainties estimated and documented, including both spatial and temporal dimensions
* Attribution of error sources
* Validation information, i.e. how the data was assessed for uncertainties by comparison with alternative measurements
* Latency from time of observation
* Resolution, including both spatial and temporal dimensions
* Usability, in the sense of being in form and content convenient to use for the specific purpose
* Simplicity, such that the data and the underlying data model are not unnecessarily complicated or difficult to understand or manipulate.

GEOSS Data User Communities should

1. Address fitness of purpose directly by recommending in a standard way and demonstrating the correct use of metadata to document dataset characteristics and quality: quality is in the eye of the beholder.

2. Provide feedback to data providers and other data users, using user feedback mechanisms (3), (6) set up for repositories distributing these data that increase the knowledge of the quality characteristics of the product's real use.

Appropriate other GEO bodies and tasks should

1. GEO tasks on data resources quality, such as the current GeoViQua (http://www.geoviqua.org), should encourage and provide mechanisms for standardized documentation of quality assessments, including the encoding of quantitative uncertainties at dataset, feature and pixel level (3), (7).

2. GeoViQua should provide mechanisms (repositories and search tools) by which data users can provide feedback to other data users on the practical usability of datasets. The mechanisms should allow association of user feedback with producer metadata for more complete assessment of fitness-for-purpose.

3. The Infrastructure Implementation Board should enable the provision of refereed standardized feedback from data users to data providers, and posting of these comments for the benefit of other users (e.g. data providers when registering data include email address). GEO provides a form to data users, including data quality elements and metrics for data quality that they can fill out and send back to provider. (6)

References

(1) QA4EO (http://qa4eo.org).
(2) ISO 19115:2003.
(3) GeoViQua's recommendations (and summarised model) for quality documentation can be found here in a GEO Data Quality Tutorial http://wiki.ieeeearth.org/Documents/GEOSS_Tutorials/GEOSS_Provider_Tutorials/Data_Quality_Tutorial_for_GEOSS_Providers/Section_0%3a_Table_of_Contents
(4) ISO 19157 Geographic Information Data Quality: has a description of data quality/fitness for purpose, and data quality "checklist" elements.
(5) ISO 19156: This standard for observations and measurements (O&M) has classes that associate quality and lineage with classes from other parts of the ISO metadata family of standards.
(6) See discussion of an example of an existing online library for climate data at http://onlinelibrary.wiley.com/doi/10.1002/2013EO130001/pdf. Title: **Climate Data Guide Spurs Discovery and Understanding (pages 121–122)** David P. Schneider, Clara Deser, John Fasullo and Kevin E. Trenberth. NCAR Climate Data Guide can be found at https://climatedataguide.ucar.edu/.
(7) UncertML (http://www.uncertml.org).