SciDataCon 2016

Conference paper

Ensuring and Improving Information Quality for Earth Science Data and Products – Role of the ESIP Information Quality Cluster

Hampapuram Ramapriyan, Science Systems and Applications, Inc.
Ge Peng, North Carolina State University
David Moroni, Jet Propulsion Laboratory/CalTech
Chung-Lin Shie, University of Maryland, Baltimore County

Summary

Quality of products is always of concern to users regardless of the type of products. The focus of this paper is on the quality of Earth science data products. There are four different aspects of quality – scientific, product, stewardship and service. All these aspects taken together constitute Information Quality. With increasing requirement on ensuring and improving information quality, there has been considerable work related to information quality during the last several years. Given this rich background of prior work, the Information Quality Cluster (IQC), established within the Federation of Earth Science Information Partners (ESIP) has been active with membership from multiple organizations. Its objectives and activities, aimed at ensuring and improving information quality for Earth science data and products, are discussed briefly.

Introduction

The quality of products is always of concern to users, whether they are buying a car or some other consumer goods, or using scientific data for research or an application. While the producers of the products are best able to assess the quality, conveying the information about quality in a manner that is understandable and usable is many times a challenge. Thus it is helpful to have a set of standards and “best practices” for collecting and conveying information about quality. The focus of this paper is on the quality of scientific data generated by Earth observation systems and their derived data products. First, four aspects of information quality will be defined. This will be followed by a section on the significant background work that has occurred over the last decade including: Quality Assurance Framework for Earth Observation (QA4EO), ISO 19157:2013 standard for geographic information data quality, NOAA Climate Data Records (CDR) Maturity Matrix, NOAA Data Stewardship Maturity Matrix, NCAR data guide, NASA MEaSUREs Product Quality Checklists, and activities of the NASA Data Quality Working Group. This will be followed by a discussion of the current activities of the Earth Science Information Partners IQC.

Information Quality

We consider four different aspects of quality. First, the *scientific quality*, defined in terms of accuracy, precision, uncertainty, validity and suitability for use (fitness for purpose) in various applications is considered paramount. Second, the *product quality* is important as well. Product quality addresses how well the scientific quality is assessed and documented, how complete the metadata and documentation are, etc. Third, *stewardship quality* addresses questions such as how well data are being managed, preserved and cared for by an archive or repository. Fourth, *service quality* deals with how easy it is for users to search, access, understand, trust, and use a given data product, as well as ensuring an archive has the requisite knowledge base and people functioning as subject matter experts available to help its data users. In general, we can refer to all these aspects of quality together as *Information Quality.*

Background

With increasing requirements on ensuring and improving information quality, there has been considerable work devoted to addressing information quality challenges over the last several years. In this section we will outline some of these activities briefly.

The Group on Earth Observations (GEO) identified the need for an internationally harmonized strategy to enable interoperability and acceptance of quality of Earth observation data at “face value”. In response to this, the Committee on Earth Observing Satellites (CEOS) established and endorsed the Quality Assurance Framework for Earth Observation (QA4EO). Following four international workshops (2007, 2008, 2009 and 2011), a framework and 10 key guidelines were established. Examples are provided (see http://qa4eo.org/case-studies/) to illustrate activities that are compliant with the QA4EO guidelines.

The standard ISO 19157:2013 was published in December 2013. See http://www.iso.org/iso/catalogue_detail.htm?csnumber=32575. It establishes the principles for describing geographic data quality and defines a set of measures for evaluating and reporting data quality. It is useful for: 1. data producers providing information on data quality, 2. data distributors providing users data quality guidance and 3. data users trying to decide whether or not a specific data product is suitable for their particular uses.

NOAA has developed an approach using a matrix to assess and document the maturity of individual Climate Data Records (CDRs) (Bates and Privette, 2012). The matrix defines six levels for maturity in each of the following six categories: Software Readiness, Metadata, Documentation, Product Validation, Public Access, and Utility. It provides a description, for each category, of what it means to be at various levels of maturity. EUMETSAT’s CORE-CLIMAX matrix is based on the CDR Maturity Matrix, and contains guidance on uncertainty measures.

The NOAA National Centers for Environmental Information (NCEI)/ Cooperative Institute for Climate and Satellites - North Carolina (CICS-NC) have developed a Data Stewardship Maturity Matrix (DSMM) (Peng et al, 2015). This matrix provides a unified framework for assessing the maturity of measurable stewardship practices applied to individual digital Earth Science data products that are publicly available. It assesses maturity in 9 categories (e.g., preservability, accessibility, data quality assessment, and data integrity) at 5 levels. It

provides understandable data quality information to users including scientists and actionable information to management.

The National Center for Atmospheric Research (NCAR) maintains a data guide with contributions from the community at the web site https://climatedataguide.ucar.edu/about/contribute-climate-data-guide. This is a resource used for gathering inputs from the climate community on a variety of observational data products and models. It takes advantage of the community's expertise to provide an assessment of data products by users for the benefit of other users. Inputs can be from both data product developers and users, self-identified as either "Expert Developers" or "Expert Users". The inputs received by this community are reviewed for quality before publication. For more details see Schneider et al (2013).

NASA's Making Earth System Data Records (ESDRs) for Use in Research Environments (MEaSUREs, https://earthdata.nasa.gov/community/community-data-system-programs/measures-projects) Program uses product quality checklists, which were developed in 2011. The product quality is considered to be a combination of scientific quality of the data and the completeness of associated documentation and ancillary information. The checklists are used to gather information on the completeness of activities needed to ensure product quality. The questions in the checklists address science quality, documentation quality, usage, and user satisfaction.

NASA's Data Quality Working Group (DQWG), one of the Earth Science Data System Working Groups (ESDSWG), was established in March 2014. Its mission is to "assess existing data quality standards and practices in the inter-agency and international arena to determine a working solution relevant to Earth Science Data and Information System Project (ESDIS), Distributed Active Archive Centers (DAACs), and NASA-funded Data Producers." The DQWG analyzed 16 use cases pertinent to data distributed by the DAACs from the point of view of users in order to identify issues related to information quality, and made nearly 100 recommendations for improvement. These were subsequently consolidated into 12 high priority recommendations, and 25 solutions to address these recommendations have been identified and assessed for operational maturity and readiness for implementation, with an initial focus on four "low-hanging fruit" recommendations; solutions that exist as open-source and in an operational environment were ranked as highest priority for implementation.

ESIP Information Quality Cluster

The Federation of Earth Science Information Partners (ESIP) is a US-based organization with international membership and "is an open, networked community that brings together science, data and information technology practitioners." (http://esipfed.org/). The ESIP initially formed the IQC in January 2011, led by Greg Leptoukh who was also taking an active role in QA4EO. With his unfortunate demise in January 2012, the IQC activities had become dormant until July 2014 when it was rejuvenated (see http://wiki.esipfed.org/index.php/Information_Quality). The current objectives of the IQC are to: 1. Actively evaluate community data quality best practices and standards; 2. Improve capture, description, discovery, and usability of information about data quality in Earth science data products; 3. Ensure producers of data products are aware of standards and best practices for conveying data quality, and data providers/distributors/ intermediaries establish, improve and evolve mechanisms to assist users in discovering and understanding

data quality information; and 4. Consistently provide guidance to data managers and stewards on how best to implement data quality standards and best practices to ensure and improve maturity of their data products.

The activities of the IQC include: 1. Identification of additional needs for consistently capturing, describing, and conveying quality information through use case studies with broad and diverse applications; 2. Establishing and providing community-wide guidance on roles and responsibilities of key players and stakeholders including users and management; 3. Prototyping of conveying quality information to users in a more consistent, transparent, and digestible manner; 4. Establishing a baseline of standards and best practices for data quality; 5. Evaluating recommendations from NASA's DQWG in a broader context and proposing possible implementations; and 6. Engaging data providers, data managers, and data user communities as resources to improve our standards and best practices.

Following the principles of openness of the ESIP Federation, IQC invites all individuals interested in improving capture, description, discovery, and usability of information about data quality in Earth science data products to participate in its activities.

Acknowledgements

This work was a result of the authors' participation in the ESIP IQC. They would like to thank the members of the IQC for their contribution to the discussions at the cluster meetings as well as comments on a draft of this paper. Ramapriyan's work was supported by NASA under a contract with Science Systems and Applications, Inc. Peng's work was supported by NOAA under a grant with CICS-NC. Moroni's work was supported by NASA under a contract with the Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA. Shie's work was supported by a NASA funding to the University of Maryland, Baltimore County.

Competing Interests

The authors declare that they have no competing interests.

References

Bates, J. J. and Privette, J. L. 2012, A maturity model for assessing the completeness of climate data records, *EOS, Transactions of the AGU*, **44**, 441.

Peng, G., et al 2015, A unified framework for measuring stewardship practices applied to digital environmental datasets, *Data Science Journal*, **13**, doi:10.2481/dsj.14-049.

Schneider, D. P., et al 2013, Climate Data Guide Spurs Discovery and Understanding, *Eos Trans*. AGU, **94(13),** 121.