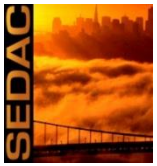
Center for International Earth Science Information Network
EARTH INSTITUTE | COLUMBIA UNIVERSITY



The Usability Subgroup of the NASA Earth Science Data Systems Working Group (ESDSWG) on Data Quality

Robert R. Downs[1]

[1] rdowns@ciesin.columbia.edu

NASA Socioeconomic Data and Applications Center (SEDAC)
Center for International Earth Science Information Network (CIESIN)
The Earth Institute, Columbia University

SUMMER 2015 MEETING OF THE
FEDERATION OF EARTH SCIENCE INFORMATION PARTNERS
Pacific Grove, California, 14-17 July 2015

Session: Information Quality Cluster
Friday 17 July 2015, 8:30 a.m.

Usability Subgroup Mission and Vision Statements

Mission:

Improve capabilities of Earth science data users to discover, assess, and use Earth science data by accessing, understanding, and evaluating quality aspects of the data.

Vision:

Users are able to easily and effectively discover, assess, and use Earth science data based on information about the quality of the data.

Source: Moroni, Ramapriyan, Downs, and Bagwell. 2015. p. 12

Usability Recommendations from Use Cases: Capturing DQ Information

Data System Related

- Data center: develop capabilities for users to comment on various categorized aspects of the quality of data sets that could be vetted and included in the publicly available information about each data set (leverage help desk such as Kayako - the help desk system being evaluated by EOSDIS, or establish public capability for commenting?).
- Data center: provide users with information on the distribution of errors for each data set, including the results of an outlier analysis for each variable.
- Data center: provide a software tool that can check for CF and ACDD metadata conformance using online CF checker at PCMDI or related tools (also being developed at PODAAC, ncdismember, and UDDC tool in the THREDDS data server, which checks and generates ACDD metadata reports and provides mapping to ISO 19115 metadata elements).
- Data center: request documentation from investigators on the extent of error introduced into data products as a result of binning and interpolation.

Science Related

- Science team: consult guidelines that describe categories of data quality and provide information and evidence about the quality of the data set for each category (ESDIS recommendations based on MEASURES 2006: Data Center and MEASURES PIs could evaluate these for potential use).
- Science team: develop capabilities to capture the distribution of errors for each data set and to conduct an outlier analysis for each variable.
- Science team and data systems researchers: collaborate to set up an appropriate scoring framework to check for CF and ACDD metadata conformance.
- Science team: for each data product, develop a data quality plan and submit it along with the data for dissemination.
- Science team: develop capabilities for investigators to describe the extent of error introduced as a result of binning and interpolation.

Derived from Downs et al., 2015

Usability Recommendations from Use Cases: Describing DQ Information

Data System Related

- Data center: provide enough publicly available information so users do not need to contact the DAAC (Self-describing documentation, Daymet, Vision 2020 goal to automate development of supplemental information).
- Data center: provide ease-to-use quality flags (like the confidence level flag used by GHRSST).

- Data Center develop capabilities to provide users with information about the correctness of the land mask that is distributed with each sea ice data product.
- Data Center develop capabilities for including and populating descriptions of quality flags for questionable ice values.
- Data Center develop capabilities to identify all deliverables that are described in the source and to verify that the data product contains all deliverables.
- Data center: provide users with documentation describing the canopy bias to enable estimations of bald Earth digital elevation models, ie; user guides & web pages.

Science Related

- Science team: describe quality flags in the data documentation and in the FAQs (keep populating FAQs).
- Science team should make quality flags publicly accessible and directly correspond to a quantifiable metric, such as the related uncertainty, confidence intervals, and confidence levels.
- Science team: describe any restrictions on the use of the data and visibly display the rights enabling the use and adaption of the data and of the data quality information.

- Science team: provide information about the correctness of the land mass distributed with each sea ice data product.
- Science team: provide users with a list of quality flags for questionable ice values along with descriptions for each quality flag, such as provided by MODIS land products.
- Science team: provide a list of all deliverables and describe how to find each deliverable and how it can be used.
- Science team: develop software to identify and describe the canopy bias to enable estimations of bald Earth digital elevation models.

Derived from Downs et al., 2015

Usability Recommendations from Use Cases: Finding DQ Information

Data System Related

- Data center: provide capabilities to present or visualize data quality indicators (e.g. use dropout percentage as sorting criteria; visualize dropout percentage map).
- Data center: provide capabilities to harmonize different spatial representations (i.e. boundaries and resolutions in different units & Spatial Reference System). For example, when a user specify 10km as the maximum spatial resolution, a data with 0.01-degree resolution shall be considered to be qualified.
- Data center: develop capabilities for data sets to be searched by selecting among values for a particular quality indicator.
- Data center: develop capabilities for users to search for data products that contain the same variables as a particular data product of interest.
- Data center: establish capabilities for users to enter terms to a list of search terms that can be selected by any user for inclusion in a search query.
- Data Center: develop capabilities for users to refine the results of search queries by selecting among choices of quantifiable data quality criteria, such as confidence levels or any quality flag derived from a quantifiable metric.

Science Related

- Science team: define or create data quality indicators that describe the quality characteristics of a data product (Data Quality Screening Service in GESDISC. Default Quality Flags and Advanced Quality Control from MODIS subset tool. Webification to extract quality indicators on the fly and also subset on the fly using quality indicators at PODAAC).
- Science team: provide definitions for each quality indicator and a description of how each quality indicator can be used (documentation, user guide, and in search system).
- Science team: identify variables that can be selected by users to search for all data sets that contain the selected variable.
- Science team: establish an authoritative list of scientific terms, such as those in the SWEET ontology and GCMD, that can be selected for inclusion in search queries to find the dataset.
- Science team: identify quantifiable data quality criteria, such as confidence levels and the values of quality flags, that can be used as criteria for refining search queries.

Derived from Downs et al., 2015

Usability Recommendations from Use Cases: Enabling Use of DQ Indicators

Data System Related

- Data center: provide storage that enables users to generate estimates for the probability of flooding for a location using Shuttle Radar Topography Mission (SRTM) Digital Elevation Models (DEM).
- Data center: provide users with a tool that identifies which inputs, such as AVHRR or MODIs Aqua/Terra, that have contributed to each pixel.
- Data center: use OpenDap or THREDDS to enable users to remotely interrogate data for the purposes of quality assessment, subsetting, aggregation, co-location, and visualization.

Science Related

- Science teams should create software that estimates the probability of flooding for a location using Shuttle Radar Topography Mission (SRTM) digital elevation models (DEM).
- Science teams should create tools that capture into a variable for Level 4 data sets, the sensor inputs, such as AVHRR or MODIS Aqua/Terra, that have contributed to each pixel.

Derived from Downs et al., 2015

References

- Downs R. R., Bagwell R., Ding F., Bennett S. D., Moroni D., Shen S., Wei Y. (2015) Data Lifecycle Recommendations for Improving the Usability of Data Quality Information for Earth Science Data. 2015 NASA Earth Science Data System Working Groups (ESDSWG) Meeting. Goddard Space Flight Center, Greenbelt, MD, March 24-26, 2015.
- Moroni, Ramapriyan, Downs, and Bagwell. Under Review. Earth Science Data System Working Groups 2014-2015 Data Quality Working Group Final Report. 30 June 2015