Version: v02r00-20200827

Case Statement for Developing Community Guidelines for Consistently Curating and Representing Dataset Quality Information

The Pre-ESIP FAIR Dataset Quality Information Workshop Organizing Committee:
Ge Peng, Carlo Lacagnina, Robert Downs, Ivana Ivánová,
Gilles Larnicol, David Moroni, Hampapuram "Rama" Ramapriyan, and Yaxing Wei

Corresponding Authors: Ge Peng, ESIP IQC and CISESS/NCSU; gpeng@ncsu.edu
Carlo Lacagnina, BSC; carlo.lacagnina@bsc.es

Document History

Version	Contributors	Contribution/What is new
v01r00 20200702	The Pre-ESIP FAIR Dataset Quality Information Workshop Organizing Committee: Ge Peng, Carlo Lacagnina, Robert Downs, Ivana Ivánová, Gilles Larnicol, David Moroni, Hampapuram "Rama" Ramapriyan, and Yaxing Wei	First baseline of this document ("FAIR DQI Case Statement")
v02r00 20200827	The Organizing Committee	Second baseline of the FAIR DQI Case Statement – implemented review comments from the workshop participants. Sections affected: all.

Recommended Citation:
Peng, G., C. Lacagnina, R. R. Downs, I. Ivanova, G. Larnicol, D. F. Moroni, H. Ramapriyan, and Y. Wei, 2020: Case Statement for Developing Community Guidelines for Consistently Curating and Representing Dataset Quality Information. Version: v02r00-20200827.
https://doi.org/10.6084/m9.figshare.12605438

User Feedback:
Please direct your suggestions to: Ge Peng; gpeng@ncsu.edu. The latest version of this document will be maintained at and can be downloaded from figshare.com with the following persistent digital object identifier (doi):10.6084/m9.figshare.12605438

Version: v02r00-20200827

Case Statement for Developing Community Guidelines for Consistently Curating and Representing Dataset Quality Information

Ge Peng, Carlo Lacagnina, Robert Downs, Ivana Ivánová,
Gilles Larnicol, David Moroni, Hampapuram "Rama" Ramapriyan, and Yaxing Wei

Knowledge about the quality of data and associated information is important to support decision-making for taking solid, informed actions. Assessment of the data quality is key for ensuring that the available information is creditable and for establishing relations of trust between the data provider and various downstream users. Quality assessments inform the reliability and usability of the data, and quality information needs to be consistently curated, fully traceable, adequately documented and able to support users to address their specific needs. This is especially important for data used to support policy-making and enabling data to be reusable, in addition to being findable, accessible, and interoperable, as defined by the FAIR data principles (Wilkinson et al. 2016). Although the importance of access to quality-assured data is well recognized, methodologies of an evaluation framework and presentation of resultant quality information to end users may not have been comprehensively addressed, especially in an operational environment.

The Cloud offers both opportunities and challenges. These can be observed when testing the ability to integrate data across different systems, i.e., interoperability. The Cloud allows users to discover, access, subset/integrate, and analyze large volumes and varieties of data in a more efficient, scalable manner compared with legacy, on-premises data storage and distribution technologies. Determining how to establish trust and enable the correct use of the data is crucial to data sharing (Digital Science et al. 2019).

Technologies such as machine learning (ML) and artificial intelligence (AI) are becoming increasingly useful as tools to uncover or gain new knowledge from data. However, sound analysis needs to build on reliable data. Thus, it is becoming increasingly important to gain access to verifiable and consistently quality-controlled data when making data-driven decisions. Managing quality through the entire data life cycle is imperative for ensuring that the information and knowledge gained are not contaminated by inaccurate or corrupted data and for facilitating accurate uncertainty estimates in the derived analyses.

Therefore, it is crucial to consistently curate and represent quality information of data and associated information and make it readily available and integratable. However, dataset quality information is not routinely curated in a human- and machine-readable manner, despite the fact that the ISO 19157 standards for describing the quality of geographic data have been in place since 2013 (ISO 19157 2013). This is because there are many challenges, including but not limited to the following:

- Dataset quality information is multi-dimensional, i.e., science, product, stewardship, and services (Ramapriyan et al. 2017), with many quality attributes (Wang and Strong 1996) and many different maturity assessment models (see Peng 2018 for an overview; also see references in Table 1). Information about dataset quality traverses different knowledge domains and curating it requires cross-disciplinary collaborations.
- Currently, dataset quality information, when available, tends to focus on scientific quality and is published in science journals that are free-text-based and are not readily integrable into data management and stewardship processes or across different systems. For example, while uncertainty information is considered valuable, there are many approaches toward quantifying, characterizing, disseminating and interpreting that information (Moroni et al. 2019).
- Dataset quality information needs to be understandable by those who plan to use the described data as well as by those who are trying to determine whether the data are appropriate for their intended use.

Community guidelines for curating and representing dataset quality information in a consistent fashion and in line with the FAIR principles can help science data centers, repositories, data producers, publishers, data managers, and stewards establish trustworthiness and maximize reuse and value of their data. This effort requires a community-wide, cross-disciplinary effort to set a solid foundation for a wide implementation.

Therefore, we formally propose to develop community guidelines and the development will be done closely with domain experts from the Information Quality Cluster (IQC) of Earth Science Information Partners (ESIP), the Barcelona Supercomputing Center (BSC) team, which is the lead contractor for the Evaluation and Quality Control (EQC) function of the Copernicus Climate Change Services (C3S) Climate Data Store (CDS), the Australia Research Data Commons (ARDC), Open Geospatial Consortium (OGC), and other international organizations. (ESIP is mostly supported by NASA, NOAA, and USGS with over 150 national and international partners. C3S is one of six services of the European Union's Copernicus Earth observation programme and is implemented by ECMWF. ARDC facilitates a coherent research environment with Australian research community and industry access to eInfrastructure, platforms, and high-quality data. OGC an international voluntary community with over 500 global organizations.)

A planned guidelines document will focus on when dataset quality information needs to be created and how it can be captured and represented consistently. The guidelines will capture recommended practices developed by the community to ensure that dataset quality information is findable, accessible, interoperable, and reusable (FAIR). Much attention will also be dedicated to the provenance of the assessments, how it is presented and on rich-content with explicit license information to enable the resulting quality information be effectively shared and reusable.

To bring together national and international subject matter experts on the quality of datasets or data products to initiate the development of community guidelines, a one-day virtual workshop was held on Monday July 13, 2020. Invited speakers shared their knowledge to help attendees gauge the complexity and multi-dimensionality of dataset quality information. This knowledge exchange allowed attendees to understand why Earth science organizations need to curate and represent data quality information throughout the entire life cycle of a dataset – from the data product design and production stage, through data and metadata curation stage for preservation and access, to data use stage by servicing data to consumers. It also helped attendees appreciate the challenges those organizations face, and learn the different approaches they take to lay the groundwork for the development of community guidelines. Additional details can be found in the workshop summary report (Peng et al. 2020).

Given the increasingly prominent role that quantitative climate information assumes in decision-making, there is no doubt that the quality of this data will come under increasing scrutiny in the future. Indeed, an increasing number of national and international initiatives focus on developing the technical solutions to provide easily accessible, timely and sustained data quality information for data-driven decision-making. Collecting challenges and lessons learned across the different communities can contribute to the development of recommended practices for data quality information curation. The outcomes of this effort will be beneficial to organizations including NASA, NOAA, USGS, JPSS, EUMETSAT, ESA, C3S, WMO, GEOSS, WGISS, etc., to improve the trustworthiness and sharing of their data and information in different Earth Science disciplines.

References

Digital Science, B. Fane, P. Ayris, M. Hahnel, I. Hrynaszkiewicz, G. Baynes, et al., 2019: The State of Open Data Report 2019. Digital Science. Report. https://doi.org/10.6084/m9.figshare.9980783.

ISO 19157, 2013: Geographic information—Data quality. Version: 2013-12. Geneva, Switzerland. Available at: https://www.iso.org/standard/32575.html

Moroni, D. F., H. Ramapriyan, G. Peng, J. Hobbs, J. C. Goldstein, R. R. Downs, R. Wolfe, C.-L. Shie, C. J. Merchant, M. Bourassa, J. L. Matthews, P. Cornillon, L. Bastin, K. Kehoe, B. Smith, J. L. Privette, A. C. Subramanian, O. Brown, & I. Ivánová, 2019: Understanding the Various Perspectives of Earth Science Observational Data Uncertainty. *Figshare*. https://doi.org/10.6084/m9.figshare.10271450

Peng, G., 2018: The state of assessing data stewardship maturity – An overview. *Data Science Journal*. 17, doi:10.5334/dsj-2018-007

Peng, G., C. Lacagnina, R. Downs, I. Ivánová, D. Moroni, H. Ramapriyan, Y. Wei, and G. Larnicol, 2020: Laying the Groundwork for Developing International Community Guidelines to Effectively Share and Reuse Digital Data Quality Information – Case Statement, Workshop Summary Report, and Path Forward. Version: v03r04-20200828. *Open Science Framework*, doi:10.31219/osf.io/75b92

Ramapriyan, H., G. Peng, D. Moroni, and C.-L. Shie, 2017: Ensuring and Improving Information Quality for Earth Science Data and Products. *D-Lib Magazine*, 23, 10.1045/july2017-ramapriyan

Wang, R.Y. and D. M. Strong, 1996: Beyond Accuracy: What Data Quality Means to Consumers. *Journal of Management Information Systems,* 12(4):5, https://doi.org/10.1080/07421222.1996.11518099

Wilkinson, M., and Coauthors, 2016: The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Scientific Data*, 3, https://doi.org/10.1038/sdata.2016.18

Acronyms

ARDC	Australian Research Data Commons
BSC	Barcelona Supercomputing Center
C3S	Copernicus Climate Change Service
CEOS	Committee on Earth Observation Satellites
CDS	Copernicus Data Store
CISESS	Cooperative Institute for Satellite Earth System Studies
DQD WG	OGC Data Quality Domain Working Group
DQI	Dataset Quality Information
DQ IG	ARDC Data Quality Interest Group
ESA	European Space Agency
ESIP	Earth Science Information Partners
EUMETSAT	European Organisation for the Exploitation of Meteorological Satellites
FAIR	Guiding principles for Findable, Accessible, Interoperable, and Reusable
GEOSS	Group on Earth Observations System of Systems
IQC	Information Quality Cluster of ESIP
JPSS	Joint Polar Satellite Systems
NASA	National Aeronautics and Space Administration
NCSU	North Carolina State University
NOAA	National Oceanic and Atmospheric Administration
OGC	Open Geospatial Consortium
USGS	United States Geological Survey
WGISS	Working Group on Information Systems and Services
WMO	World Meteorological Organization

Glossary

Data can refer to anything that is collected, observed, or derived and used as a basis for reasoning, discussion, or calculation. Data can be either structured or unstructured, and can be represented in quantitative, qualitative, or physical forms.

Scientific or research data is defined as the recorded factual material commonly accepted in the scientific community as necessary to validate research findings.

Digital data, distinguished from physical records, such as paper weather reports, are represented in discrete numerical form that can be used by a computer or electronic device.

Environmental data are defined as the recorded and/or derived observations and measurements of the physical, chemical, biological, geological, or geophysical properties or conditions of the oceans, atmosphere, space environment, sun, and solid earth, as well as correlative data and related documentation or metadata.

Geospatial data describe the state and impact of environmental systems and include information on the geographic location and characteristics of constructed features and boundaries of the earth.

Climate data is a subset of environmental data that is particularly sensitive to stewardship due to the length of its practical life cycle, breadth of scope and the required consistency across its period of record. Climate is the historical behavior of atmosphere and ocean systems, weather is the day-to-day conditions of atmosphere and oceans in a region and their short-term (from minutes to weeks) variation.

Data product refers to a product that facilitates an end goal through the use of data, usually with a well-thought out algorithm or approach. Data products tend to be structured and can be raw measurements or scientific products derived from raw measurements or other products. Products can also be statistical or numerical model outputs, including analyses, reanalyses, predictions, or projections. Earth Science data products may be further categorized based on their processing levels.

Dataset is an identifiable collection of physical records, a digital rendition of factual materials, or a product of a given version of an algorithm/model. A dataset may contain one or many physical samples or data files in an identical format, having the same geophysical variable(s) and product specification(s), such as the geospatial location or spatial grid. Dataset and data product may be used interchangeably.

Information is considered as data being processed, organized, structured, or presented in a given context, while *knowledge* is gained from an understanding of the significance of information. Data and information may overlap and may be used interchangeably.

Dataset quality includes quality of both data and associated information.

Dataset quality information includes quality of both data and associated information such as data quality descriptive information such as those captured in documents, e.g., papers or reports, and data quality metadata that is captured in a metadata record, throughout the entire life cycle of a dataset.

Dataset quality management refers to a data quality function is made of different groups: scientific quality, technical quality, dissemination of the quality information to the users, possibly more groups. All these aspects together are considered as dataset quality management, which is part of data governance. What guarantees that data is not corrupted and is accurate are the quality control procedures + a series of protocols to avoid new quality issues in the system. This latter is part of the quality assurance. Both quality control and quality assurance can consider scientific (e.g. uncertainty) and technical (e.g. temporal completeness) aspects. The information acquired during the quality control/assurance processes has to be disseminated to improve "usability" of the data and "verifiability" of the quality procedures applied, this to reach the ultimate goal of increasing trustworthiness in data and information disseminated by the operational services. How to disseminate this quality-related information is another piece of quality management.

Knowledge is an abstract concept, defined as a familiarity, awareness, or understanding of someone or something, gained through education, experience, or association. It can refer to a theoretical or practical understanding of a subject.

Provenance is information about the origin and history of entities, activities, and people involved in producing a piece of data or thing.

M*aturity model* refers to a maturity reference or assessment model with desired evolution in discrete stages from a certain aspect or perspective of dataset quality.

Uncertainty information refers to the information extracted from data representing a quantitative and/or qualitative assessment of the errors (both random and systematic) pertaining to an estimated quantity of interest (QOI) and is particularly useful at both estimating and inferring the confidence of a particular QOI (e.g., sea surface temperature).