COMMITTEE ON DATA
CODATA
INTERNATIONAL
SCIENCE COUNCIL

Global Community Guidelines for Documenting, Sharing, and Reusing Quality Information of Individual Digital Datasets

RESEARCH PAPER

]u[ubiquity press

GE PENG
CARLO LACAGNINA
ROBERT R. DOWNS
ANETTE GANSKE
HAMPAPURAM K. RAMAPRIYAN
IVANA IVÁNOVÁ
LESLEY WYBORN
DAVE JONES
LUCY BASTIN
CHUNG-LIN SHIE
DAVID F. MORONI

*Author affiliations can be found in the back matter of this article

ABSTRACT

Open-source science builds on open and free resources that include data, metadata, software, and workflows. Informed decisions on whether and how to (re)use digital datasets are dependent on an understanding about the *quality* of the underpinning data and relevant information. However, quality information, being difficult to curate and often context specific, is currently not readily available for sharing within and across disciplines. To help address this challenge and promote the creation and (re)use of freely and openly shared information about the quality of individual datasets, members of several groups around the world have undertaken an effort to develop international community guidelines with practical recommendations for the Earth science community, collaborating with international domain experts. The guidelines were inspired by the guiding principles of being findable, accessible, interoperable, and reusable (FAIR). Use of the FAIR dataset quality information guidelines is intended to help stakeholders, such as scientific data centers, digital data repositories, and producers, publishers, stewards and managers of data, to: i) capture, describe, and represent quality information of their datasets in a manner that is consistent with the FAIR Guiding Principles; ii) allow for the maximum discovery, trust, sharing, and reuse of their datasets; and iii) enable international access to and integration of dataset quality information. This article describes the processes that developed the guidelines that are aligned with the FAIR principles, presents a generic quality assessment workflow, describes the guidelines for preparing and disseminating dataset quality information, and outlines a path forward to improve their disciplinary diversity.

CORRESPONDING AUTHOR:
Ge Peng
Earth System Science Center/
NASA MSFC IMPACT, The
University of Alabama in
Huntsville, Huntsville, AL, US
gpeng93@gmail.com

KEYWORDS:
quality; trust; open-source science; FAIR; guidelines; metadata

TO CITE THIS ARTICLE:
Peng, G, Lacagnina, C, Downs, RR, Ganske, A, Ramapriyan, HK, Ivánová, I, Wyborn, L, Jones, D, Bastin, L, Shie, C-L and Moroni, DF. 2022. Global Community Guidelines for Documenting, Sharing, and Reusing Quality Information of Individual Digital Datasets. *Data Science Journal*, 21: 8, pp. 1–20. DOI: *https://doi.org/10.5334/dsj-2022-008*

1. BACKGROUND

Informed decisions on whether and how to (re)use particular digital datasets rely on knowledge about aspects of data and metadata quality, including their completeness, accuracy, provenance and timeliness (Digital Science et al. 2019; Peng et al. 2021a). Quality assessments also improve the reliability and usability of both data and metadata (Callahan et al. 2017) and are crucial for supporting open-source science and data-driven policy-making processes (Peng et al. 2020a; 2021a).

A dataset in this article refers to a collection of data that is identifiable (ISO 19115-1 2014), and has the potential to be curated or published by a single actor (W3C 2020). A particular dataset can digitally represent a group of observations, a data product from a specific version of a processing algorithm based on observations, output of numerical model(s), or outcomes of laboratory experiments.

Dataset quality information embodies information about the quality or state of data (input, output, and ancillary), metadata, documentation, software, procedures, processes, workflows, and infrastructure that were created or utilized during the entire lifecycle of a dataset (Peng et al. 2021a). Therefore, the focus of this article is on *dataset* quality – not just data quality.

To be effectively shared and utilized, quality information needs to be consistently curated, preferably traceable, and appropriately documented (Peng et al. 2021a). The granularity of this quality documentation may vary – sometimes be very fine (e.g., per-observation in the case of volunteered observations) but the critical common resolution required to support FAIR data publishing is the individual dataset level.[1] Quality assessment results also need to be represented consistently, updated regularly, and should be integrable across systems, services, and tools to enable improved data sharing (Henzen et al. 2021; Wagner et al. 2021; Peng et al. 2021a).

While the needs for assessments about the quality of data and related information for a particular dataset are well recognized, an approach for a framework to evaluate and present such quality information to data users (e.g., Figgemeier et al. 2021) may not have been sufficiently developed and/or sufficiently addressed for disciplinary or interdisciplinary use. In response, an international workshop was held virtually on 13 July 2020 to pursue the needs and challenges for preparing and documenting dataset quality information consistently during the complete dataset lifecycle by a group of global Earth science, interdisciplinary domain experts. A number of challenges were identified in Peng et al. (2020b), and three are highlighted below.

First, the selection of relevant quality attribute(s) (e.g., accuracy, completeness, relevancy, timeliness, etc.) is largely dependent upon context and can yield multiple quality categories and practical dimensions (Lee et al. 2002; Ramapriyan et al. 2017; Redman 1996; Wang & Strong 1996). This multi-dimensionality makes the assessment of dataset quality a complex endeavor. For example, the quality attribute of completeness can refer to the completeness of data values in both spatial and temporal spaces, or the completeness of metadata elements or content. The multi-dimensionality of dataset quality has been discussed in detail by Peng et al. (2021a).

An example of grouping dataset quality into four aspects (i.e., science, product, stewardship, and service) through the entire dataset lifecycle is shown in ***Figure 1***. For each aspect, three important stages are listed along with selected quality attributes which do not constitute an exhaustive list. Those dataset lifecycle stages do not necessarily cover all activities. They may not necessarily happen sequentially, and also may occur in more than one quality aspect. For example, the 'Evaluate' part of the lifecycle in the 'Product' quadrant may overlap with the 'Science' by influencing the 'Validate' part. However, generally speaking, activities in the dataset lifecycle identified in the 'Science' quadrant occur before those in the 'Product' quadrant as noted by the direction of the arrows in ***Figure 1***. Note that the term 'Develop' used in the 'Science' quadrant also includes data observation/acquisition. The feedback and improvement cycle can occur in any one of the stages.

[1] *https://www.gbif.org/data-quality-requirements*.

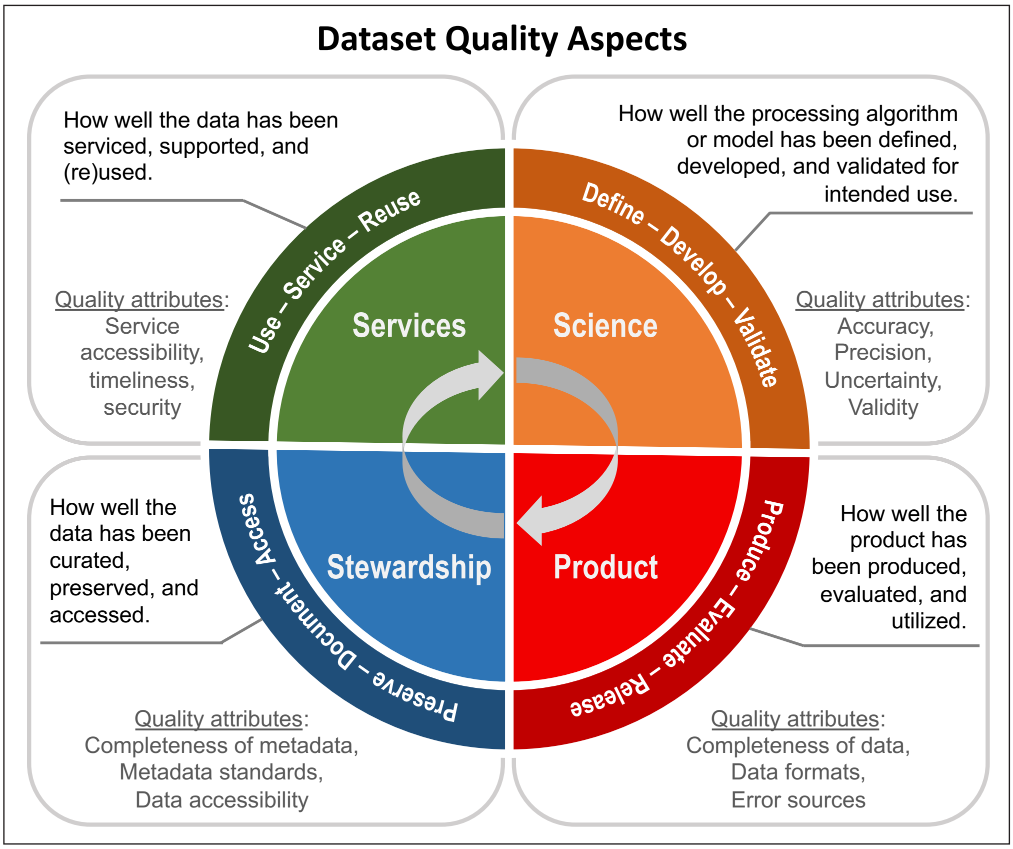

Dataset Quality Aspects
How well the data has been serviced, supported, and (re)used.
Quality attributes: Service accessibility, timeliness, security
How well the processing algorithm or model has been defined, developed, and validated for intended use.
Quality attributes: Accuracy, Precision, Uncertainty, Validity
How well the data has been curated, preserved, and accessed.
Quality attributes: Completeness of metadata, Metadata standards, Data accessibility
How well the product has been produced, evaluated, and utilized.
Quality attributes: Completeness of data, Data formats, Error sources
Use – Service – Reuse
Define – Develop – Validate
Preserve – Document – Access
Produce – Evaluate – Release
Services
Science
Stewardship
Product

Figure 1 Brief description of four quality aspects (i.e., science, product, stewardship and service) throughout a dataset lifecycle, three key stages and a few quality attributes associated with each quality aspect (e.g., define, develop, and validate stages for the science quality aspect). The quality aspects and associated stages are based on Ramapriyan et al. (2017) with the following changes, based on feedback from the ESIP community and the International FAIR Dataset Quality Information (DQI) Community Guidelines Working Group: i) 'Assess' replaced by 'Evaluate' in the Product aspect; ii) 'Deliver' replaced by 'Release' in the Product aspect; and iii) 'Maintain' replaced by 'Document' in the Stewardship aspect. Additionally, completeness of metadata is moved from the Product to Stewardship aspect. Creator: Ge Peng; Contributors to conceptualization: Lesley Wyborn and Robert R. Downs.

Second, quality attributes are often not defined, measured, or captured consistently, even within one discipline. Moroni and colleagues recently observed such complexity as it pertains to the uncertainty of Earth science data (Moroni et al. 2019). Consistency in defining quality attributes and converging to standardized assessment models may be optimal for sharing, but more progress needs to be made, and whether such consistency is achievable remains to be seen. A step towards cross-domain interoperability, however, may be achieved by thorough documentation of domain-specific quality assessment techniques and metrics and the full provenance of the quality assessment. This allows transformations to be applied to dataset quality scores when this is possible and appropriate, e.g., computation of an exceedance value or quantile from a mean and standard deviation (Bastin et al. 2013, Section 5.1).

The third challenge is associated with the paradigm shift in the capabilities of the designated community of scientific data: from domain literate with familiarity of the scientific context and intended use of data products, to potential users representing diverse fields of inquiry (Baker et al. 2016), with increasing demand for machine interoperability. Therefore, the existence of a wide range of stakeholders and data users, including those with very little or no science background, should be considered to facilitate the analysis, interpretation, understanding of research data and related information and in some cases acted upon (Peng et al. 2021a).

Any effort to maximize the sharing of quality information requires collaboration among members of the entire community across science, data management, and technology domains. Recognizing that, 32 workshop participants – all international domain experts – issued an open 'call-to-action for global access to and harmonization of quality information of individual Earth science datasets' (Peng et al. 2021a). In response to that action call and further motivated by the needs of and interest from the global Earth science community, the International FAIR Dataset Quality Information (FAIR-DQI) Community Guidelines Working Group was formed.

Working group members comprise international domain experts, such as data producers and contributors, data managers and curators from scientific institutes and data centers, and data consumers and publishers. Given their common interest in dataset quality information, this group of people can be regarded as a 'Community of Practice (CoP)' (E. Wenger-Trayner & B. Wenger-Trayner 2015). Together, the members of this group possess valuable first-hand knowledge and expertise in dealing with the challenges of developing, managing, disseminating, and using a variety of Earth science data products and services, such as data products obtained from surface, airborne, and satellite observations as well as output from numerical models.

Since September 2020, the members of this working group have been working collaboratively to develop practical guidelines for data managers and repositories to follow when preparing, representing, and reporting on the quality of individual datasets. These guidelines build on the success of the FAIR Guiding Principles for data sharing (Wilkinson et al. 2016) and on the extensive expert knowledge and practical experiences of working group members, while leveraging community practices. This article describes the development principles and processes, captures the outcomes of this international community effort, and presents a path forward toward enhancing the coverage of disciplines beyond Earth sciences.

This article is organized as follows. A background has been provided in this section. The principles, scope, goals, and intended audience for the development of the guidelines are provided in Section 2, while the development process is described in Section 3. The guidelines developed are presented in Section 4, with a workflow for initiating and carrying out quality assessment, as well as a description of crosswalks to elements of the FAIR Principles. Potential impact of the guidelines, benefits of CoP, and path forward are discussed in Section 5, with a conclusion in Section 6.

2. DEVELOPMENT PRINCIPLES, SCOPE, GOALS, AND INTENDED AUDIENCE

2A. DEVELOPMENT PRINCIPLES

The following principles are utilized to guide the development of the guidelines, based on feedback from the Earth science community:

i. A holistic dataset life-cycle approach should be adopted for developing guidelines.

ii. Guidelines should be produced in an iterative manner with continuous community engagement for feedback.

iii. Guidelines should be independent of specific quality attributes, assessment types, and context of applications.

iv. Any methodology that is utilized to evaluate certain dataset quality attribute(s) should be findable and accessible, and preferably be interoperable and reusable for both human users and machine users.

v. The assessment results should be openly available findable, accessible, interoperable, and reusable to both human users and machine users.

vi. Transparent and quantifiable quality assessments should be a part of a dataset quality management framework.

vii. Guidelines should be regularly updated and version controlled.

2B. SCOPE

Given the complexity of dataset quality attributes and different contexts of their fitness for use, the guidelines will focus on providing guidance for capturing and representing dataset quality information consistently, adapting the FAIR Guiding Principles. Preparing such guidance will foster data use by providing users with consistent, timely, and accessible information that is available to effectively make educated data (re)use decisions for their unique application requirements. The guidelines do not focus on what quality attributes, aspects, or dimensions to assess; what assessment models to use; or how to assess dataset quality. However, a basic workflow has been developed, and practical examples are provided as references to help organizations and data stewards get started.

A dataset lifecycle in the context of this article starts at the planning and designing stage of developing a data product (***Figure 1***).[2] It will not touch on sensor algorithms or model development and deployment. However, it is also important to capture and describe quality

[2] It is possible that planning of data products starts long before data are collected, as for satellite missions.

information such as algorithm model parameters (e.g., accuracy, precision, uncertainty) during these development and deployment stages, because the quality information from these stages is critical for identifying error sources; estimating data product uncertainty (Moroni et al. 2019); and examining error progression to downstream applications (e.g., Matthews et al. 2013).

2C. GOALS

This international community effort has been undertaken to develop guidelines for the Earth science community, in collaboration with international domain experts on data and information quality. The primary objective of the guidelines has been to offer the Earth science community actionable recommendations that can be adopted by a variety of stakeholders to consistently capture, represent, and integrate dataset quality information. Treating dataset quality information as a digital object and being consistent with the FAIR Guiding Principles, improves its potential for sharing and reuse with more targeted practicality. Care was taken so that the guidelines would be general enough to be readily adopted or adapted by other research science communities. The optimal goal is to foster global access to and harmonization of quality information of datasets as a critical step towards facilitating open-source science in both machine- and human-friendly environments as called for by Peng et al. (2021a).

2D. INTENDED AUDIENCE

All data stakeholders may benefit from the community guidelines:

- *Data producers* will find these useful to ensure at the point of acquisition that critical attributes are captured. Such attributes will later be used to ascertain the quality of the data they are capturing (e.g., uncertainty of location/measurements, instrument parameters, metadata attributes on the instrument used to acquire the data).
- *Data publishers and data curators* may find the community guidelines valuable for improving the quality information associated with the data that they publish and manage.
- *Sponsors and funders* may find the guidelines helpful when reviewing data management plans in proposals for the support of projects and programs that will be creating, curating, disseminating, and supporting the use of data. They will also find them useful during the project closure phase when assessing the quality of the data products generated against the initial project goals and data management plans.
- *Data users* may find that the guidelines improve their understanding of quality issues when determining whether a particular data product or service is appropriate for their intended use and what the limitations may be for using the data. This could support the application of 'confidence levels' to certain information derived from the data.

3. DEVELOPMENT PROCESS: TIMELINES AND WORKFLOW

This section provides a detailed description of the process of developing a framework through an international collaboration with the expectation that it will be useful for other groups or communities that may be considering similar endeavors.

The idea of potentially developing a framework for consistently capturing quality information for enabling the use of Earth science datasets was initiated in September 2019 (***Figure 2***). Follow-on discussions on community needs and the prospect of developing community guidelines for documenting and reporting dataset quality information as described in Peng et al. (2020b), were carried out among several groups across the globe. These groups include the Earth Science Information Partners (ESIP) Information Quality Cluster (IQC), the Barcelona Supercomputing Center (BSC) Evaluation and Quality Control (EQC) team, and the Australia/New Zealand Data Quality Interest Group (AU/NZ DQIG).

ESIP, which was founded in 1998, is primarily supported by United States Earth science governmental agencies, including the National Aeronautics and Space Administration (NASA), the National Oceanic and Atmospheric Administration (NOAA), and the United States Geological Survey (USGS). ESIP members include over 150 national and international partner organizations. The ESIP IQC fosters cross-disciplinary collaborations to evaluate various facets of Earth science data and information quality and produces recommended practices for the community. The

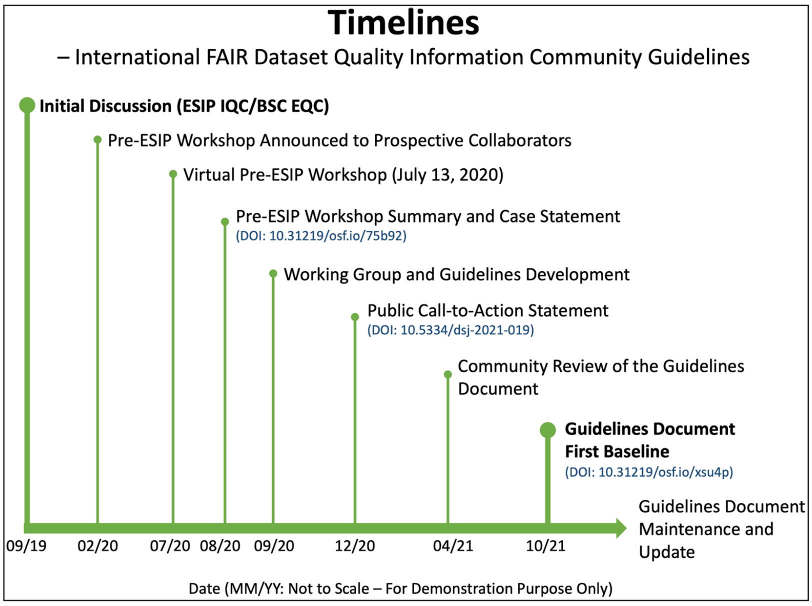
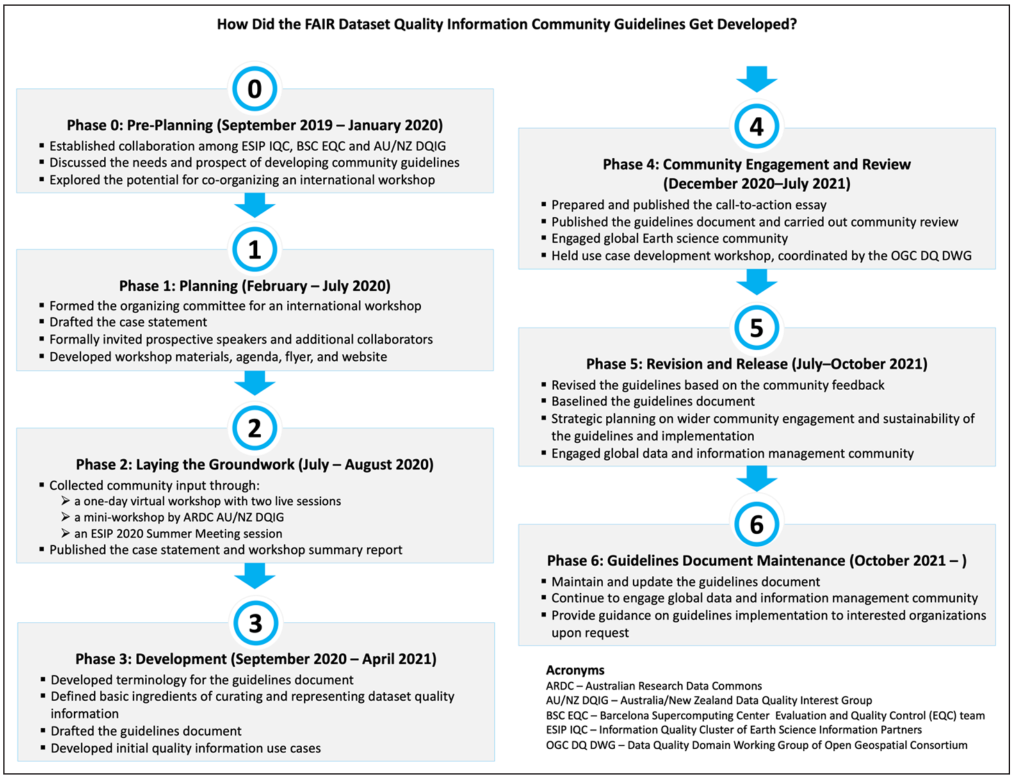

Timelines
– International FAIR Dataset Quality Information Community Guidelines
Initial Discussion (ESIP IQC/BSC EQC)
Pre-ESIP Workshop Announced to Prospective Collaborators
Virtual Pre-ESIP Workshop (July 13, 2020)
Pre-ESIP Workshop Summary and Case Statement
(DOI: 10.31219/osf.io/75b92)
Working Group and Guidelines Development
Public Call-to-Action Statement
(DOI: 10.5334/dsj-2021-019)
Community Review of the Guidelines Document
Guidelines Document First Baseline
(DOI: 10.31219/osf.io/xsu4p)
Guidelines Document Maintenance and Update
09/19
02/20
07/20
08/20
09/20
12/20
04/21
10/21
Date (MM/YY: Not to Scale – For Demonstration Purpose Only)

Figure 2 Schematic diagram of timelines of the initiation, planning, development, community review, and first baseline of the guidelines document. The guidelines document will be updated in the future to improve its coverage in diverse disciplines. ESIP IQC: Information Quality Cluster of the Earth Science Information Partners. BSC EQC: Barcelona Supercomputing Center (BSC) Evaluation and Quality Control (EQC) team.

BSC EQC team supports the EQC function of Copernicus Climate Change Service (C3S) Climate Data Store, one of six services of the European Union's Earth observation programme. The AU/NZ DQIG is a forum for AU/NZ data providers, repository operators and data consumers and it is facilitated by the Australian Research Data Commons (ARDC). The AU/NZ DQIG was founded in late 2019 by ARDC, Curtin University and the Australian National University (ANU).

Support from the ESIP leadership was committed in early 2020 to sponsor a whole-day, in-person, international workshop prior to the ESIP 2020 summer meeting (SM20) with an additional report-out session during the SM20. The goal of the pre-ESIP workshop was to convene international domain experts to kick off the development of the guidelines by exploring the needs, challenges and current state of documenting and reporting dataset quality information. Invitations for participation were sent to prospective collaborators.

In the wake of the COVID-19 pandemic, the in-person workshop was changed to a virtual event, allowing it to be extended to a wider audience. A case statement was drafted and published to help set the stage and communicate the effort (Peng et al. 2020a). The workshop website[3] was established to host the workshop materials and additional resources (***Figure 3***).


How Did the FAIR Dataset Quality Information Community Guidelines Get Developed?

Phase 0: Pre-Planning (September 2019 – January 2020)
- Established collaboration among ESIP IQC, BSC EQC and AU/NZ DQIG
- Discussed the needs and prospect of developing community guidelines
- Explored the potential for co-organizing an international workshop

Phase 1: Planning (February – July 2020)
- Formed the organizing committee for an international workshop
- Drafted the case statement
- Formally invited prospective speakers and additional collaborators
- Developed workshop materials, agenda, flyer, and website

Phase 2: Laying the Groundwork (July – August 2020)
- Collected community input through:
  - a one-day virtual workshop with two live sessions
  - a mini-workshop by ARDC AU/NZ DQIG
  - an ESIP 2020 Summer Meeting session
- Published the case statement and workshop summary report

Phase 3: Development (September 2020 – April 2021)
- Developed terminology for the guidelines document
- Defined basic ingredients of curating and representing dataset quality information
- Drafted the guidelines document
- Developed initial quality information use cases

Phase 4: Community Engagement and Review (December 2020–July 2021)
- Prepared and published the call-to-action essay
- Published the guidelines document and carried out community review
- Engaged global Earth science community
- Held use case development workshop, coordinated by the OGC DQ DWG

Phase 5: Revision and Release (July–October 2021)
- Revised the guidelines based on the community feedback
- Baselined the guidelines document
- Strategic planning on wider community engagement and sustainability of the guidelines and implementation
- Engaged global data and information management community

Phase 6: Guidelines Document Maintenance (October 2021 –)
- Maintain and update the guidelines document
- Continue to engage global data and information management community
- Provide guidance on guidelines implementation to interested organizations upon request

Acronyms
ARDC – Australian Research Data Commons
AU/NZ DQIG – Australia/New Zealand Data Quality Interest Group
BSC EQC – Barcelona Supercomputing Center Evaluation and Quality Control (EQC) team
ESIP IQC – Information Quality Cluster of Earth Science Information Partners
OGC DQ DWG – Data Quality Domain Working Group of Open Geospatial Consortium

Figure 3 Flowchart outlining different phases of the guidelines development process, including the initiation, planning, development, community review and engagement, and baseline of the guidelines.

[3] *https://wiki.esipfed.org/Pre-ESIP_Workshop*.

About 80 ESIP and invited international domain experts, affiliated with over 40 private, academic and governmental institutions from nine countries within North America, Oceania, and Europe registered for the workshop (Peng et al. 2020b). Two live 90-minute virtual workshop sessions were held on July 13, 2020, to accommodate attendees from different time zones. More than 45 workshop registrants attended the first live session while approximately 25 attended the second. About 45 ESIP SM20 registrants attended the subsequent report-out session. Prior to this workshop, a mini-workshop had been held by AU/NZ DQIG on July 6, 2020, where 57 had registered and 27 participated actively.

Eleven invited speakers presented during the two virtually live workshop sessions and additional three presented at the 90-minute report-out session during SM20. Invited speakers represented diverse international organizations, including major international space agencies and satellite programs, such as the NOAA Joint Polar Satellite System (JPSS) program (Goldberg & Zhou 2020), European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT) (Schulz 2020), and European Space Agency (ESA) (Albani & Maggio 2020). Presentations described data stewardship activities at global organizations, such as the Group on Earth Observations (Downs 2020) and the World Meteorological Organization (Lief et al. 2020); as well as major national Earth science data and service centers, including those for NASA (Wei et al. 2020), NOAA (Ritchey 2020), USGS (Hou 2020) and Copernicus Marine Environment Monitoring Service (CMEMS) (Drévillon et al. 2020). (See Table 1 of Peng et al. 2020b, for the full list of presentation titles, affiliated organizations, and citations).

The speakers shared their knowledge to help participants ascertain the complexity and multi-dimensionality of curating dataset quality information. This knowledge exchange allowed participants to understand why Earth science organizations need to prepare and describe data quality information throughout the entire dataset lifecycle – covering stages from data product design and production, through data and metadata curation for preservation and access, to data use by servicing data to consumers. It also helped attendees appreciate the challenges those organizations face and learn about the different approaches taken. These informative presentations provided perspective for productive discussions among participants during the live sessions. Notes were recorded online in a collaborative Google Doc and offline discussions continued following the workshop during the two weeks of the virtual SM20. For many of the over twenty non-US pre-ESIP workshop attendees, this was their first time engaging with the ESIP community (Peng et al. 2020b).

The strong need for practical guidelines was recognized as an opportunity to provide the community with guidance to improve data sharing by consistently preparing and representing information about the quality of datasets. The absence and limitations of currently available guidance also was recognized (Peng et al. 2021a). Participants of both the pre-ESIP workshop and the subsequent SM20 session have stressed the need for such guidelines to be created by the community and for the community through an iterative process with community feedback (Peng et al. 2020b).

Several community calls to participate voluntarily in an international working group were announced during the pre-ESIP workshop and the subsequent SM20 session, along with messages to relevant Earth science email lists, including the ESIP community list. Since September 2020, over twenty international domain experts have joined the working group, which has begun developing the guidelines by consolidating community recommendations first (***Figures 2*** and ***3***). A white paper on the guidelines was published for community review in April 2021 (Peng et al. 2021b, version 3). Extensive outreach was conducted by working group members to share the initial draft of the guidelines document with the Earth science and geospatial data community (e.g., Downs et al. 2021; Lacagnina et al. 2021a–b; Peng et al. 2020c; Peng et al. 2021c–i; Wyborn et al. 2021). The guidelines document, partially reproduced below, has since been revised to release the first baseline version, which reflects community comments and suggestions (Peng et al. 2021b).

4. FAIR DATASET QUALITY INFORMATION GUIDELINES

In this section, we first define a basic workflow with relevant elements to consider when setting out to assess dataset quality and curate quality information. A set of the guidelines developed by the International FAIR-DQI Community Guidelines Working group are then presented, followed by crosswalks between the guidelines to the FAIR Guiding Principles.

4A. BASIC WORKFLOW FOR CURATING DATASET QUALITY INFORMATION

While assessing dataset quality is multi-dimensional (Peng et al. 2021a), there are common aspects. Knowledge about these common aspects may help to set the direction for the right approach in each specific case of assessing quality and reporting assessment results.

To help organizations and data stewards address the challenge of where to start when curating and reporting dataset quality information, we have developed a typical workflow (***Figure 4***). This approach is inspired by the quality evaluation procedures defined in ISO 19157 (2013) and Six Sigma (e.g., Cordy & Coryea 2006), and follows the steps outlined below to define, measure, analyze, and improve, as presented in Lee et al. (2002) for organizing data quality management.

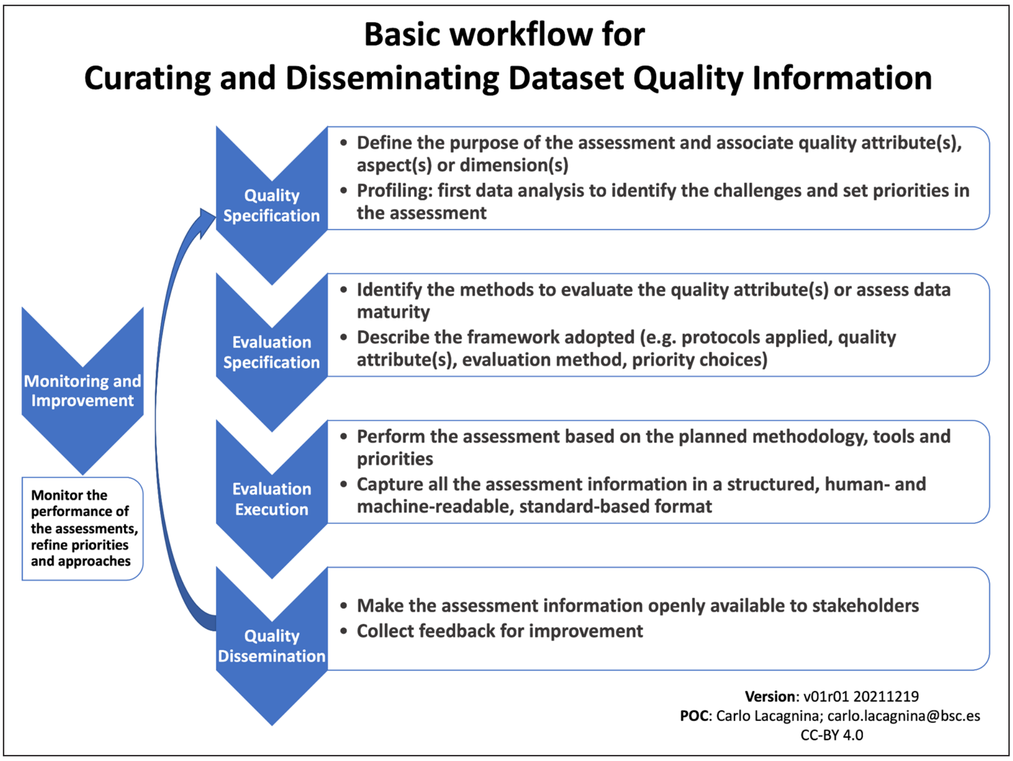

Basic workflow for
Curating and Disseminating Dataset Quality Information
Quality Specification
• Define the purpose of the assessment and associate quality attribute(s), aspect(s) or dimension(s)
• Profiling: first data analysis to identify the challenges and set priorities in the assessment
Evaluation Specification
• Identify the methods to evaluate the quality attribute(s) or assess data maturity
• Describe the framework adopted (e.g. protocols applied, quality attribute(s), evaluation method, priority choices)
Evaluation Execution
• Perform the assessment based on the planned methodology, tools and priorities
• Capture all the assessment information in a structured, human- and machine-readable, standard-based format
Quality Dissemination
• Make the assessment information openly available to stakeholders
• Collect feedback for improvement
Monitoring and Improvement
Monitor the performance of the assessments, refine priorities and approaches
Version: v01r01 20211219
POC: Carlo Lacagnina; carlo.lacagnina@bsc.es
CC-BY 4.0

Figure 4 A schematic diagram of a basic workflow with relevant elements for curating and disseminating dataset quality information. Creator: Carlo Lacagnina. Contributor: Ge Peng.

The workflow highlights some of the basic ingredients and elements to be considered at each step when curating dataset quality information. We add the dissemination, a.k.a. 'reporting' in ISO 19157 (2013), of dataset quality information, which is becoming an increasingly important task for building trust between data providers and end-users and for improving data usability.[4]

As shown in ***Figure 4***, the following two steps are needed prior to carrying out any assessment activity.

Step 1: Quality specification – Curating dataset quality information should start with defining the quality attribute(s), aspect, or dimension that will be assessed, determining the level of granularity (variable, ensemble member, model or algorithm), and identifying which data and quality attribute should be prioritized. This step will need some profiling, that is, an initial analysis of the available data to understand the challenges and the most critical issues to set priorities and determine the appropriate strategy to deploy (e.g., Cosoli & Grcic 2019; Woo & Gourcuff 2021).

Step 2: Evaluation specification – The next step involves identifying or developing an approach (or method) to evaluate the identified quality attribute(s) or assess its maturity. Example approaches could include a statistical analysis approach (Wu et al. 2017) or a scientific maturity matrix (Zhou et al. 2016). In this step, the framework for the evaluation is defined. It is important to describe the identified quality attribute or dimension, the evaluation method used, and the protocols, standards and workflows applied (e.g., Cosoli & Grcic 2019; Lemieux

[4] *https://is.enes.org/* (the Infrastructure for the European Network for Earth System Modelling (IS-ENES) programme).

et al. 2017; Popp et al. 2020; Woo & Gourcuff 2021; Wu et al. 2017; Zhou et al. 2016). A well-documented quality evaluation helps to increase transparency, verifiability, reproducibility, and resilience of the quality evaluation process.

The next two steps are important to capture and convey the resultant quality information.

Step 3: Evaluation execution – During this stage, the actual assessments are performed based on the tools, approaches and priorities defined in the previous steps. While doing this, the assessments should be captured in structured, human- and machine-readable, and standard-based formats (e.g., Heydebreck et al. 2020; Peng et al. 2019a).

Step 4: Quality dissemination – The results of the assessments represent the core of the dataset quality information and need to be disseminated with the data for the benefit of end-users. Feedback from users on data quality is beneficial to data producers to initiate data improvement processes. For reproducibility purposes, it is recommended that the operations performed to produce the quality information also be published (e.g., Davies & Sommerville, 2020). In this step, the mechanism for quality information dissemination (e.g., metadata, web page, API) is implemented and put into practice.

Finally, feedback from users on dataset quality information should be sought and evaluated to improve the quality information provided along with how the information is disseminated.

Step 5: Monitoring and improvement – The feedback collected in the previous step and the experience gained during the assessments are rationalized to consider improvements of the protocols, tools, and approaches and to redefine priorities in the assessment process (e.g., Cosoli & Grcic 2019; Wu & Gourcuff, 2021). This step is completed continuously throughout the assessment to dissemination steps, as it helps to improve the curation of quality information.

4B. GUIDELINES FOR ENABLING FAIR DATASET QUALITY INFORMATION

The following five guidelines are developed by the International FAIR-DQI Community Guidelines Working Group to enable curated dataset quality information to be FAIR (i.e., findable, accessible, interoperable, and reusable), for both human users and machines. A description of crosswalks to relevant elements of the FAIR Principles, which are denoted as F1-F4 for Findable, A1-A2 for Accessible, I1-I3 for Interoperable, and R1 for Reusable, is provided (see Wilkinson et al. 2016 for the definitions of the FAIR Principles).

The current state of dataset compliance with these guidelines varies. Most, if not all, datasets do not yet fully satisfy these guidelines. While it is difficult to find examples of datasets that comply with all the guidelines, it is still useful to provide examples that illustrate how each individual guideline is being met. This is the approach followed below. Additional examples can be found in Peng et al. (2021b).

Guideline 1: Describe dataset (title, persistent identifier [PID] with a comprehensive landing page, e.g., digital object identifier [DOI], product uniform resource identifier [URI], version, data producer, publication/update date, publisher, date accessed, usage license, e.g., CC-BY 4.0 or CC0).

This guideline aims to ensure that the underlying dataset is findable, comprehensively described, and potentially reusable by cross-walking to all the F1-F4 principles of *Findability*, and the R1 (rich metadata with a plurality of relevant attributes) and R1.1 principles (data usage license) of *Reusability*, either directly or indirectly, denoted by solid and dashed lines in ***Figure 5***, respectively.

Specifically, having a dataset PID leads to satisfying F1 (data are assigned a unique and persistent identifier). Given the nature of PID and the required landing page ensures that the (meta)data are indexed and resolvable (F4). To have a comprehensive landing page of a dataset, both data and metadata need to be described with numerous pertinent attributes, which leads to satisfy F2 (data are described with rich metadata) and R1 principles, respectively. Including a usage license leads to supporting the R1.1 principle.

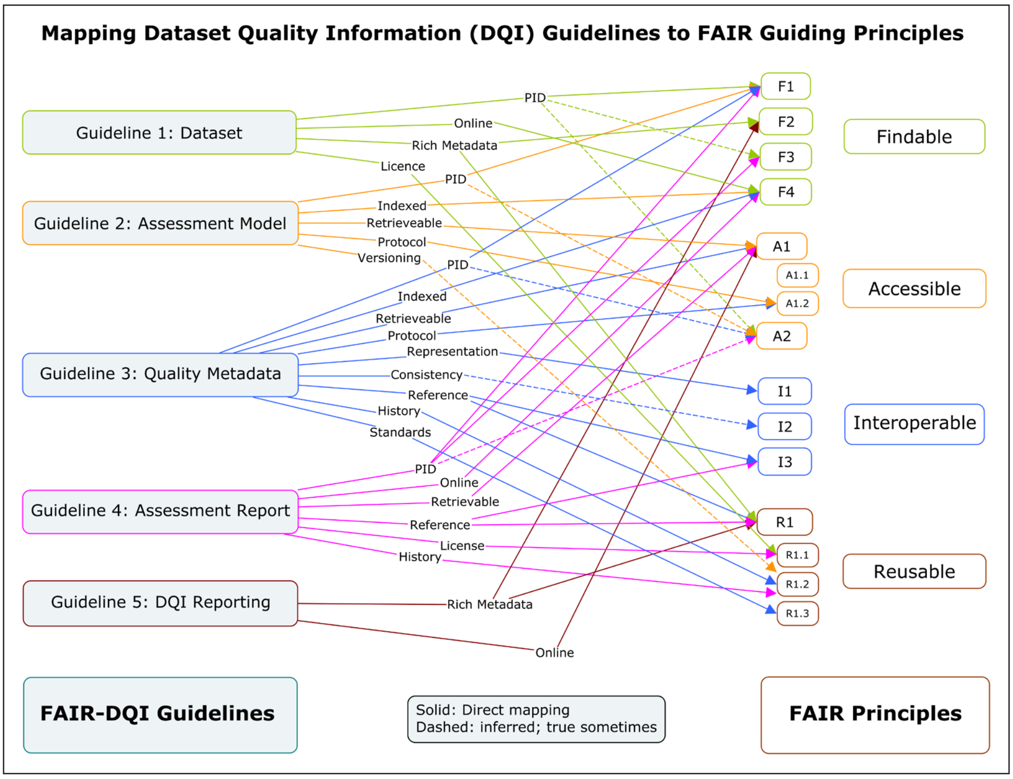


Figure 5 Diagram mapping the guidelines to the FAIR Guiding Principles as defined in Wilkinson et al. (2016).[5] Solid lines represent direct mapping while the dashed lines represent indirect or weak mapping that are either inferred or may not always hold. {F, A, I, R}n denotes the nth element of the findable, accessible, interoperable, and reusable principles, respectively. Based on Table 1 in Peng et al. (2021b), with additional weak mappings represented by the dashed lines. Creator: Ge Peng. Contributor: Anette Ganske.

The current common practice is to include the data PID in the metadata (F3) as part of the process of assigning and minting that PID. If the data PID is minted by a service provider such as DataCite, metadata should continue to be accessible even beyond the availability of the data (A2). However, since it is largely up to practices implemented by individual organizations, it yields only an indirect crosswalk from the guideline 1 to these two FAIR principles (F3, A2).

There are many examples of published datasets that meet this guideline by following community data citation standards. Two of them are shown below:[5]

> Neumann, D, Matthias, V, Bieser, J and Aulinger, A (2017). Concentrations of gaseous pollutants and particulate compounds over northwestern Europe and nitrogen deposition into the north and Baltic Sea in 2008. World Data Center for Climate (WDCC) at DKRZ. License: CC BY 4.0. Created: 2017–06–08. *https://doi.org/10.1594/WDCC/CMAQ_CCLM_HZG_2008*.
>
> Maggi, F, Tang, F H M, la Cecilia, D and McBratney, A (2020). Global Pesticide Grids (PEST-CHEMGRIDS), Version 1.01. Created: September 2020. License: CC-BY 4.0 International. Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC). *https://doi.org/10.7927/weq9-pv30*

Guideline 2: Utilize a one- (or more) dimensional, structured quality assessment metric that is:

2.1. versioned and publicly available with a globally unique, persistent and resolvable identifier (PID) such as digital object identifier (DOI) and universally unique identifier (UUID);

2.2. registered or indexed in a searchable resource that supports authentication and authorization, such as Figshare, Zenodo, GitHub, and Dryad; and

2.3. retrievable by their identifier using an open, free, standardized and universally implementable communications protocol such as Hypertext Transfer Protocol Secure (HTTPS) or Open Archives Initiative – Protocol for Metadata Harvesting (OAI-PMH).

This guideline aims to ensure that the assessment model is searchable and retrievable (***Figure 5***). Requirement 2.1 leads to satisfying F1 (assignment of PID), while Requirements 2.2 and 2.3 ensure that F4 and A1 (registered (meta)data and their retrievability) are satisfied, respectively. The authentication and authorization requirements in 2.2 meet A1.2. The requirements for

[5] FAIR Principles, meaning of each element and examples: *https://www.go-fair.org/fair-principles/*.

protocol in 2.3 lead to satisfying A1.2. The versioning itself is far short of the information required for assessment of model provenance. However, it helps support provenance (R1.2). Therefore, an indirect crosswalk to R1.2 is indicated in ***Figure 5***.

Examples of existing dataset quality assessment models and their compliance with Guideline 2 are provided in ***Table 1***. Additional assessment model examples can be found in Peng et al. (2021b).

If no suitable assessment model is available, one may need to develop a new one. In this case, above requirements 2.1–2.3 should be satisfied to make the assessment model findable and accessible. Individual researchers can also use the Registry of Research data Repositories (re3Data) at *https://doi.org/10.17616/R3D* to search for appropriate repositories based on their own requirements. A CoreTrustSeal certified repository demonstrates more matured organizational processes and capabilities in managing its holdings of digital objects (CoreTrust Seal 2019).

ASSESSMENT MODEL	SCIENTIFIC DATA STEWARDSHIP MATURITY MATRIX (PENG ET AL. 2015)	STEWARDSHIP MATURITY MATRIX FOR CLIMATE DATA (PENG ET AL. 2019B)	FAIR DATA MATURITY MODEL (RDA FAIR DATA MATURITY MODEL WORKING GROUP 2020)	METADATA QUALITY FRAMEWORK (BUGBEE ET AL. 2021)	DATA QUALITY ANALYSES AND QUALITY CONTROL FRAMEWORK (WOO & GOURCUFF 2021)
Quality Entity (i.e., attribute, aspect, or dimension)	Stewardship	Stewardship	FAIRness	Metadata	Data
2.1 – Publicly Available	Yes	Yes	Yes	Yes	Yes
2.1 – PID	DOI	DOI	DOI	DOI	DOI
2.2 – Indexed	Data Science Journal	Figshare	Zenodo	Data Science Journal	Integrated Marine Observing System Catalog
2.3 – Retrievable **Using Free, Open, Standard-Based Protocol**	Yes	Yes	Yes	Yes	Yes

Table 1 Examples of dataset quality assessment models and their compliance with Guideline 2.

Minimally, a published paper (with DOI) that describes a quality assessment model is necessary to provide access to the model. We highly recommend publishing the assessment model itself (with DOI), for example, in one of the aforementioned repositories. A project website tends to be a common place currently, but is often not sustainable or persistent due to the limited lifespan of projects. For example, a broken link as a result of organizational system migration will lead to inaccessibility of the assessment model.

Guideline 3: Capture the quality attribute(s)/aspect(s)/dimension(s), assessment method and results in a dataset-level metadata record using a consistent framework/schema that:

3.1. is semantically and structurally consistent and follows community standards – preferably compliant with national or international metadata standards that satisfy the conditions of Guideline 2 (i.e., 2.1–2.3),

3.2. includes a description of the quality attribute(s), aspect(s), or dimension(s) to be assessed,

3.3. includes a description of the assessment method and assessment model structure and version, and access date if applicable,

3.4. includes a description of the assessment results, and

3.5. includes versioning and the history of the assessments.

This guideline aims to ensure that the quality information is captured or referenced in the dataset metadata and that it is findable, accessible, interoperable, and reusable by machine end-users (***Figure 5***).

Utilizing a metadata framework/schema that satisfies the conditions 2.1–2.3 of Guideline 2 ensures that it is findable and accessible.

The requirements of capturing quality entity (i.e., attribute, aspect, or dimension), assessment method and results and that in 3.1 help ensure that the dataset-level metadata is richly described (R1) following metadata standards (R1.3) and is machine interoperable (I1). Capturing the assessment method is often accomplished by referencing it in the metadata record, which satisfies I3; as is capturing assessment results in the form of a published report.

Specifically, including a description of the information related to assessments, that is, quality entity, method, and results as required in 3.2–3.5, leads to rich metadata with a plurality of relevant attributes (R1). The semantically and structurally consistent metadata record that is compliant with standards (3.1) and crosswalks to I1 and R1.3. It may also potentially meet the requirements of I2 (FAIR-compliant vocabularies) in a best-case scenario but may fall short in most of the cases, so only a weak mapping is denoted by the dashed line (***Figure 5***). The requirements in 3.5 support the provenance of the assessment results (R1.2).

Examples of existing approaches in representing quality entities, assessment models and assessment results in machine-readable quality metadata and their compliance with Guideline 2 are provided in ***Table 2***. Additional examples can be found in Peng et al. (2021b).

QUALITY METADATA FRAMEWORK	NOAA *ONESTOP* DSMM QUALITY METADATA (PENG ET AL. 2019A)	ATMODAT MATURITY INDICATOR (HEYDEBRECK ET AL. 2020)	METADATAFROMGEODATA (WAGNER ET AL. 2021)
Quality Entity	Stewardship	Any Quality Entity	Data and Metadata
3.1 – Semantically and Structurally Consistent	Yes	Yes	Yes
3.1 – Metadata Framework/Schema	International	Domain	Domain
3.2 – Quality Entity Description	Yes	Yes	Yes
3.3 – Assessment Method/Structure Description	Yes	Yes	Partly (contains evaluation of quality description and not description of quality assessment)
3.4 – Assessment Results Description	Yes	Yes	Yes
3.5 – Versioning and the History of the Assessments	Yes	Versioning	Creation & Last Update Dates

Table 2 Examples of representing quality entities, assessment models and assessment results in machine-readable quality metadata and their compliance with Guideline 3.

Adopting or adapting (including information about the adaptation) existing quality metadata frameworks also is recommended. If that is not possible**,** a new quality metadata framework or schema may be developed. In this case, the framework should have the capability to allow for requirements in 3.1–3.5 to be satisfied.

Using a consistent metadata tag and including it in a schema is recommended, if applicable. For example, Peng et al. (2019a) uses MM-Stew as a metadata tag to denote stewardship maturity assessment. Once the new schema is stable, registering it with schema.org or other relevant metadata schema host entities, such as DataCite, is recommended.

Guideline 4: Describe comprehensively the assessment method, workflow, and results in at least a human-readable quality report that:

4.1. preferably follows a template that is published and satisfies the conditions of Guideline 2 (i.e., 2.1–2.3),

4.2. is published with an explicit open license and history of the report, satisfying the conditions of Guideline 2, and

4.3. links the report PID to the dataset-level metadata record.

This guideline aims to at least ensure the quality information is findable, accessible, citable, reusable and understandable to human end-users (***Figure 5***). However, we strongly encourage quality reports to be also machine readable.

Comprehensively describing the relevant information yields human-readable metadata with multiple attributes (R1: richly described metadata). Publishing the assessment report following the criteria 2.1–2.3 with an explicit open license (4.2) leads to F1 (PID), F4 ((meta)data registered in a searchable resource), A1 ((meta)data retrievable via standardized protocol), and R1.1 (clear data usage license). The inclusion of the report history (4.2) supports R1.2. Linking the report PID to the dataset-level metadata record (4.3) satisfies the F3 (PID in metadata) and I3 (references to other metadata) principles, respectively.

Examples of existing approaches in representing quality entities, assessment models, and assessment results in human-readable quality reports and their compliance with Guideline 4 are provided in ***Table 3***. Additional examples can be found in Peng et al. (2021b).

QUALITY REPORT	LEMIEUX ET AL. (2017)	HÖCK ET AL. (2020)	COWLEY (2021)
Quality Entity	Stewardship	Data	Data
4.1 – Follow Template	Yes	Yes	Yes
4 – Quality Entity Description	Yes	Yes	Yes
4 – Assessment Method Description	Yes	Yes	Yes
4 – Assessment Results Description	Yes	Yes	Yes
4.2 – License	Yes	Yes	Yes
4.2 – Assessment History	Yes	Yes	Yes
4.3 – Linked Report PID	Yes	No	Yes

Table 3 Examples of human-readable dataset quality assessment reports and their compliance with Guideline 4.

Guideline 5: Report/disseminate the dataset quality information in an organized way via a web interface with a comprehensive description of:

5.1. the dataset according to the Guideline 1,

5.2. assessed quality attribute(s)/aspect(s)/dimension(s),

5.3. the evaluation method and process including the review process, if applicable, and

5.4. how to understand and use the information.

This guideline aims to ensure that the quality information is online and comprehensively described, findable, and easily understood and trusted by providing the assessment provenance (***Figure 5***).

A comprehensive description of the dataset (requirement 5.1), the assessed quality attribute/aspect (5.2), the evaluation method (5.3), and how to understand and use the quality information (5.4) leads to rich metadata with a plurality of relevant attributes (F2 and R1). The nature of reporting or disseminating and being online indicates it is retrievable via a standardized communication protocol (A1).

Examples of existing approaches in representing assessment results online and their compliance with Guideline 5 are provided in ***Table 4***. Additional examples can be found in Peng et al. (2021b).

There is a large diversity in current approaches to disseminate data and metadata quality information because of the dependency on the knowledge-base of the designated community for data. Data users should provide feedback on which disseminated quality information is most relevant and how it can be improved. Therefore, user engagement activities are quite relevant at this stage, including prompt responses to questions and suggestions received from users.

ONLINE PORTAL	JPSS DATA PRODUCT ALGORITHM MATURITY PORTAL[6]	C3S CLIMATE DATA STORE DATASET QUALITY ASSESSMENT PORTAL[7]	ROLLINGDECK TO REPOSITORY (R2R) QA DASHBOARD[8]
Quality Entity	Algorithm	Technical and Scientific Quality	Sensor
5 – Report information in an organized way	Yes	Yes	Yes
5.1 – Dataset Description	Minimal	Yes	Minimal
5.2 – Assessed Quality Entity Description	Yes	Yes	Yes
5.3 – Evaluation Method and Review Process Description	Yes	Yes	Yes
5.4 – Description of How to Understand and Use Description	Some	Some	Minimal

Table 4 Examples of disseminating assessment results online and their compliance with Guideline 5.

Likewise, it is also recommended to convey dataset quality information in a manner that is easily understood and usable by data users and provide a mechanism for user feedback.

5. DISCUSSION

This section provides a brief discussion of the potential impact of the guidelines provided above, benefits of CoP, and the path forward to increasing community awareness of the guidelines and promoting their adoption.

5A. POTENTIAL IMPACT OF THE GUIDELINES

Improving practices for documenting, sharing, and reusing information about the quality of datasets will help advance scientific progress and contribute to societal benefits through open-source science. When dataset quality information enables potential users to discover a dataset and determine whether it is appropriate for an intended use, FAIR data quality information also helps to achieve FAIR data (Peng et al. 2021a). Likewise, when information describing the quality of a dataset fosters its interoperability and reusability, the guidelines further help to make the data FAIR. Those elements of the guidelines which focus on documentation of quality assessment strategies have the additional potential to make FAIR not just the data, but also those evaluation processes. This articulation and communication of domain-specific models, protocols and assumptions can support robust interdisciplinary re-use of data.

In addition, adoption of the guidelines for dataset quality information by the Earth science community, as well as by other disciplinary communities, offers an opportunity to improve the trust that potential users have in the underlying datasets. From a user's perspective, finding relevant, trusted data is critical to driving decisions. By improving practices for documenting, sharing, and reusing information about the quality of datasets, data providers and users will have increased confidence and improve consistency when disparate datasets are accessed, overlayed, and shared to drive impact-based decisions. These guidelines can assist in establishing trusted approaches for enabling diverse in-situ observing platforms to be used with confidence when assessing, for example, water quality information in estuaries, rivers, bays, and oceans when those sensors may have been installed and funded by different state and federal agencies.

Furthermore, providing sufficient information, including quality information, for using datasets within data collections has the potential to improve trust in the data repositories that are responsible for curating and sharing data (Lin et al. 2020). Clearly, community guidelines for dataset quality information would also benefit disciplines beyond the Earth sciences and efforts are underway to increase their discipline diversity.

[6] *https://www.star.nesdis.noaa.gov/jpss/AlgorithmMaturity.php*.

[7] *https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-pressure-levels-monthly-means?tab=eqc*.

[8] *https://www.rvdata.us/qa_info*.

5B. BENEFITS OF A COMMUNITY OF PRACTICE (COP)

With common interests and passions about sharing quality information, the members of the International FAIR-DQI Community Guidelines Working Group have come together to essentially form a loosely organized CoP. The development of the guidelines benefited from the common advantages of a CoP. These include knowledge sharing on needs, challenges, and practices in curating and representing quality information from diverse Earth science domains. There are also added benefits of participating in a CoP throughout the development process. Two will be highlighted below.

One is that we are all learning together. Knowledge about other perspectives broadens our own point of view that comes from our own experiences. A large part of developing knowledge is developing consensus through learning from each other.

Another is that we bring what we have learned back to our jobs, organizations, and communities. Changing is a long process of learning, accepting, and adapting – the first and hardest part is culture change. The subtle changes we make through knowledge we learned can become the seeds that lead to much-needed culture change in our organizations and communities towards sharing quality information at large.

5C. PATH FORWARD

The guidelines should help organizations and data stewards get started on providing dataset quality information to data consumers – an important step to close the chasm between data producers and users. However, adoption often requires culture change, which demands continued engagement with the Earth science community (e.g., Lacagnina et al. 2021a).

The effective sharing and (re)use of dataset quality information needs cross-disciplinary integration. Efforts are underway to engage and collaborate with other communities and disciplines beyond Earth science, such as:

- Open Geospatial Consortium (OGC; Ivánová et al. 2021 – OGC Data Quality Workshop – citizen science, Earth science, geospatial science, machine learning, social science, urban planning),
- World Data System (Ramapriyan et al. 2021 – SciDataCon session – astronomy, citizen science, Earth science, social science), and
- Research Data Alliance (RDA) (Peng et al. 2021h – RDA18th Plenary session – astronomy, Earth science, genomics, social science). Activities are underway towards forming an RDA working group on making dataset quality information FAIR for the RDA community.

It has been pointed out by the community during our on-going engagement that it will be beneficial to develop and provide use cases for data quality and implementation of the guidelines. The OGC Data Quality Domain Working Group (OGC DQ DWG)[9] is currently working towards the development of a catalog of data quality use cases and we will be contributing to the effort.

6. CONCLUSION

The FAIR Guiding Principles described by Wilkinson et al. (2016) provide a succinct and measurable set of concepts to be used as a guideline for improving the access and reusability of data for human users and machines. Although the FAIR Principles have provided an effective way to enable data sharing, they do not explicitly describe how dataset quality information should be curated and shared.

Inspired by the FAIR Guiding Principles, a set of guidelines for curating and reporting dataset quality information were developed for both human users and machines, as a global community effort. The guidelines development effort was carried out by a Community of Practice through an iterative process guided by community feedback. The process of developing the guidelines has been described, which may be of use to inspire similar activities requiring large community consensus and uptake.

[9] *https://www.ogc.org/projects/groups/dqdwg*.

The guidelines aim to improve the availability and usability of quality information at the individual digital dataset level. Utilizing a structured quality assessment model helps to ensure the consistency of evaluation methods and results, which in turn will make it easier to capture them consistently. Capturing the assessment results in the dataset-level metadata using a consistent framework improves machine interoperability and supports integration across systems and tools. Disseminating the dataset quality information in a transparent and user-friendly way will help end users to understand and effectively use or integrate the information.

Community guidelines developed as a result of this effort bring the Earth science community one step closer to standardizing the curation and representation of dataset quality information. The guidelines described in this article offer opportunities to enable or improve the transparency and interoperability of dataset quality information. Adopting all or part of the guidelines can contribute to the ecosystem that supports open-source science. An excellent byproduct of streamlining the curation and representation of dataset quality information is the improved likelihood of automating the curation and reporting process, leading to international access to and usability of information about the quality of individual digital datasets (Peng et al. 2021a).

Utilizing the guidelines also helps improve the overall FAIRness of a dataset by providing community-standard-based rich metadata with a plurality of relevant quality attributes and qualified references. It establishes the trustworthiness of data and ultimately improves the maturity of a dataset in multiple quality dimensions or aspects including product, stewardship, and services by improving the completeness and usability of metadata and documentation.

The international FAIR-DQI community guidelines document (Peng et al. 2021b) is a living document and is expected to evolve over time to accommodate user feedback and emerging community best practices. As indicated in Section 5c, use cases will be developed, in collaboration with OGC DQ DWG, to further improve the maturity and comprehensiveness of the guidelines and provide implementation examples for the global Earth science and geospatial community. Furthermore, in collaboration with the RDA community, an effort is underway to improve the discipline diversity of the guidelines.

ACKNOWLEDGEMENTS

The development and baseline of the community FAIR-DQI guidelines document would not have been possible without the voluntary and dedicated effort of the domain experts of the International FAIR-DQI Community Guidelines Working Group. We would like to thank all members of the working group for their interest, participation, and contribution.

The members of the International FAIR-DQI Community Guidelines Working Group are:

Ge Peng, Carlo Lacagnina, Ivana Ivánová, Robert R. Downs, Hampapuram K. Ramapriyan, Anette Ganske, Dave Jones, Lucy Bastin, Lesley Wyborn, Irina Bastrakova, Mingfang Wu, Chung-Lin Shie, David Moroni, Gilles Larnicol, Yaxing Wei, Nancy Ritchey, Sarah Champion, C. Sophie Hou, Ted Habermann, Gary Berg-Cross, Kaylin Bugbee, and Jeanné le Roux.

Kaylin Bugbee reviewed the manuscript with beneficial comments that have improved the clarity of the manuscript. Constructive comments from two DSJ anonymous reviewers have improved the clarity and readability of the paper.

We thank ESIP for sponsoring the virtual pre-ESIP workshop held on July 13, 2020, and for providing infrastructure support for the workshop and maintaining the FAIR-DQI website. We acknowledge the pre-ESIP workshop organizing committee members: G. Peng, C. Lacagnina, R. R. Downs, I. Ivánová, D. F. Moroni, H. K. Ramapriyan, Y. Wei, and G. Larnicol. The Australian participants acknowledge the support of the ARDC.

We thank all participants in community discussions during the development of the guidelines document, especially the following people who have provided beneficial information and/or comments: Mitch Goldberg, Jörg Schulz, Mirko Albani, Christina Lief, Shelley Stall, Lihang Zhou, Iolanda Maggio, Marie Drévillon, Brian Westra, Siri Jodha Khalsa, Miles Nicholls, Erin Kenna, Kerstin Lehnert, Paul Lemieux, Donald Collins, Gastil Gastil-Buhl, Danie Kinkade, Chantel Ridsdale, Shawn Smith, Shannon Leslie, Rebecca Hudak, Tyler Christensen, Venice Bayrd, David Green, Karen Moe, and Boyi Shangguan.

We thank all the guidelines document reviewers who have taken their time and efforts in reviewing the guidelines document and provided us with valuable comments and suggestions, which have improved the clarity and comprehensiveness of the document. Special thanks goes to Siri Jodha Khalsa, Erin Kenna, Stefan Schliebner, Alfred Stein, Peter Strobl, Tyler Christensen, Jasmine Muir, Jessie Oliver, Kenneth Casey, and Christin Henzen and her colleagues from the Geoinformatics/TU Dresden and the GeoKur project team.

Contents from the FAIR-DQI case statement (Peng et al. 2020a), the workshop summary report (Peng et al. 2020b), and the guidelines document (Peng et al. 2021b) are reused in this article.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

GP is the main contributor to conceptualization, project planning, and writing of the original draft. CL contributed significantly to conceptualization, project planning, writing of the original draft, reviewing, and editing. RD contributed significantly to conceptualization, writing of the original draft, reviewing, and editing. AG contributed significantly to writing of the original draft, reviewing, and editing. HR and LW contributed to conceptualization, writing of the original draft, reviewing, and editing. II contributed to conceptualization, project planning, reviewing, and editing. DJ and LB contributed to conceptualization, writing, reviewing, and editing. CS contributed significantly to reviewing and editing. DM contributed to project planning, reviewing, and editing.

All authors have reviewed the manuscript and approved the submission.

AUTHOR AFFILIATIONS

Ge Peng *orcid.org/0000-0002-1986-9115*
Earth System Science Center/NASA MSFC IMPACT, The University of Alabama in Huntsville, Huntsville, AL, US
Carlo Lacagnina *orcid.org/0000-0001-9434-9809*
Barcelona Supercomputing Center (BSC), Barcelona, ES
Robert R. Downs *orcid.org/0000-0002-8595-5134*
Center for International Earth Science Information Network (CIESIN), Columbia University, Palisades, NY, US
Anette Ganske *orcid.org/0000-0003-1043-4964*
TIB – Leibniz Information Centre for Science and Technology, Hannover, DE
Hampapuram K. Ramapriyan *orcid.org/0000-0002-8425-8943*
Science Systems and Applications, Inc., Lanham, MD, USA and NASA Goddard Space Flight Center, Greenbelt, MD, US
Ivana Ivánová *orcid.org/0000-0001-6836-3463*
Curtin University, Perth, AU
Lesley Wyborn *orcid.org/0000-0001-5976-4943*
National Computational Infrastructure, Australian National University, ACT, AU
Dave Jones *orcid.org/0000-0003-4573-2400*
StormCenter Communications | GeoCollaborate, Halethorpe, MD, US
Lucy Bastin *orcid.org/0000-0003-1321-0800*
Aston University, Birmingham, UK
Chung-lin Shie *orcid.org/0000-0002-1115-1029*
University of Maryland at Baltimore County, Baltimore, MD, USA and NASA Goddard Space Flight Center, Greenbelt, MD, US
David F. Moroni *orcid.org/0000-0003-2994-557X*
David F. Moroni, Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA

REFERENCES

Albani, M and **Maggio, I.** 2020. CEOS WGISS Data Management and Stewardship Maturity Matrix and Application at ESA. *Pre-ESIP Workshop*, 13 July 2020, Virtual. DOI: *https://doi.org/10.6084/m9.figshare.12711896*

Baker, KS, Duerr, RE and **Parsons, MA.** 2016. Scientific knowledge mobilization: Co-evolution of data products and designated communities. *International Journal of Digital Curation*, 10(2): 110–135. DOI: *https://doi.org/10.2218/ijdc.v10i2.346*

Bastin, L, Cornford, D, Jones, R, Heuvelink, GBM, Pebesma, E, Stasch, C, Nativi, S, Mazzetti, P, and **Williams, M.** 2013. Managing uncertainty in integrated environmental modelling: The UncertWeb framework. *Environmental Modelling and Software*, 39: 116–134. DOI: *https://doi.org/10.1016/j.envsoft.2012.02.008*

Bugbee, K, le Roux, J, Sisco, A, Kaulfus, A, Staton, P, Woods, C, Dixon, V, Lynnes, C and **Ramachandran, R.** 2021. Improving discovery and use of NASA's Earth observation data through metadata quality assessments. *Data Science Journal,* 20(1): 17. DOI: *https://doi.org/10.5334/dsj-2021-017*

Callahan, T, Barnard, J, Helmkamp, L, Maertens, J, and **Kahn, M.** 2017. Reporting data quality assessment results: Identifying individual and organizational barriers and solutions. *eGEMs*, 5(1). DOI: *https://doi.org/10.5334/egems.214*

Cordy, CE and **Coryea, LR.** 2006. Champion's Practical Six Sigma Summary. Version: 27 January 2006. Xlibris Corporation. 65 pp. ISBN 978-1-4134-9681-9

CoreTrustSeal. 2019. Core Trustworthy Data Repository Requirements 2020–2022 – Extended Guidance. Version 2.0 November 2019. *Zenodo*. *https://zenodo.org/record/3638211#.YCfqv89Ki7M*.

Cosoli, S and **Grcic, B.** 2019. Quality control procedures for IMOS Ocean Radar Manual Version 2.1. *Integrated Marine Observing System*. DOI: *https://doi.org/10.26198/5c89b59a931cb*

Cowley, R. 2021. Report on the quality control of the IMOS East Australian Current (EAC) deep water moorings array. Deployed: April/May 2018 to September, 2019. Version 1.1. Hobart, Australia: CSIRO Oceans and Atmosphere, 56pp. DOI: *https://doi.org/10.26198/5r16-xf23*

Davies, C and **Sommerville, E** (eds) 2020. National Reference Stations Biogeochemical Operations Manual Version 3.3.1. Hobart, Australia. *Integrated Marine Observing System*, 66pp. DOI: *https://doi.org/10.26198/5c4a56f2a8ae3*

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

Downs, R, Moroni, C, Peng, G, Ramapriyan, HK and **Wei, Y.** 2021. Documentation to Foster Sharing and Use of Open Earth Science Data: Quality Information. *International Digital Curation Conference (IDCC)*, 19 April 2021. Virtual. *Zenodo*. DOI: *http://doi.org/10.5281/zenodo.4701361*

Downs, RR. 2020. GEOSS Data Management & Data Sharing Principles and TRUST – Implications for Information Quality. *Pre-ESIP Workshop*, 13 July 2020, Virtual. DOI : *https://doi.org/10.6084/m9.figshare.12711989*

Drévillon, M, García-Hermosa, I, Sotillo, MG, Régnier, C and **the CMEMS Product Quality Working Group.** 2020. Production of Quality Information at the Copernicus Marine Environment Monitoring Service (CMEMS). *ESIP Summer Meeting*, 22 July 2020, Virtual. DOI: *https://doi.org/10.6084/m9.figshare.12721592*

Figgemeier, H, Henzen, C and **Rümmler, A.** 2021. A geo-dashboard concept for the interactively linked visualization of provenance and data quality for geospatial datasets. *AGILE GIScience Ser.*, 2: 25. DOI: *https://doi.org/10.5194/agile-giss-2-25-2021*

Goldberg, M and **Zhou, L.** 2020. JPSS & Product Algorithm Maturity Matrix and Application. *Pre-ESIP Workshop*, 13 July 2020, Virtual. DOI: *https://doi.org/10.6084/m9.figshare.12711869*

Henzen, C, Della Chiesa, S and **Bernard, L.** 2021. Recommendations for Future Data Management Plans in Earth System Sciences. *AGILE GIScience Ser.*, 2: 31. DOI: *https://doi.org/10.5194/agile-giss-2-31-2021*

Heydebreck, D, Ganske, A, Kraft, A, Kaiser, A, Thiemann, H, Habermann, T and **Peng, G.** 2020. Maturity Indicator – potential extension to the DataCite Metadata Schema. *GitHub*. Version 7.1. Available at: *https://github.com/AtMoDat/maturity-indicator*.

Höck, H, Toussaint, F and **Thiemann, H.** 2020. Fitness for use of data objects described with quality maturity matrix at different phases of data production. *Data Science Journal*, 19(1): 45. DOI: *https://doi.org/10.5334/dsj-2020-045*

Hou, Y. 2020. Interpreting and Applying the FAIR Principle Checks. *ESIP Summer Meeting*, 22 July 2020, Virtual. DOI: *https://doi.org/10.6084/m9.figshare.12721628*

ISO 19115-1. 2014. Geographic Information—Metadata – Part 1: Fundamentals. Version: 2014–04. International Organization for Standardization. Geneva, Switzerland. Available at: *https://www.iso.org/standard/53798.html.*

ISO 19157. 2013. Geographic information – Data quality. Geneva, Switzerland. Available at: *https://www.iso.org/standard/32575.html.*

Ivánová, I, Peng, G, Lacagnina, C and **ODG Data Quality Domain Working Group.** 2021. OGC quality makes data FAIR workshop. 15 June 2021. Virtual. Presentations are available from: *https://www.ogc.org/projects/groups/dqdwg*.

Lacagnina, C, Peng, G, Downs, RR, Ramapriyan, H, Ivánová, I and others. 2021a. Towards Developing Community Guidelines for Sharing and Reusing Quality Information of Earth Science Datasets. *European Geosciences Union General Assembly*, 19–30 April 2021. EGU21–23, 27 April 2021. Virtual. DOI: *https://doi.org/10.5194/egusphere-egu21-23*

Lacagnina, C, Peng, G, Ivánová, I and **others.** 2021b. Global Community Effort on Sharing Dataset Quality Information. *OGC "Quality makes data FAIR' Workshop*, 15 June 2021. Virtual.

Lee, YW, Strong, DM, Khan, BK and **Wang, RY.** 2002. AIMQ: A methodology for information quality assessment. *Information & Management*, 40: 133–146. DOI: *https://doi.org/10.1016/S0378-7206(02)00043-5*

Lemieux, P, III, Peng, G and **Scott, DJ.** 2017. Data Stewardship Maturity Report for NOAA Climate Data Record (CDR) of Passive Microwave Sea Ice Concentration, Version 2. *Figshare*. DOI: *https://doi.org/10.6084/m9.figshare.5279932*

Lief, C, Wright, W, Peng, G, Baddour, O, Siegmund, P, Berod, D, Dunn, R, Cazenave, A and **Brunet, M.** 2020. The High-Quality Global Data Management Framework for Climate – Improving the Quality of Climate Data Management for Better Climate Monitoring. *Pre-ESIP Workshop*, 13 July 2020, Virtual. DOI: *https://doi.org/10.6084/m9.figshare.12712001*

Lin, D, Crabtree, J, Dillo, I, Downs, RR, Edmunds, R, Giaretta, D, De Giusiti, M, L'Hours, H, Hugo, W, Jenkyns, R, Khodiyar, V, Martone, M, Mokrane, M, Navale, V, Petters, J, Sierman, B, Sokolova, DV, Stockhause, M and **Westbrook, J.** 2020. The TRUST principles for digital repositories. *Scientific Data,* 7: 144. DOI: *https://doi.org/10.1038/s41597-020-0486-7*

Matthews, JL, Mannshardt, E and **Gremaud, P.** 2013. Uncertainty quantification for climate observations. *Bulletin of the American Meteorological Society,* 94: ES21–ES25. DOI: *https://doi.org/10.1175/BAMS-D-12-00042.1*

Moroni, DF, Ramapriyan, HK, Peng, G, Hobbs, J, Goldstein, JC, Downs, RR, Wolfe, R, Shie, C-L, Merchant, CJ, Bourassa, M, Matthews, JL, Cornillon, P, Bastin, L, Kehoe, K, Smith, B, Privette, JL, Subramanian, AC, Brown, O and **Ivánová, I.** 2019. Understanding the Various Perspectives of Earth Science Observational Data Uncertainty. *Figshare*. DOI: *https://doi.org/10.6084/m9.figshare.10271450*

Peng, G, Lacagnina, C, Downs, RR, Ivánová, I, Larnicol, G, Moroni, DF, Ramapriyan, H and **Wei, Y.** 2020a. Case Statement for Community Guidelines for FAIR Dataset Quality Information. *Figshare*. DOI: *https://doi.org/10.6084/m9.figshare.12605438*

Peng, G, Lacagnina, C, Downs, RR, Ivánová, I, Moroni, DF, Ramapriyan, H, Wei, Y and **Larnicol, G.** 2020b. 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. *Open Science Framework.* DOI: *https://doi.org/10.31219/osf.io/75b92*

Peng, G, Lacagnina, C, Downs, RR, Ramapriyan, Ivánová, I and **others.** 2020c. Towards Developing Community Guidelines for Sharing and Reuse of Digital Data Quality Information. IN012-04. *AGU Fall Meeting 2020*. Talk. Dec 8, 2020. Virtual.

Peng, G, Downs, RR, Lacagnina, C, Ramapriyan, H, Ivánová, I and **others.** 2021a. Call to action for global access to and harmonization of quality information of individual Earth science datasets. *Data Science Journal*, 20. DOI: *https://doi.org/10.5334/dsj-2021-019*

Peng, G, Lacagnina, C, Ivánová, I, Downs, RR, Ramapriyan, H, Ganske, A and **others.** 2021b. International Community Guidelines for Sharing and Reusing Quality Information of Individual Earth Science Datasets. *Open Science Framework*. DOI: *https://doi.org/10.31219/osf.io/xsu4p*

Peng, G and **the International FAIR-DQI Community Guidelines Working Group.** 2021c. Developing Community Guidelines for FAIR Dataset Quality Information. *OGC Data Quality Domain Working Group Meeting*, March 22, 2021. Virtual.

Peng, G, Downs, R, Ramapriyan, HK, Moroni, D and **Wei, Y.** 2021d. Introducing Community Guidelines for FAIR Dataset Quality Information. *AU/NZ Data Quality Interest Group-ESIP IQC Meeting*, March 31, 2021. Virtual.

Peng, G, Lacagnina, C, Ivánová, I and **others.** 2021e. Global Community Effort on Sharing Dataset Quality Information. *Barcelona Supercomputing Center Evaluation and Quality Control Workshop*. June 6, 2021. Virtual.

Peng, G and **the International FAIR-DQI Community Guidelines Working Group.** 2021f. FAIR Dataset Quality Information. October 19, 2021. *SciDataCon 2021*, Virtual.

Peng, G and **the International FAIR-DQI Community Guidelines Working Group.** 2021g. Developing Community Guidelines to Promote Global Access to and Harmonization of Quality Information of Individual Earth Science Datasets. October 21, 2021. *CEOS WGISS #52 Meeting*, Virtual.

Peng, G, Wyborn, L, Downs, RR, Ramapriyan, HK, Ivánová, I, Lacagnina, C and **Wu, M.** 2021h. Representing and Communicating Data Quality Information. Session. November 3, 3021. *RDA 18th Plenary*. Virtual. *https://www.rd-alliance.org/representing-and-communicating-data-quality-information*.

Peng, G, Lacagnina, C, Ivánová, I, Downs, RR, Ramapriyan, HK, Wyborn, L, Wu, M and **others.** 2021i. Making Dataset Quality Information FAIR – Guidelines for Sharing and Reusing Quality Information of Individual Earth Science Datasets. Nov 3, 2021. *RDA 18th Plenary*, Virtual.

Peng, G, Milan, A, Ritchey, N, Partee II, RP, Zinn, S, McQuinn, Lemieux, PE, III, Ionin, R, Collins, D, Jones, P, Jakositz, A, and **Casey, KS.** 2019a. Practical application of a stewardship maturity matrix for the NOAA OneStop Program. *Data Science Journal*, 18. DOI: *https://doi.org/10.5334/dsj-2019-041*

Peng, G, Privette, JL, Kearns, EJ, Ritchey, NA and **Ansari, S.** 2015. A unified framework for measuring stewardship practices applied to digital environmental datasets. *Data Science Journal*, 13: 231–253. DOI: *https://doi.org/10.2481/dsj.14-049*

Peng, G, Wright, W, Baddour, O, Lief, C and **the SMM-CD Work Group.** 2019b. The Guidance Booklet on the WMO-Wide Stewardship Maturity Matrix for Climate Data. *Figshare*. DOI: *https://doi.org/10.6084/m9.figshare.7002482*

Popp, T, Hegglin, MI, Hollmann, R, Ardhuin, F, Bartsch, A and **others.** 2020. Consistency of satellite climate data records for Earth system monitoring. *BAMS*. DOI: *https://doi.org/10.1175/BAMS-D-19-0127.1*

RDA FAIR Data Maturity Model Working Group. 2020. FAIR Data Maturity Model: specification and guidelines. DOI: *https://doi.org/10.15497/rda00050*

Ramapriyan, H, Downs, R, Peng, G and **Wei, Y.** 2021. The State of Documenting and Reporting Data and Information Quality for Supporting Open Science, Session 285. *SciDataCon 2021*, *https://www.scidatacon.org/virtual-2021/sessions/285/*

Ramapriyan, H, Peng, G, Moroni, D and **Shie, C-L.** 2017. Ensuring and improving information quality for Earth science data and products. *D-Lib Magazine*, 23. DOI: *https://doi.org/10.1045/july2017-ramapriyan*

Redman, CT. 1996. *Data quality of the information age.* Artech House, Boston. 303 pp.

Ritchey, N. 2020. NOAA/NCEI Data Stewardship Maturity Assessment. *Pre-ESIP Workshop*, 13 July 2020, Virtual. DOI: *https://doi.org/10.6084/m9.figshare.12711929*

Schulz, J. 2020. System Maturity and Application Performance for Climate Data Record. *Pre-ESIP Workshop*, 13 July 2020, Virtual. DOI: *https://doi.org/10.6084/m9.figshare.12711875*

W3C (World Wide Web Consortium). 2020. Data Catalog Vocabulary (DCAT), Version 2. Available at: *https://www.w3.org/TR/vocab-dcat-2/#Class:Dataset*.

Wagner, M, Henzen, C, and **Müller-Pfefferkorn, R.** 2021. A research data infrastructure component for the automated metadata and data quality extraction to foster the provision of FAIR data in Earth system sciences. *AGILE GIScience Ser.*, 2: 41. DOI: *https://doi.org/10.5194/agile-giss-2-41-2021*

Wang, RY and **Strong, DM.** 1996. Beyond accuracy: What data quality means to consumers. *Journal of Management Information Systems*, 12(4): 5. DOI: *https://doi.org/10.1080/07421222.1996.11518099*

Wei, Y, Moroni, D, Ramapriyan, H, Downs, RR, Liu, Z, Scott, D and **NASA ESDSWG.** 2020. NASA ESDSWG Data Quality Working Group. *Pre-ESIP Workshop,* 13 July 2020, Virtual. DOI: *https://doi.org/10.6084/m9.figshare.12711962*

Wenger-Trayner, E and **Wenger-Trayner, B.** 2015. Introduction to communities of practice. Available at: *https://wenger-trayner.com/introduction-to-communities-of-practice*.

Wilkinson, MD, Dumontier, M, Aalbersberg, IJ, Appleton, G, Axton, M, Baak, A and **others.** 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data,* 3: 160018.. DOI: *https://doi.org/10.1038/sdata.2016.18*

Woo, LM and **Gourcuff, C.** 2021. Delayed Mode QA/QC Best Practice Manual Version 3.0. *Integrated Marine Observing System*. DOI: *https://doi.org/10.26198/5c997b5fdc9bd*

Wu, F, Cornillon, P, Boussidi, B and **Guan, L.** 2017. Determining the pixel-to-pixel uncertainty in satellite-derived SST fields. *Journal of Remote Sensing*, 9(9). DOI: *https://doi.org/10.3390/rs9090877*

Wyborn, L, Wu, M, Ivánová, I, Bastrakova, I, Peng, G, Wei, Y, Moroni, D and **Downs, RR.** 2021. International Efforts to Develop Community Guidelines for FAIR Quality Information of Earth Science Datasets. *2021 Australia Collaborative Conference on Computational & Data Intensive Science (C3DIS)*. 5–9 July 2021. Virtual.

Zhou, L, Divakarla, M and **Liu, X.** 2016. An overview of the joint polar satellite system (JPSS) science data product calibration and validation. *Remote Sensing*, 8. DOI: *https://doi.org/10.3390/rs8020139*

TO CITE THIS ARTICLE:
Peng, G, Lacagnina, C, Downs, RR, Ganske, A, Ramapriyan, HK, Ivánová, I, Wyborn, L, Jones, D, Bastin, L, Shie, C-L and Moroni, DF. 2022. Global Community Guidelines for Documenting, Sharing, and Reusing Quality Information of Individual Digital Datasets. *Data Science Journal*, 21: 8, pp. 1–20. DOI: *https://doi.org/10.5334/dsj-2022-008*

Submitted: 03 January 2022
Accepted: 28 February 2022
Published: 31 March 2022

Data Science Journal is a peer-reviewed open access journal published by Ubiquity Press.