International Community Guidelines for Sharing and Reusing Quality Information of Individual Earth Science Datasets

International FAIR-DQI Community Guidelines Working Group[1]

Ge Peng, Carlo Lacagnina, Ivana Ivánová, Robert R. Downs, Hampapuram 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

Document ID: FAIR-DQI_Guidelines
Version: v01r02-20220326

Document History

Version	Contributors	What Is New
v0r05-20210417	Members of International FAIR-DQI Community Guidelines Working Group (FAIR-DQI WG)	First complete draft of the FAIR dataset quality information community guidelines document (FAIR-DQI Guidelines) for community review.
v01r00-20211001	Members of the FAIR-DQI WG	First baseline of the FAIR-DQI Guidelines document. Changes were made to address edits and comments provided by reviewers from the community and working group members. Review comments and responses are captured in Appendix I.
V01r01-20220316; v01r02-20220326	Ge Peng	Implemented comments from the FAIR-DQI WG members including minor editorial edits and additional examples for the guidelines 1-5. Detailed FAIR-DQI WG member contributions. Impacted: Section 4 and Acknowledgement.

[1] A list of author names, affiliations, sectors, roles and/or subject areas, and ORCIDs can be found in Appendix G

Disclaimer

This document facilitates the development, update, and effective use of the international community guidelines for promoting the representation and sharing of quality information by enabling global access to and harmonization of quality information at the level of individual Earth science datasets. The document is provided without any representations or warranties, express or implied. The views expressed herein are those of the authors and do not necessarily reflect that of the Earth Science Information Partners (ESIP) and the affiliated organizations of the authors. The use cases and examples are provided for references with no endorsement or preference intended. The document is subjected to future evolution without notification as knowledge improves and community requirements expand and/or change.

Feedback and Maintenance of the Document

This document is a living document. Please provide your comments and suggestions to: Ge Peng at gpeng93@gmail.com; Carlo Lacagnina at carlo.lacagnina@bsc.es, or Ivana Ivánová at ivana.ivanova@curtin.edu.au. The latest version can be downloaded from the Open Science Framework (osf.io) with the following persistent digital object identifier (DOI): https://doi.org/10.31219/osf.io/xsu4p. The document along with additional resources are also maintained at https://wiki.esipfed.org/FAIR_Dataset_Quality_Information

Recommended Citation for This Document

Peng, G., C. Lacagnina, I. Ivánová, R. R. Downs, H. Ramapriyan, A. Ganske, D. Jones, L. Bastin, L. Wyborn, I. Bastrakova, M. Wu, Chung-Lin Shie, D. Moroni, G. Larnicol, Y. Wei, N. Ritchey, S. Champion, C. Hou, T. Habermann, G. Berg-Cross, K. Bugbee, and J. le Roux, and International FAIR-DQI Community Guidelines Working Group, 2021: International Community Guidelines for Sharing and Reusing Quality Information of Individual Earth Science Datasets. Document ID: FAIR-DQI-Guidelines. Updated: 2022. Version: v01r02 20220326. *Open Science Framework*. DOI: https://doi.org/10.31219/osf.io/xsu4p

TABLE OF CONTENTS

1. EXECUTIVE SUMMARY

Under the auspices of the Earth Science Information Partners (ESIP) and with collaboration among members of the 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), a community effort has been undertaken by international Earth Science domain experts. The objective of this effort is to develop global community guidelines with practical recommendations to promote the representation, sharing and reuse of quality information at the dataset level, leveraging the experiences and expertise of a team of interdisciplinary domain experts and community best practices. The community guidelines are inspired by the guiding principles of findability, accessibility, interoperability, and reusability (FAIR) and aim to help stakeholders such as science data centers, repositories, data producers and publishers, data managers and stewards, etc., i) to capture, describe, and represent quality information of their datasets in a way that is in line with the FAIR guiding principles; ii) to allow for the maximum discovery, trust, sharing, reuse and value of their datasets; and iii) to enable global access to and integration of dataset quality information. The vision of developing these guidelines is to promote the creation and use of freely and openly shared dataset quality information that is consistently described, readily available in community standardized formats, and capable of being integrated across commonly-used Earth science systems and tools for search and access with explicitly expressed usage licenses.

2. BACKGROUND

Knowledge about the quality of data and metadata is important to support informed decisions on the (re)use of individual datasets and is an essential part of the ecosystem that supports open science. Quality assessments reflect the reliability and usability of data and need to be consistently curated, fully traceable, and adequately documented, as these are crucial for sound decision- and policy-making efforts that rely on data. Quality assessments also need to be consistently represented and readily integrated across systems and tools to allow for improved sharing of quality information at the dataset level (Henzen et al. 2021, Wagner et al. 2021; Peng et al. 2021) for individual quality attributes, aspects, or dimensions such as those defined in Wang and Strong (1996) and Ramapriyan et al. (2017).[2] The need to improve data quality information also extends to crowdsourcing data products such as those from citizen science projects (Downs et al. 2021), which have become increasingly important for augmenting traditional Earth and geospatial science for a wide range of research and applications.

Although the need for assessing the quality of data and associated information at the individual dataset level is well recognized, methodologies for an evaluation framework and presentation of resultant quality information to end users (e.g., Figgemeier et al. 2021) may not have been comprehensively addressed within and across disciplines. Global interdisciplinary domain experts have therefore come together in a workshop to systematically explore needs, challenges and impacts of consistently curating and representing quality information through the entire lifecycle of a dataset. The outcomes of the workshop were reported in Peng et al. (2020). A call-to-action statement paper (Peng et al. 2021) has been published by the Committee on Data of the International Science Council (CODATA) Data Science Journal in response to the CODATA's call for a special open science collection: Open Science for a Global Transformation.[3]

Motivated by the needs and interest from the global Earth science community, a working group was formed as the result of several open calls to the community, especially to the ESIP community, on various occasions. The current members of the working group consist of international domain experts such as data producers, publishers, managers or stewards from national science and/or data centers, domain or institutional repositories, as well as data users from the academic and private sectors. Collectively, they bring together many years of valuable experience in production, management, services, and applications of various types of Earth science data including satellite, in situ, and model data, along with knowledge of the challenges and best practices in their domains.

Since September 2020, the working group members have been working collaboratively to develop practical guidelines to curate, represent, and report dataset quality information in a manner that is consistent with the FAIR guiding principles of being findable, accessible, interoperable, and reusable as defined in Wilkinson et al. (2016). These guidelines build on the success of the FAIR guiding principles on data sharing and the extensive expert knowledge and working experiences of the working group members, leveraging community best practices. This document captures the outcomes of this international community effort.

---

[2] Additional information on data quality attributes, aspects, and dimensions can be found in Appendix C.

[3] http://codata.org/blog/2020/10/28/open-science-for-a-global-transformation-call-for-papers-for-a-special-collection-in-data-science-journal/

The document is organized as follows. An executive summary has been provided in Section 1. A brief background is provided in this section. The scope, rationale, goals, and intended audiences for the community guidelines are described in Section 3. Guidelines developed are quality-attribute agnostic and assessment-type agnostic. There are needs for a full lifecycle approach and common terminology with controlled vocabulary. Additional information about those points, basic elements to consider, and the guidelines are provided in Section 4. Summary and Discussion are presented in Section 5. Followed by acknowledgement and references. Additional information such as key terms and definitions, the FAIR data guiding principles, explanation, and Earth sciences community implementation examples, quality attributes and dimensions, quality assessment types, and additional examples of quality assessment models, as well as acronyms used in this document are included in Appendices.

3. SCOPE, RATIONALE, GOALS, AND INTENDED AUDIENCES

3a. Scope

A dataset in this document refers to an identifiable collection of data (ISO 19115-1 2014), and it can be published or curated by a single agent (W3C 2020). A dataset can be the digital rendition of a data product of a given version of an algorithm, model, experiment, or observations (descriptions of key terms used in this document can be found in Appendix A). Here we focus on dataset quality, not just on data quality, because we consider information about quality or the state of data (input and output), metadata/documentation, software, procedures, processes, and infrastructure throughout the entire lifecycle of a dataset. A dataset lifecycle in this document starts during the planning and design stage of a data product after data collection (Figure 1). Therefore, it will not touch on sensor algorithms or model development and deployment. However, it is important to note that quality information during these development and implementation stages is important to be captured and described as well, because they are critical to estimating data product uncertainty and error progression to downstream applications (e.g., Matthews et al. 2013).

The dataset lifecycle shown in Figure 1 (the outermost circle) includes three important stages for each one of the four quality aspects defined in Ramapriyan et al. (2017) (the second outermost circle), which are represented by the blue circle.[4]) Those dataset lifecycle stages are not necessarily linear or in sequence and may also occur in more than one quality aspect. The feedback and improvement cycle can occur in any one of the stages. One may turn to Peng et al. (2018) for a conceptual framework for managing scientific data stewardship activities, utilizing the Plan-Do-Check-Act (PDCA) cycle (Deming 1986).

[4] Additional information on data quality attributes, aspects, and dimensions can be found in Appendix C.

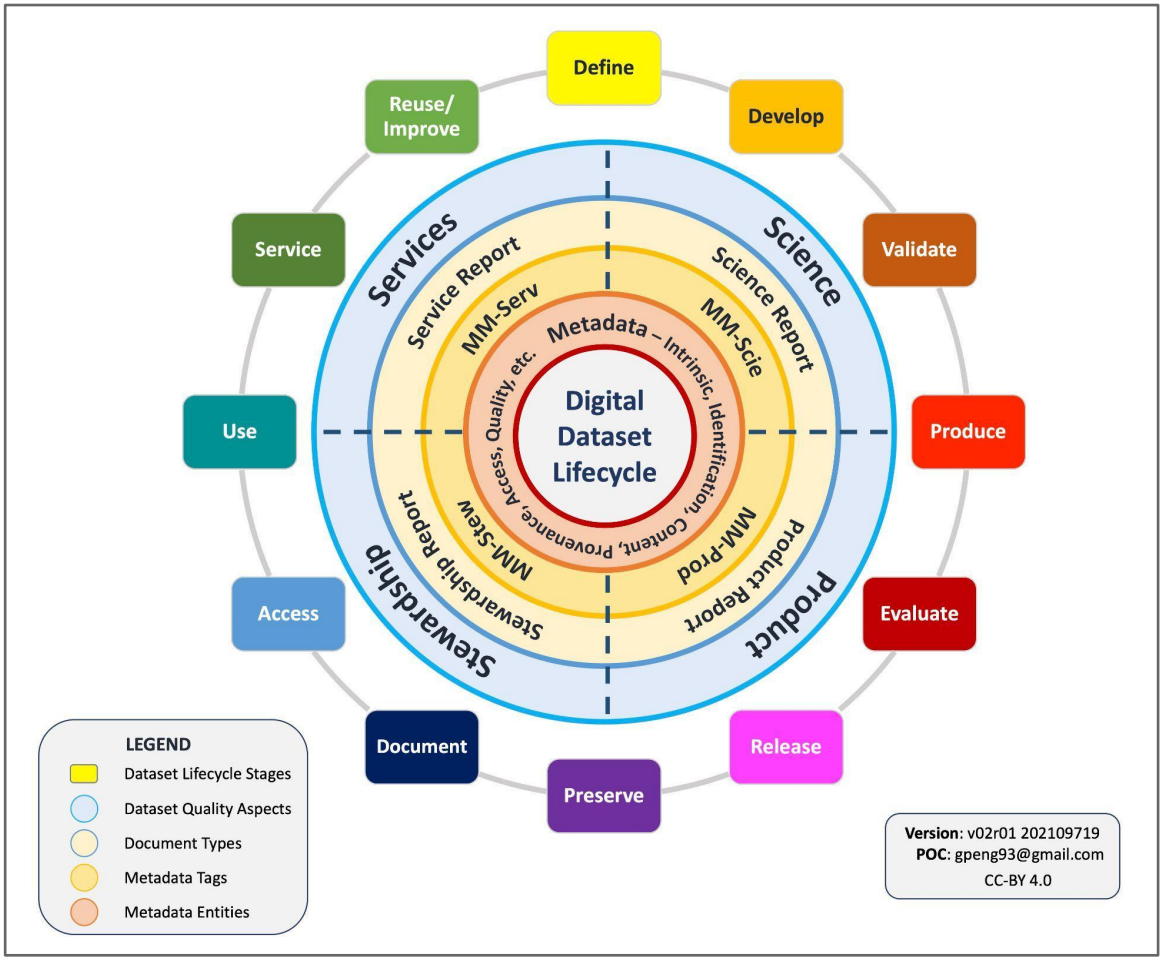

Figure 1: A schematic diagram of dataset lifecycle stages, quality aspects and associated documentation types and metadata tags (MM-*), and metadata entities. See Table C1 for a description of quality aspects and associated documentation types. Creator: Ge Peng; Contributors: Anette Ganske, Lesley Wyborn, and Mingfang Wu.

3b. Rationale

Quality information at the individual dataset level (hereinafter referred to as dataset quality information) such as accuracy, completeness, timeliness, and provenance, is imperative for establishing trust and enabling accurate use and reuse of data (Digital Science et al. 2019). To support effective decision- and policy-making processes, dataset quality information, such as information about the state of data, metadata, documentation, software, workflows, and tools used for producing and managing the data; and how the data being curated and serviced, should be consistently captured in the metadata and be a part of the ecosystem that supports open science (e.g., WMO 2019; Henzen et al. 2021; RDA Discipline-specific Guidance for Data Management Plans Working Group;[5] WMO expert team on metadata standards [6]).

[5] https://www.rd-alliance.org/groups/discipline-specific-guidance-data-management-plans-wg

[6] https://community.wmo.int/governance/commission-membership/commission-observation-infrastructures-and-information-systems-infcom/commission-infrastructure-officers/infcom-management-group/sc-imt/ET-Metadata

Assessment of data quality is key for ensuring that the available data and information are credible and such assessments are essential when establishing trust for reuse of the data (Callahan 2017). Trusted data are perceived as worthy of use in decision making environments where the metadata is sufficient to adequately describe the data, e.g., information about the dataset author and data timeliness. Describing the quality of a data product and providing access to such quality information can support potential users of a particular dataset to determine whether it is appropriate for their planned usage, i.e., fitness for purpose.

Peng et al. (2021) describes a real-life use case of how trusted, timely and readily integrable data and quality information is critical to disaster responses by utility companies with billions of dollars at stake. For disaster response managers, any information needs to be trusted and readily integrated, and understood in layman's terms in a matter of a minute. Accuracy and timeliness of data and information is extremely important. Any data that are selected to be a part of their decision-making processes need to be trusted. Managers will not trust just any available datasets as their decisions can have an impact on the safety and survival of at-risk populations, can cost up to millions of dollars, and influence the reputation of their organizations.

For this use case, datasets are pre-vetted with an operational readiness level (ORL) ranking that is readily available and easily understood by decision makers who are generally non-data experts. An assigned ORL leads such decision makers to rapidly trust datasets. Data and information are integrated into a system to provide an easy-to-understand dashboard to disaster response managers to allow them to make decisions promptly. Thus, providing quality information with the data establishes the trust needed for supporting such potentially life-saving emergency response activities and maximizes the benefit of sharing data.

Pre-vetting datasets and developing the dashboard requires years of work and ongoing effort in addition to cultivated human relationships. Readily available and consistently curated quality information of an individual dataset will help improve the process of establishing trust necessary to support tools and services provided to disaster preparation and response efforts, saving time and money. It will also support effective (re)use of the dataset for other applications, resulting in wide community utilization and therefore maximizing the value of the dataset.

To estimate the annual cost of not sharing data, a systematic analysis was carried out under the European Commission. The cost of not sharing data was found to be a minimum of €10.2bn per year (European Commission and PwC EU Services 2018). On average, data scientists spent 60-70% of their effort on dealing with data quality related issues (e.g., Press 2016). Thus, not sharing data quality information will compound that loss, especially productivity loss due to redundancy in assessing data quality or having to reject the data in the absence of quality information.

For dataset quality information to be effectively (re)used, it needs to be consistently curated, fully traceable, adequately documented, updated timely, able to support users to address their specific needs, and where possible, machine readable. This is, however, a daunting objective because it necessitates both a wide range of data quality attributes and heuristic information to ascertain fitness for purpose, while facing challenges in cross-disciplinary (and even in-discipline specific) knowledge integration (Peng et al. 2020).

It also needs to be considered throughout the whole dataset life cycle, from creation to archiving to servicing, and covering multiple dimensions (e.g., product generation, scientific content curation, data stewardship, services used to access the data) (Peng et al. 2021; see additional discussion in section 4d).

However, dataset quality information is not routinely curated, albeit some efforts (Figgemeier et al. 2021, Wagner et al. 2021) have been recently made, and much less represented in a human- and machine-readable manner, despite the fact that international standards or vocabularies for describing the quality of geographic data have been in place since 2003 (e.g., ISO 19157: 2013; ISO 19115-1:2014; W3C 2016). The lack of adoption of one or more data quality standards may in part reflect the diversity of approaches, availability of resources, technologies, networks, and research questions of investigators (Leonelli 2017), as well as the context for the planned purpose and use of the data (Canali 2020; Illari 2014). Lack of motivation to document quality can be caused by the lack of prescriptiveness of existing standards - documentation of data quality metadata has always been optional in the ISO 19100 series and as of 2014, the ISO 19115-1 standard for metadata does not define a minimum set of discovery metadata. Several other issues also may contribute to challenges for assessing and reporting data quality, and ultimately for the curation of dataset quality information. A frequently cited barrier against documenting the quality of spatial data is that it mostly requires special domain-expert technical knowledge, while documenting general metadata can be done automatically or by non-specialists (Coetzee 2018, Wagner et al. 2021). Therefore, practical community guidelines on how to curate, report and disseminate dataset quality information in a way that enables sharing and harmonizing the information across Earth science communities are needed and beneficial.

The FAIR data guiding principles defined by Wilkinson et al. (2016) emphasize the importance of data sharing in a machine-friendly environment. Since their inception, the FAIR data guiding principles have been adopted by international entities and have had a major impact in promoting data sharing and reuse globally (e.g., G20 Leaders 2016; Australia FAIR Access Working Group 2017; European Commission 2018; 2020; Mons 2018; U.S. Public Law 115-435 2019; CODATA 2019). See Appendix B for the definition of the FAIR data guiding principles and implementation examples in the Earth science community.

However, the FAIR data guiding principles are somewhat limited in that they call for only meta(data) to be associated with detailed provenance and “richly described with a plurality of accurate and relevant attributes” (Wilkinson et al. 2016). They do not explicitly address the sharing of meta(data) quality information. For example, if the FAIRness of a dataset has been evaluated, what can be done to ensure the method used and assessment results are readily findable, accessible and (re)usable to end users? What can be done to ensure that the information about the method and assessment results can be readily integrated across different tools and systems within and out of individual organizations?

Therefore, building on the direction that the FAIR guiding principles have provided for data sharing, we would like to go one step further and set a stage for all dataset quality information to be FAIR to enable or improve the sharing of quality information of individual datasets.

3c. Goals

This effort aims to develop guidelines for the Earth science community, in collaboration with international domain experts on data and information quality. The primary goal of the guidelines is to offer the Earth science community actionable recommendations that can be adopted by a variety of stakeholders to consistently capture and represent dataset quality information. This should be done in a way that is in line with the FAIR guiding principles to improve its sharing and reuse with more targeted practicality. The optimal goal is to allow for global access to and harmonization of quality information of individual datasets as an important step towards open science in both machine- and human-friendly environments.

3d. Potential Impacts of the Guidelines

Availability of quality information can assist users when deciding whether and how to use a dataset. This is achieved by informing users about the selection of the methodology to be applied, including particular approaches, tools, and techniques for analyzing the data or for using the dataset in conjunction with other datasets, for example.

Improving reuse of data by providing access to quality information about a dataset primarily contributes to the **T**ransparency, **R**esponsibility, and **U**ser focus aspects of the TRUST Principles for Data Repositories. It could also contribute to its **S**ustainability, by enabling future use, and to the application of **T**echnology for using the data (see Lin et al. 2020 for a detailed description of the TRUST Principles). In particular, the disclosure of quality information at the dataset level helps those who developed and curated the data to achieve transparency, which in turn will improve the stewardship maturity of the dataset (Peng et al. 2015). Providing clear guidance facilitates understanding about the quality of data products and their corresponding metadata. Providing such information about a dataset is necessary for its correct use and, therefore, it is a fundamental responsibility of data producers and curators. Describing quality aspects of the data also provides a crucial service to support potential data users when they are deciding whether a particular data product could meet their needs.

FAIR dataset quality information can also help improve the sharing of the data in several ways. When data can be discovered based on information about certain quality attributes, the findability of the data is improved for users who need data that contain such attributes. Accessibility and usability of data is improved by describing issues and conditions that could affect the use of the data. Describing quality information in standardized formats, schemas, and terminology with controlled and even harmonized vocabularies, improves the interoperability of the data. The reusability of data is facilitated by describing limitations on use as well as appropriate and inappropriate uses and usage of the data.

3e. Intended Audiences of the guidelines

All data stakeholders may benefit from the community guidelines:

- *Data producers* will find these guidelines useful to ensure that, at the point of acquisition, they are capturing those critical attributes that will later be used to ascertain the quality of the data that they are capturing (e.g., uncertainty of location/measurements, instrument parameters, metadata attributes on the instrument used to acquire the data). The same applies to those producers that generate multiple products from raw measurements. Data

quality guidelines also will help data producers to clean up their data prior to submission to a repository and to document data quality and share data quality information.

- *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;
- *Journal editors and reviewers* may refer to the guidelines when assessing data that are associated with manuscripts under evaluation for potential publication;
- *Sponsors and funders* may find the guidelines helpful when reviewing data management plans in proposals for support of projects and programs that will be creating, curating, disseminating, and supporting the use of data. They will also find them useful at project closure phase when assessing the quality of the data produced against the initial project aims and data management plans; and
- *Data users/consumers* also 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.

4. FAIR DATASET QUALITY INFORMATION GUIDELINES

4a. Basic Elements to Consider When Curating Dataset Quality Information

Assessing dataset quality is a multi-dimensional problem. Despite the multi-dimensionality of quality, there are aspects that are common when assessing datasets. Knowledge about the common aspects may help to set the direction for the right approach to follow in each specific case. In this section, the basic elements are outlined for developing data quality management processes and curating dataset quality information.

Lee et al. (2002) presented how data quality management (DQM) may be organized into four phases: define, measure, analyze, and improve. Inspired by the quality evaluation procedures defined in ISO 19157 (2013), Six Sigma (e.g., Cordy et al. 2006) and to help organizations and data stewards address the challenge of where to start, we have developed a typical workflow (Figure 2). It highlights the basic ingredients and elements to be considered 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 to build trust between data providers and end-users and to increase data usability.[7]

As shown in Figure 2, the following two phases are needed prior to carrying out any assessment activity.

- **Quality specification** - Curating dataset quality information should start with defining what quality attribute(s), aspect, or dimension will be assessed and at which level of granularity (variable, ensemble member, model or algorithm), and which data and quality attribute should be prioritized. This step will need some profiling, i.e., an initial analysis of the available data to understand the challenges and the most critical issues to set priorities and figure out the best strategy forward.

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

- **Evaluation specification** - The next step is to identify or develop an approach or method to evaluate the identified quality attribute(s) or assess its maturity, for example, statistical analysis approach (Wu et al. 2017) or scientific maturity matrix (Zhou et al. 2016). This is the step where the framework for the evaluation is defined. It is important to describe in this step the identified quality attribute or dimension, the evaluation method used, and the protocols, standards and workflows applied (e.g., Wu et al. 2017; Zhou et al. 2016; Lemieux et al. 2017; Popp et al. 2020). 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.

- **Evaluation execution** – At this stage the actual assessments are performed based on the tools, approaches and priorities defined in the previous phases. While doing this, the assessments should be captured in structured, human- and machine-readable, and standard-based formats (e.g., Peng et al. 2019a; Heydebreck et al. 2020).
- **Quality dissemination** – The results of the assessments represent the core of the dataset quality information and need to be disseminated along with the data for the benefit of the end-users and data providers to enable effective data discovery and use. For reproducibility purposes, the operations performed to produce the quality information also are recommended to be published. In this step, the way that quality information is disseminated (e.g., metadata, web page) is implemented and put into practice.

The feedback from the users should be sought and harvested to improve the assessment provided.

- **Monitoring and improvement** – The feedback collected in the previous steps 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. This step is continuous during the assessment to dissemination steps and helps to improve the curation of quality information, thereby increasing the reliability of the data repository.

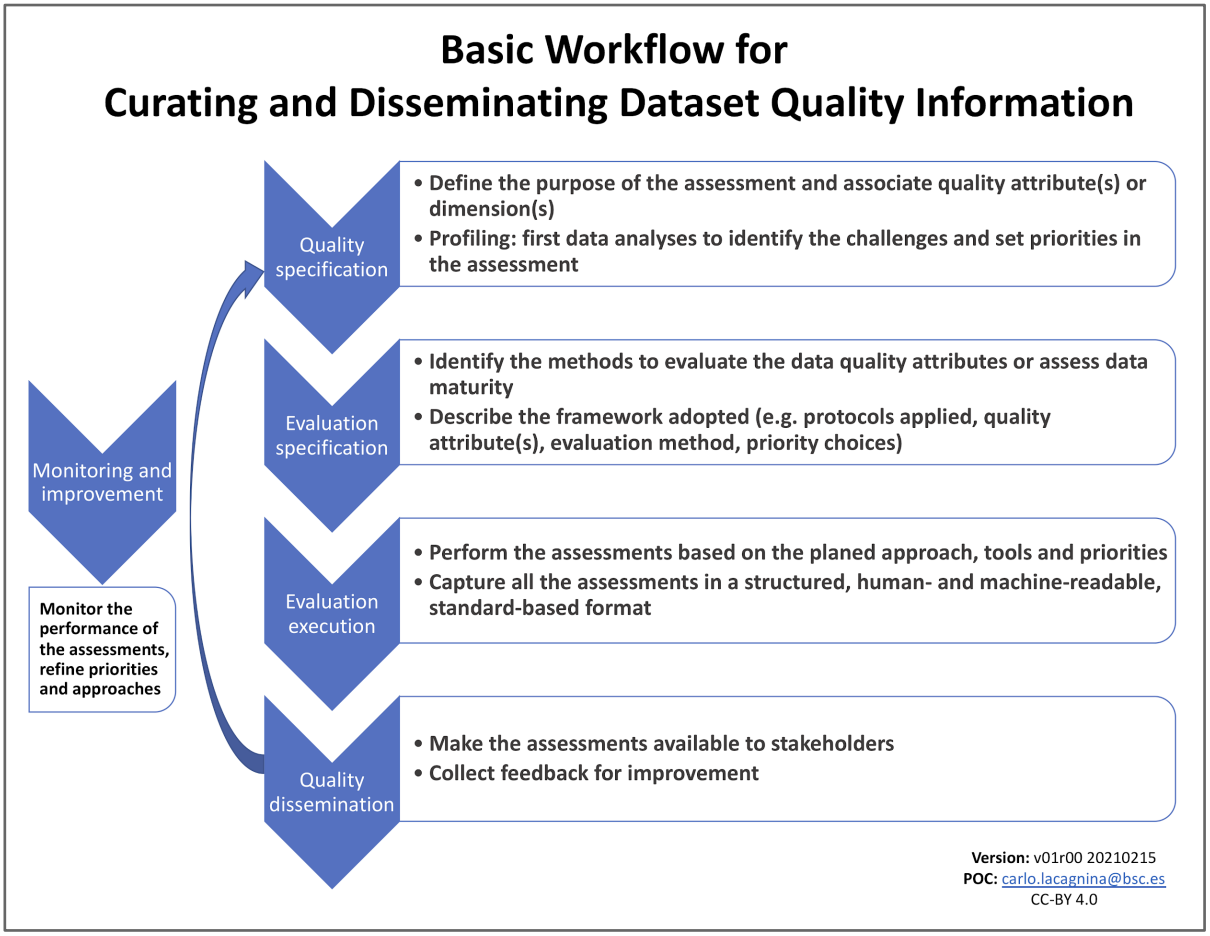

Figure 2: A schematic diagram of workflow and basic ingredients for curating and disseminating dataset quality information. Creator: Carlo Lacagnina; Contributor to conceptual design: Ge Peng.

4b. Quality-Attribute Agnostic Guidelines

As mentioned previously and in Appendix C, assessing dataset quality is a complex, multi-dimensional problem. The selection of the relevant attributes is context-dependent and it leads to different categorizations and practical dimensions (Redman 1996) (See Appendix C for an outline of relevant quality attributes, aspects, and dimensions). The complexity exists even for a single quality attribute within one discipline in terms of its definition and how to measure and represent it consistently. An example of this is on data uncertainty as explored by Moroni et al. (2019).

The selection of the relevant attributes is context-dependent because datasets are often crafted for specific designated communities. Traditionally, designated communities of data consumers are domain literate and have some familiarity with the scientific context, data generation, or intended data use. As discussed by Baker et al (2016), designated communities can change over time. With the increasing availability of data today, the existence of interested audiences representing a variety of scientific backgrounds outside the domain of data collection must be considered in order for scientific knowledge to be widely conveyed and understood by broader audiences.
Lessons learned from various data curation initiatives indicate that an alternative approach is to equip data consumers with readily available information that is consistently captured, in a way that can be easily understood by a variety of end-users and integrated across tools and systems,

regardless of the quality attribute and the assessment approach. How to tailor the dataset quality information to the designated community is left to the specific entities who serve that particular community. To support such a decision process, a description of what quality attribute or dimension is assessed and what assessment approach is utilized, should be captured in machine-readable quality metadata and/or a human-readable quality report in a consistent way for transparency and enhanced interoperability.

4c. Assessment-Type Agnostic Guidelines

Similar complexity exists for the types of assessments that one can carry out throughout the entire data lifecycle (see Appendix D for a comprehensive description of relevant assessment types). The assessments can be individual checks for different dimensions such as:

- Accuracy checks for data values, variable names, corresponding units of the variables, content of metadata elements, etc.;
- Completeness checks for data values, metadata entities, temporal and spatial coverage, etc.;
- Conformity checks for compliance to community standards for data format or variable standard names, metadata structure or syntax, keywords, etc.;
- Consistency checks for data values, data fixation (integrity), metadata syntax, data and metadata representation, etc.

Accuracy, completeness, conformity and consistency checks for data values are usually performed during the data production process (e.g., Durre et al. 2010). On the other hand, accuracy, completeness, conformity and consistency checks for metadata are usually performed during the data preservation stages, often associated with data management and stewardship (e.g., O'Brien et al. 2016, Stockhause et al. 2012). Automating such checks, when possible, can improve the efficiency of the data quality process.

4d. Full Dataset Lifecycle Approach

Dataset quality can be affected by activities that are conducted throughout the data lifecycle. For example, lack of a quality assurance and control (QA/QC) procedure when generating the data product may impact the scientific quality of the data, while lack of the information about the QA/QC procedure will influence the quality of metadata and potentially reduce the level of user's confidence in trusting and using the data. In addition, data can be corrupted at any stage of the data lifecycle. In order for users and decision-makers to trust the data and the scientific findings resulting from analysis of this data, it is essential to establish and demonstrate, in a consistent and transparent way, the credibility of not only the data itself, but also the whole process of producing, managing, stewarding, analyzing, and servicing the data (Tilmes et al. 2015a). Therefore, a dataset lifecycle approach to dataset quality assessment is necessary. Furthermore, a dataset lifecycle approach to dataset quality assessment can facilitate effective recording of dataset quality information during various dataset lifecycle activities. The details of such information could be lost during later stages in the dataset lifecycle if they are not recorded in a timely fashion when the dataset quality events or assessments occurred. Moreover, managing quality throughout the entire dataset lifecycle is imperative for ensuring that the information and knowledge gained are not contaminated by inaccurate or corrupt data, as well as for facilitating accurate uncertainty estimates in the derived analyses. The value of lifecycle approaches to data quality has been recognized for various kinds of data, including remote sensing observations (Barsi et al. 2019),

health services (Kahn et al. 2015), and health and biomedical citizen science (Borda et al. 2020). Data lifecycle approaches to quality assessment also could be informed by lifecycle approaches to software quality (Lenhardt et al. 2014).

4e. Common Terminology

Consistent terminology is essential for enabling interoperability. The lack of an overarching consistent dataset quality vocabulary is identified as one of the priority gaps in data quality (Nightingale et al. 2019). It is recommended to seek out and adopt existing common and controlled vocabularies and content standards for the Earth science community. Examples can be found in Appendix F.

Being semantically and syntactically consistent is another complex challenge facing various domains during the whole dataset quality information curation and representation workflow that is conducted throughout the entire dataset lifecycle. Semantic and syntactic consistency are critical for enabling interoperability across systems and tools within and across individual organizations. It is beyond the scope of this document to provide exhaustive research in this area and detailed guidance. Nevertheless, an example is given in Figure 1 to demonstrate how dataset lifecycle stages (the outermost circle) can be mapped to the four dataset quality aspects (shaded blue circle) and how each quality aspect can be matched with associated document (light yellow shaded circle) and metadata tags (MM-*, light orange shaded circle). Metadata is developed throughout the entire dataset lifecycle with core metadata entities listed (light red shaded circle). (Descriptions of the terms in Figure 1 can be found in Table C1 in Appendix C.) Figure 1 is intended to depict an example of a holistic and systematic view to potentially change the culture and the way that quality information evaluation and curation is approached. Taking such an approach can serve as a first step in moving towards community consensus for developing terminology and syntax to describe, report, and disseminate dataset quality information.

4f. Guidelines for Enabling FAIR Dataset Quality Information

The following guidelines are developed by the International FAIR-DQI Community Guidelines Working Group to enable curated dataset quality information to be FAIR, namely, findable, accessible, interoperable, and reusable, to both human and machine end users. Examples are provided for references with no endorsement or preference intended. Some of the examples are quite preliminary and still maturing as they represent the current state of our awareness. The list is not necessarily exhaustive, but can serve as a starting point to help interested parties to improve the data quality process and optimally lead to community convergence. Additional examples may be added when they are provided by the Earth science community.

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).

Examples:

- Ensemble Mean of CMIP5 TOS for the Period 1971 to 2000. http://doi.org/10.5281/zenodo.12843, v1, Bruno Combal. November 24, 2014. Zenodo, CC-BY 4.0 International.

- Airborne Sea Ice Plus Snow Thickness During the PAMARCMIP 2017 Aircraft Campaign in the Arctic Ocean. https://doi.org/10.1594/PANGAEA.924848, v1, Hendricks et al. 2020. PANGAEA. CC-BY 4.0 International.
- NOAA/NSIDC Climate Data Record of Passive Microwave Sea Ice Concentration, Version 3. Meier et al. 2017. Boulder, Colorado, USA: National Snow and Ice Data Center (NSIDC). doi: https://doi.org/10.7265/N59P2ZTG

Notes: It is recommended to use a consistent title for the dataset throughout the entire workflow, including that in the dataset-level metadata record and quality assessment report document. Sometimes it may be desirable to include a short name or unique identifier which may be used as an internal tag for the dataset and/or a part of naming convention for a quality assessment report document or rating diagrams, especially if they are generated in an automatic fashion.

The square brackets denote elements that are optional because the information may not always be available but it is strongly recommended to include the usage license with the data, if known. Generally speaking, information about the license or permissions associated with the data is necessary to determine whether the intended use of the data would be allowed.

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).

Examples:
- Data Stewardship Maturity Matrix (DSMM) (Peng et al. 2015) and Self-Evaluation DSMM Template (Peng 2014);
- WMO Stewardship Maturity Matrix for Climate Data (SMM-CD) (Peng et al. 2019b) and Self-Evaluation SMM-CD Template (Lief and Peng 2019);
- RDA FAIR Data Maturity Model Working Group (2020);
- Analysis and Review of NASA Common Metadata Repository (ARC; Bugbee et al. 2021). The ARC metadata curation dashboard tool is publicly available at: https://github.com/nasa/cmr-metadata-review;
- Delayed Mode QA/QC Best Practice Manual (Woo and Gourcuff 2021).

Additional examples of assessment models in a form of metric or matrix are available in Appendix E. Peng (2018) describes maturity assessment models associated with each of four dataset quality aspects defined by Ramapriyan et al. (2017).

Notes: If no suitable assessment model is available, a quality assessment model may need to be developed or adapted. In this case, conditions (2.1-2.3), above, should be satisfied to make the

model findable and accessible. Individual researchers can utilize the Registry of Research Data Repositories (re3data) at https://doi.org/10.17616/R3D to find appropriate repositories for their assessment models based on their particular requirements. A CoreTrustSeal certified repository demonstrates more matured organizational processes and capabilities in managing its holdings of digital objects (CoreTrust Seal 2019).

A published, peer-reviewed paper, with an associated DOI, that describes the model is preferred to ensure that the model has been assessed by the user community. Also, publishing the assessment model itself, with a DOI, in a community-based repository can facilitate discovery and persistent accessibility. Publishing a quality assessment model on a project website is not recommended, since such locations are not necessarily sustainable and often cannot guarantee persistent access. For example, a broken link can result from a system reorganization or migration, leading to an inaccessible 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.

Examples:

- DSMM OneStop ISO Quality Metadata Implementation (Peng et al. 2019a);
- AtMoDat Maturity Indicator (Heydebreck et al. 2020);
- Automated metadata and data quality extraction and Geo-dashboard concept (Wagner et al. 2021).

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

Furthermore, including a consistent metadata tag is recommended when developing a quality assessment metadata schema, if applicable. For example, Peng et al. (2019a) uses MM-Stew as a metadata tag to denote stewardship maturity assessment (Table 1). 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.

Examples:
- DSMM OneStop - Data Stewardship Maturity Report (DSMR; e.g., Sea Ice CDR (Lemieux et al. 2017));
- Quality Maturity Matrix used at DKRZ (Höck et al. 2020);
- An example of a report on how quality control was applied during the deployment of the IMOS East Australian Current (EAC) Deep Water moorings array (Cowley 2021).

Notes: The National Oceanic Atmospheric Administration (NOAA) National Centers for Environmental Information **(**NCEI) has developed workflows and tools to automate the processes of i) generating International Standards Organization (ISO) quality metadata to be integrated into the dataset-level metadata record, and ii) creating Data Stewardship Maturity Reports (DSMR) documents to be published by the NOAA Central Library with assigned DOIs (Peng et al. 2019a).

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.

Conveying dataset quality information in a manner that is easily understood and usable by data users is recommended along with providing a mechanism to obtain user feedback.

Examples:
- JPSS Data Product Algorithm Maturity Portal: https://www.star.nesdis.noaa.gov/jpss/AlgorithmMaturity.php
  (This portal also provides a timeline of algorithm maturity status (Beta, Provisional, Validated) and supporting documents for each data product with a defined and coordinated review process.)
- Copernicus Climate Change Service (C3S) Climate Data Store Dataset Quality Assessment Portal: https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels?tab=eqc
- NOAA *OneStop* Data Discovery Portal: https://data.noaa.gov/onestop/collections/details/ad83c4df-6f2d-447c-8b43-bafa0a91d10d
  (The average assessment rating is displayed as filled stars at the bottom of the page; click on the question icon next to the stars to see the rating value and a list of nine components of the assessment model along with an embedded link to the reference.)
- Rollingdeck to Repository (R2R) QA Dashboard: https://www.rvdata.us/qa_info

(R2R QA dashboard provides information on the status of quality assessment tests conducted on individual cruise datasets that are readily readable and actionable by machines. Figgemeier et al 2021 propose a Geo-dashboard to visualize quality information and related provenance information.)

- Fourth National Climate Assessment (NCA4) Metadata Viewer: https://nca2018.globalchange.gov/chapter/1/ (scroll to Figure 1.2, click eye icon for metadata record)
  (The NCA4 metadata viewer captures provenance of individual figures, for example, Figure 1.2 in NCA4. It not only has a significant amount of data and information underpinning the figure with 13 individual panels, but this figure also persists with each NCA. The original is traced back to a published version where panels are composed of time series, and where most recent versions are now rooted in the United States Global Change Research Program (USGCRP) Indicators Working Group.)

Notes: Organized reporting of dataset quality information is the most challenging area with diverse practices. The examples shown above demonstrate how different approaches can be applied when disseminating dataset quality information. The challenges for diverse practice also come from the dependencies of the audience for which this information is intended. Data users can provide feedback on which and how disseminated information is most relevant and what can be improved. Therefore, user engagement activities are very relevant at this stage, including prompt response to user requirements.[8] See Appendix D for additional discussion on user engagement and feedback.

4g. Map the Guidelines to the FAIR Guiding Principles

To provide an abstractive view of how each guideline is related to the FAIR principles, crosswalks between the guidelines and the FAIR principles are provided in Table 1.

Table 1: Crosswalks between the Guidelines to the FAIR Guiding Principles
(Creator: Anette Ganske; Contributor: Ge Peng.)

FAIR Dataset Quality Information Guideline	**Corresponding FAIR Principles** (Wilkinson et al. 2016; also see Table B1 of Appendix B)
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).	F*, R1
Guideline 2. Utilize a one- (or more) dimensional, structured quality assessment metric that is:	

[8] https://public.wmo.int/en/bulletin/what-do-we-mean-climate-services

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)	F1
2.2. registered or indexed in a searchable resource that supports authentication and authorization, such as Figshare, Zenodo, GitHub, and Dryad	F4, A1.2
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)	A1
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:	I3, R1
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)	I1, R1.3, F1, F4
3.2. includes a description of the quality attribute(s), aspect(s), or dimension(s) to be assessed	R1
3.3. includes a description of the assessment method and assessment model structure and version, and access date if applicable	R1
3.4. includes a description of the assessment results	R1
3.5. includes versioning and the history of the assessments	R1.2
Guideline 4. Describe comprehensively the assessment method, workflow, and results in at least a human-readable quality report that:	R1
4.1. preferably follows a template that is published and satisfies the conditions of Guideline 2 (i.e., 2.1–2.3)	F1, F4, A1, R1
4.2. is published with an explicit open license and history of the report, satisfying the conditions of Guideline 2	F1, F4, A1, R1.1, R1.2
4.3. links the report PID to the dataset-level metadata record	I3/F3
Guideline 5. Report/disseminate the dataset quality information in an organized way via a web interface with a comprehensive description of:	F2, A1, R1

5.1. the dataset according to the Guideline 1	F2, R1
5.2. assessed quality attribute(s)/aspect(s)/dimension(s)	F2, R1
5.3. the evaluation method and process including the review process, if applicable	F2, R1
5.4. how to understand and use the information	F2, R1

* The letter F, A, I or R in the right column denotes that the guideline from the left column can crosswalk to all criteria of being findable, accessible, interoperable, or reproducible, respectively, while the number (*n*) after the letter of F, A, I, or R refers to the *n*th criterion in that aspect of the FAIR.

4h. Additional Examples on Representing Assessment Results in Metadata

Here we present two methods for providing quality information in metadata. These are practical examples presented for illustrative purposes, with no specific associated endorsement included.

Adopting ISO 19115-1 (2014) data quality metadata standards, Peng et al. (2019a) has proposed and utilized the implementation practices and framework to curate and represent stewardship maturity of individual datasets assessed with a stewardship maturity matrix, which can be adapted for any quality aspect and potentially other metadata standards (Figure 3 and Table 2). The provenance of the assessment can be captured by using the history of assessments, e.g., original assessment date, modification date, along with a description of major changes associated with each version change if applicable. It is worth noting that Dataset Quality Dimension, DQ Attribute, and DQ Metadata in Figure 3 correspond to Data quality, Data quality element, and Metaquality in ISO 19157 (2013).

Another framework of representing data maturity ratings is also provided based on Heydebreck et al. (2020). This framework has been proposed for inclusion in DataCite's metadata standards (Table 3). The crosswalks between these two frameworks are captured in Table 4.

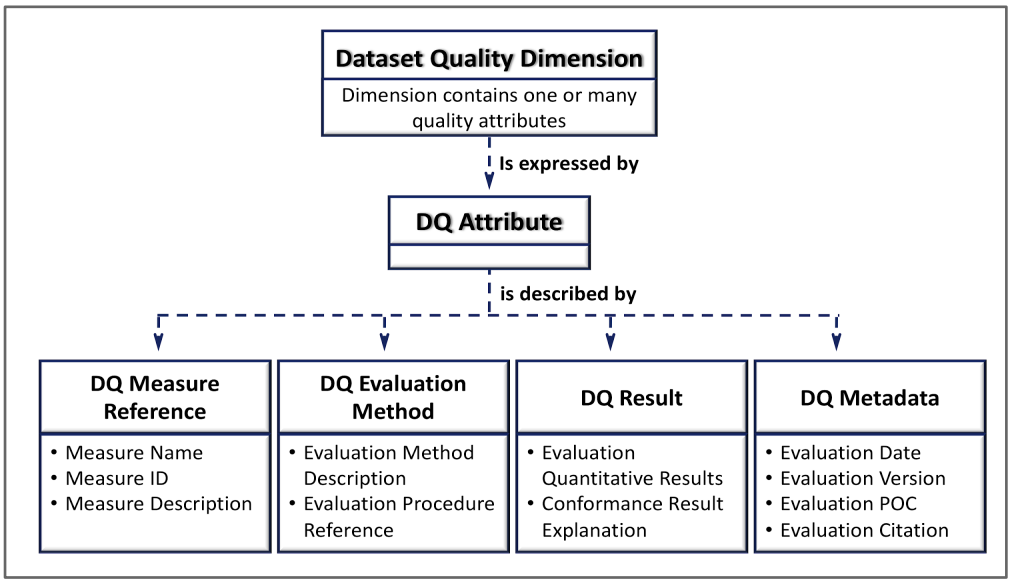

Figure 3: A structural diagram and practices for capturing and representing dataset quality information for a given quality dimension, attribute, evaluation method and results, along with evaluation metadata. Adapted from Figure 6 in Peng et al (2019a).

Table 2: Elements of a conceptual framework for representing dataset quality information in dataset-level quality metadata. An example is provided based on Peng et al. (2019a), which is associated with the stewardship quality aspect as shown in Figure 1. Please refer to Peng et al. (2015) for a description and scope of the data stewardship maturity matrix utilized.

Element	**Description**	**Example**
Dataset Quality Dimension[1]	Quality dimension or aspect of dataset	Stewardship
Dataset Quality Attribute[2]	Quality attribute under the dataset quality dimension or aspect	Accessibility
Measure Name	Name of assessment	Data Stewardship Maturity Assessment
Measure ID	Identifier of assessment	MM-Stew
Measure Description	Description of the measure utilized to assess the quality attribute and the rating system of assessment outcomes if applicable	The Data Stewardship Maturity Matrix (DSMM) is a unified framework that defines criteria for each of nine components based on measurable practices, which can be used to apply a progressive, 6-level rating to an individual dataset, representing stewardship maturity stages rated as Not Assessed or Not Available (Level 0), Ad Hoc (Level 1), Minimum (Level 2), Intermediate (Level 3), Advanced (Level 4), and Optimal (Level 5).
Evaluation Method Description	Description of how evaluation was carried out.	Data Stewardship Maturity Assessment was evaluated by the metadata content editor for the NOAA OneStop project using the Scientific Data Stewardship Maturity Assessment Model Template v4.0.
Evaluation Procedure Reference	Citation for the evaluation procedure, template, and tools.	Peng, Ge. The Scientific Data Stewardship Maturity Assessment Model Template. 2015-06-23. doi:10.6084/m9.figshare.1211954
Evaluation Date	Date of the evaluation being competed and results being summarized	2016-12-08
Evaluation Quantitative Results[2]	Quantitative results of assessment by the measure for	Advanced

	the attribute of the quality dimension or aspect	
Conformance Results Explanation	Explanation of the evaluation process and the rationale behind conformance evaluation result (e.g., comparison against quality requirements)	Data stewardship maturity assessment was carried out by NOAA OneStop metadata content editor, in collaboration with subject matter experts of the product and the maturity matrix.
Evaluation Version[3]	Current version of the evaluation results	v03r00
Evaluation Point-Of-Contact (POC)	An entity services as a POC for the evaluation	Ge Peng, gpeng@ncsu.edu [4]
Evaluation Citation	Citation of the evaluation report if applicable	Lemieux, P., G. Peng, and D.J. Scott, 2017: Data Stewardship Maturity Report for NOAA Climate Data Record (CDR) of Passive Microwave Sea Ice Concentration, Version 2. figshare, doi:10.6084/m9.figshare.5279932

[1] Dataset quality dimension or aspect contains one to many quality attributes.
[2] This element will be repeated for each quality attribute contained in the quality dimension or aspect.
[3] It is recommended to change the evaluation version only when the maturity ratings are modified. For real-time or dynamically changing data, it may be difficult to manually track the versioning.
[4] Now at ge.peng@uah.edu

Table 3: Elements of the suggested schema for representing data maturity ratings, which is proposed to DataCite to be included in the DataCite metadata schema (Based on Heydebreck et al. 2020).

Element	**Definition**	**Example: NOAA-DSMM***
MaturityCheck		Data Stewardship Maturity Matrix (MM-Stew)
maturityCheckSchemaVersion	Version of this schema	NCDC-CICS-SMM_0001_Rev.1 12/09/2014
maturityCheckName	Name of the maturity check	Data Stewardship Maturity Assessment
maturityCheckDescription	Description of the maturity check.	The Data Stewardship Maturity Matrix (DSMM) is a unified framework that defines criteria for each of nine components based on measurable

		practices, which can be used to apply a progressive, 6-level rating to an individual dataset, representing stewardship maturity stages rated as Not Assessed or Not Available (Level 0), Ad Hoc (Level 1), Minimum (Level 2), Intermediate (Level 3), Advanced (Level 4), and Optimal (Level 5).
maturityCheckResourceType	Type of the resource	Web Questionnaire; Manual
maturityCheckIdentifier	PID of the metric definition	https://doi.org/10.6084/m9.figshare.1211954
maturityCheckVersion	Version of the maturity check	v03r00
maturityCheckPerformedBy	Information on who performed the maturity check	Ge Peng
maturityCheckReport	Provide result report for the check	Lemieux, P., G. Peng, and D.J. Scott, 2017: Data Stewardship Maturity Report for NOAA Climate Data Record (CDR) of Passive Microwave Sea Ice Concentration, Version 2. figshare, doi:10.6084/m9.figshare.5279932
ReportDate	Date when the result was produced	2016-12-08
MetricName	MetricName	Usability
MetricResult	Results of the metric	Advanced
Unit	unit of the result	Level 5 of 6

Table 4: Crosswalks of the Elements in Tables 2 and 3.

Elements from Heydebreck et al. (2020)	***Elements from Peng et al. (2019a)***
MaturityCheck	Measure ID
maturityCheckSchemaVersion	
maturityCheckName	Measure Name
maturityCheckDescription	Measure Description
maturityCheckResourceType	

maturityCheckIdentifier	Evaluation Procedure Reference
maturityCheckVersion	Evaluation Version
maturityCheckPerformedBy	Evaluation Point-Of-Contact (POC)
maturityCheckReport	Evaluation Citation
ReportDate	Evaluation Date
MetricName	Dataset Quality Attribute
MetricResult	Evaluation Quantitative Results
Unit	

5. CONCLUSIONS AND DISCUSSION

The FAIR guiding principles defined by Wilkinson et al. (2016) have provided an effective way to enable data sharing. Inspired by the FAIR guiding principles for curating and reporting dataset quality information, these guidelines are being developed, iteratively, through a community effort by leveraging the experiences and expertise of an international team of interdisciplinary domain experts who are engaged in aspects of data quality and recommended community practices. The guidelines are aimed to improve the availability and sharing of quality information at the individual dataset level.

Prioritizing the collection, representation, and sharing of data quality information is necessary to improve the scientific process. Utilizing a structured quality assessment model helps ensure the consistency of evaluation methods and results, which in turn will make it easier to capture them systematically. 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.

The guidelines developed as a result of this international community effort bring the Earth science community one step closer to standardizing the curation and representation of dataset quality information. These guidelines presented here offer opportunities to enable or improve the transparency and interoperability of dataset quality information. Adopting the guidelines can contribute to the evolving ecosystem that supports open 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 global access to and harmonization of quality information of individual digital datasets.

Making dataset quality information FAIR also helps to improve the overall FAIRness of a dataset by providing standard community-based rich metadata with relevant quality attributes and qualified references. Providing such quality information with each dataset will also help establish the trustworthiness of data and ultimately improve the maturity of such datasets in multiple quality

dimensions or aspects including product, stewardship, and services by improving the completeness and usability of metadata and documentations.

The following two important issues have been raised during the community review:

- uncertainty propagation and,
- how to ensure the data, software, and information are being used correctly and ethically, i.e., the risk of abuse.

Both issues represent big challenges that face most if not all disciplines and are beyond the scope of this document. However, making quality information such as uncertainty estimates available in a consistent way during various stages or aspects of the dataset lifecycle is a first step towards addressing the uncertainty propagation issue. Likewise, improving data quality information can inform the use of data and reduce the potential for misuse.

Some aspects of the following key points, also raised by reviewers, may have been touched on at a high level in this document. However, providing the thorough exploration that these topics deserve is beyond the scope of the current effort. Nevertheless, these topics could serve as excellent candidates for future work:

- the role of technology design in influencing data quality, with specific discussion around different approaches to capturing knowledge;
- an in-depth discussion of intersections between classic geospatial data and citizen science, how they can augment each other, and how information, if provided via FAIR principles, can empower scientists and communities alike;
- a minimum set of reporting requirements for the decision maker.

This international FAIR dataset quality information community guidelines document is a living document and is expected to evolve over time to accommodate user feedback and emerging community best practices. Use cases will be developed to further improve the maturity of the guidelines and to provide implementation examples and lessons learned for the community.

ACKNOWLEDGMENT

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 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.

Specifically, G. Peng, C. Lacagnina, and I. Ivánová, as working group co-leads, contributed significantly to strategic planning and oversaw the overall guidelines development and review process.

The following working group members contributed to conceptualization and writing of the guidelines document: G. Peng, C. Lacagnina, R. R. Downs, I. Ivánová, H. Ramapriyan, A. Ganske, D. Jones, and L. Wyborn.

The following working group members contributed content to and reviewed the guidelines document: S. Champion, L. Bastin, C.-L. Shie, I. Bastrakova, G. Berg-Cross, and K. Bugbee.

The following working group members contributed to strategic planning: N. Ritchey, M. Wu, D. Moroni, and Y. Wei. They have also actively reviewed the guidelines document with beneficial edits.

All other working group members have reviewed the guidelines document.

G. Peng, R. R. Downs, H. Ramapriyan, Y. Wei and D. Moroni contributed to planning for and participated in community engagement events that were organized by the ESIP IQC. L. Wyborn, I. Ivánová, M. Wu and I. Bastrakova did so for the events organized by the AU/NZ DQIG. C. Lacagnina and G. Larnicol did so for the events organized by the BSC EQC team. We thank ESIP IQC, AU/NZ DQIG, and BSC EQC team members for their beneficial feedback. In addition, G. Peng, C. Lacagnina, I. Ivánová, R. R. Downs, H. Ramapriyan, and L. Wyborn organized or participated in community engagement events on behalf of the working group.

The community interest and support has helped propel us to the finish line. Contents from Peng et al. (2020; 2021) are reused in this document. We thank Janet Fennema for the excellent job in putting together the acronym list.

We thank ESIP for sponsoring the virtual workshop held on July 13, 2020 and all participants of the workshop and the ESIP 2020 Summer Meeting session for helping lay the ground for initiating and developing the guidelines. In particular, we thank the following presenters and attendees who have contributed significantly to the discussions: Mitch Goldberg, Jörg Schulz, Mirko Albani, Christina Lief, Shelley Stall, Lihang Zhou, Iolanda Maggio, Marie Drévillon, Brian Westra, Siri Jodha Khalsa, Kerstin Lehnert, Paul Lemieux, Donald Collins, Gastil Gastil-Buhl, and Danie Kinkade.

We thank participants in the ESIP 2021 Winter Meeting session, especially Chantel Ridsdale, Shawn Smith, Shannon Leslie, Rebecca Hudak, Tyler Christensen, and Venice Bayrd, for beneficial discussions and information, particularly on the top 3 stages that are associated with each quality aspect within the dataset lifecycle shown in Figure 1.

We thank all the reviewers who have taken their time and efforts in reviewing the draft document and provided us with valuable comments and suggestions which have improved the clarity and comprehensiveness of the document. Special thank 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.

REFERENCES

Austin, C, Cousijn, H, Diepenbroek, M, Petters, J, and Soares E Silva, M 2019 WDS/RDA Assessment of Data Fitness for Use WG Outputs and Recommendations. DOI:https://doi.org/10.15497/rda00034

Australia FAIR Access Working Group 2017 Policy Statement on FAIR Access to Australia's Research Outputs. Version: Jan 2017. Available at: https://www.fair-access.net.au/fair-statement

Baker, K S, Duerr, R E, and Parsons, M A 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

Barsi, Á, Kugler, Z, Juhász, A, Szabó, G, Batini, C, Abdulmuttalib, H, Huang, G, and Shen, H 2019 Remote sensing data quality model: from data sources to lifecycle phases. *International Journal of Image and Data Fusion*, 10(4):280-99. DOI:https://doi.org/10.1080/19479832.2019.1625977

Bates, J J and Privette, J L 2012 A maturity model for assessing the completeness of climate data records. *EOS, Trans*. American Geophysical Union, 93(44), 441. DOI:https://doi.org/10.1029/2012EO440006

Borda, A, Gray, K, and Fu, Y 2020 Research data management in health and biomedical citizen science: practices and prospects. *JAMIA Open*, 3(1):113-25. DOI:https://doi.org/10.1093/jamiaopen/ooz052

Bruce, T R and Hillmann, D I 2004 The Continuum of Metadata Quality: Defining, Expressing, Exploiting. In: Metadata in Practice. Cornell University Library: ALA Editions. Available at https://hdl.handle.net/1813/7895

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), p.17. DOI:http://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

Canali, S 2020 Towards a Contextual Approach to Data Quality. Data. 5(4):90. DOI:https://doi.org/10.3390/data5040090

CODATA 2019 The Beijing Declaration on Research Data. Version: 7 November 2019. Available at: http://www.codata.org/uploads/Beijing%20Declaration-19-11-07-FINAL.pdf

Coetzee, S 2018 Implementing Geospatial Data Quality Standards - Motivators and Barriers, 2nd International Workshop on Spatial Data Quality, Valletta, Malta 6-7 February 2018, https://eurogeographics.org/wp-content/uploads/2018/06/4-SDQ2018_Coetzee_V1e.pdf

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

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

Cunningham, J A, Speybroeck, M V, Kalra, D, and Verbeeck, R 2016 Nine Principles of Semantic Harmonization. *AMIA Annu Symp Proc. 2016,* 451–459. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5333211/

Deming, WE 1986 Out of the Crisis. MIT Center for Advanced Engineering Study. The 2010 MIT Press edition, 507 pp. Cambridge, MA, USA.

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

Dretske, F I 1981 Knowledge and the Flow of Information, 273 pp. Basil Blackwell, Oxford. Available at: https://www.amazon.com/Knowledge-Flow-Information-Fred-Dretske/dp/157586195X

Downs, RR, Ramapriyan, HK, Peng, G, and Wei, Y. 2021. Perspectives on Citizen Science Data Quality. Frontiers in Climate. 3. DOI: https://doi.org/10.3389/fclim.2021.615032

EUMETSAT 2013 CORE-CLIMAX Climate Data Record Assessment Instruction Manual. Version 2, 25 November 2013. Available from: https://www.eumetsat.int/search?text=CORE-CLIMAX

European Commission 2018 Turning FAIR into reality - Final Report and Action Plan from the European Commission Expert Group on FAIR data, EC: Brussels, DOI:https://doi.org/10.2777/1524

European Commission 2020 Recommendations on FAIR Metrics for EOSC, European Commission: Brussels, DOI: https://doi.org/10.2777/70791

European Commission and PwC EU Services 2018 Cost-benefit analysis for FAIR research data: Cost of not having FAIR research data. Version: March 2018. Available at: https://op.europa.eu/en/publication-detail/-/publication/d375368c-1a0a-11e9-8d04-01aa75ed71a1/language-en

Evans, B, Druken, K, Wang, J, Yang, R, Richards, C, and Wyborn, L 2017 A data quality strategy to enable fair, programmatic access across large, diverse data collections for high performance data analysis. *Informatics*, 4(4), 45. DOI:https://doi.org/10.3390/informatics4040045

FGDC (Federal Geographic Data Committee) 2002 Content standard for digital geospatial metadata – extension for remote sensing data. Version: FGDC-STD-012-2002. *Federal Geographic Data Committee.* Washington, D.C. Available at: https://www.fgdc.gov/standards/projects/csdgm_rs_ex/MetadataRemoteSensingExtens.pdf

Figgemeier, H, Henzen, C, 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

Ganske, A, Kraft, A, Kaiser, A, Heydebreck, D, Lammert, A, Höck, H, Thiemann, H, Voss, Grawe, V D, Leitl, B, Schlünzen, K H, Kretzschmar, J, and Quaas, J 2020a ATMODAT Standard (v3.0). World Data Center for Climate (WDCC) at DKRZ. Available at: https://cera-www.dkrz.de/WDCC/ui/cerasearch/entry?acronym=atmodat_standard_en_v3_0

Ganske, A, Heydebreck, D, Höck, H, and Kaiser, A 2020b A short guide to increase FAIRness of atmospheric model data. *Meteorol. Z. (Contrib. Atm. Sci.),* 29, DOI:https://doi.org/10.1127/metz/2020/1042 (supplement includes a JSON example for DOI metadata.)

GCMD 2020 GCMD Keywords, Version 9.1. Greenbelt, MD: Earth Science Data and Information System, Earth Science Projects Division, Goddard Space Flight Center (GSFC)

National Aeronautics and Space Administration (NASA). Available at: https://earthdata.nasa.gov/earth-observation-data/find-data/gcmd/gcmd-keywords
G20 Leaders 2016 G20 Leaders' Communique Hangzhou Summit. Version: 5 September 2016. Available at: https://ec.europa.eu/commission/presscorner/detail/en/STATEMENT_16_2967
Haiden, T, Janousek, M, Vitart, F, Ferranti, L, Prates, F, and Prates, F 2019. Evaluation of ECMWF forecasts, including the 2019 upgrade. DOI:https://doi.org/10.21957/mlvapkk
Henzen, C, Della Chiesa, S, 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
Hewitt, C D, Stone, R C, and Tait, A B 2017 Improving the use of climate information in decision-making. Nature Climate Change, 7(9), 614–616. DOI:https://doi.org/10.1038/nclimate3378
Hey, T, Tansley, S, and Tolle, K 2009 The Fourth Paradigm: Data-Intensive Scientific Discovery. *Microsoft Research*, USA. 252 pp.
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 and Toussaint, F 2019 Quality Maturity Matrix Checklist for Levels 4 and 5 with Protocols. World Data Center for Climate (WDCC) at DKRZ. DOI:https://doi.org/10.2312/WDCC/TR_QMM_Checkl_Levels_4-5_Prots
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, DOI:http://doi.org/10.5334/dsj-2020-045
Illari, P 2014 IQ: Purpose and Dimensions. In The Philosophy of Information Quality; Floridi, L, Illari, P, Eds.; Springer: Berlin, Germany pp. 281–302. DOI:https://doi.org/10.1007/978-3-319-07121-3_14
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 19131 2007 Geographic information — Data product specifications. Version: 2007-04. International Organization for Standardization. Geneva, Switzerland. Available at: https://www.iso.org/standard/36760.html
ISO 19157 2013 Geographic information - Data quality, Geneva, Switzerland, Available at: https://www.iso.org/standard/32575.html
ISO 19165-2:2020 Geographic information — Preservation of digital data and metadata — Part 2: Content specifications for Earth observation data and derived digital products, Geneva, Switzerland, Available at: https://www.iso.org/standard/73810.html
Jones, M and Slaughter, P 2019 Quantifying FAIR: metadata improvement and guidance in the DataONE repository network. *DataONE Webinar*, May 14, 2019. Slides ae available at: https://www.dataone.org/uploads/dataonewebinar_jonesslaughter_fairmetadata_190514.pdf
Kahn, M G, Brown, J S, Chun, A T, Davidson, B N, Meeker, D, Ryan, P B, Schilling, L M, Weiskopf, N G, Williams, A E, and Zozus, M N 2015 Transparent reporting of data quality in distributed data networks. *Egems*, 3(1). DOI:https://doi.org/10.13063/2327-9214.1052
Lawrence, B., Jones, C., Matthews, B., Pepler, S., and Callaghan, S. 2011 Citation and Peer Review of Data: Moving Towards Formal Data Publication, *Int. J. Digi. Cur.*, 2, 4–37.

Lee, Y W, Strong, D M, Khan, B K and Wang, R Y 2002: AIMQ: a methodology for information quality assessment, *Information & Management*, 40, 133-146. DOI:https://doi.org/10.1016/S0378-7206(02)00043-5

Lenhardt W, Ahalt S, Blanton B, Christopherson L, and Idaszak R 2014 Data management lifecycle and software lifecycle management in the context of conducting science. *Journal of Open Research Software*, 2(1). DOI:http://doi.org/10.5334/jors.ax

Leonelli, S 2017 Global Data Quality Assessment and the Situated Nature of "Best" Research Practices in Biology. *Data Science Journal,* 16, p.32. DOI:http://doi.org/10.5334/dsj-2017-032

Lemieux III, P, 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 and Peng, G 2019 The WMO Stewardship Maturity Matrix for Climate Data (SMM-CD) Template. Document ID: WMO-SMM-CD-0003. Updated 2020. Version: v04r01 20200615. *Figshare*. DOI:https://doi.org/10.6084/m9.figshare.7003709

Lin, D, Crabtree, J, Dillo, I, Downs, R R, 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, D V, Stockhause, M, Westbrook, J 2020. The TRUST Principles for Digital Repositories. *Scientific Data* 7, 144. DOI:https://doi.org/10.1038/s41597-020-0486-7

Mankins, J 1995 Technical Readiness Level - A White Paper. Version: April 6, 1995. Available at: http://www.artemisinnovation.com/images/TRL_White_Paper_2004-Edited.pdf

Mankins, J 2009 Technology readiness assessments: A retrospective. *Acta Astronautica*. 65. Available at: http://www.onethesis.com/wp-content/uploads/2016/11/1-s2.0-S0094576509002008-main.pdf

Matthews, J L, Mannshardt, E, and Gremaud, P 2013 Uncertainty Quantification for Climate Observations, *Bulletin of the American Meteorological Society,* **94**, ES21-ES25. DOI:http://doi.org/10.1175/BAMS-D-12-00042.1

Mislan, K A S, Heer, J M, and White, E P, 2015 Elevating The Status of Code in Ecology. *Trends Ecol. Evol*., 31(1). DOI:https://doi.org/10.1016/j.tree.2015.11.006

Moe, K, Jones, D, Bermudez, L E, and Fayne, J V 2018 Operational Readiness Levels - A Trust Metric for Operational Data. *American Geophysical Union Fall Meeting 2018,* abstract #IN52B-08.

Mons, B, 2018 Data Stewardship for open science: implementing FAIR principles. 1st Edition. Chapman and Hall/CRC Press, Taylor & Francis, New York. 244 pp. Available at: https://www.taylorfrancis.com/books/9781315380711

Moroni, D F, Ramapriyan, H, Peng, G, Hobbs, J, Goldstein, J C, Downs, R R, Wolfe, R, Shie, C-L, Merchant, C J, Bourassa, M, Matthews, J L, Cornillon, P, Bastin, L, Kehoe, K, Smith, B, Privette, J L, Subramanian, A C, 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

Mosely, M, Brackett, M, Early, S and Henderson, D. (eds) 2009 The Data Management Body of Knowledge (DAMA-DMBOK Guide). Bradley Beach, NJ, USA: Technics Publications, LLC. 2nd Print Edition, 406 pp.

NCEI/ESIP-DSC MM-Serv Working Group 2018 NCEI/ESIP-DSC Data Use and Services Maturity Matrix (MM-Serv). *Figshare*. DOI:https://doi.org/10.6084/m9.figshare.6855020

Nightingale, J, Mittaz, J P D, Douglas, S, Dee, D, Ryder, J, Taylor, M, Old, C, Dieval, C, Fouron, C, Duveau, G, and Merchant, C 2019 Ten priority science gaps in assessing climate data record quality. *Remote Sensing*, 11(8), 986. DOI:https://doi.org/10.3390/rs11080986

O'Brien, M, Costa, D, Servilla, M 2016 Ensuring the quality of data packages in the LTER network data management system. *Ecological Informatics,* 36. DOI:https://doi.org/10.1016/j.ecoinf.2016.08.001

OGC (Open Geospatial Consortium) 2020 OGC Discussion Paper 2020 On the definition of dataset. Available at: https://github.com/heidivanparys/discussion_paper_dataset/releases/download/v20200312/DiscussionPaperDataset.pdf

Pearlman, J, Bushnell, M, Coppola, L, Karstensen, J, and others 2019 Evolving and Sustaining Ocean Best Practices and Standards for the Next Decade. *Front. Mar. Sci.*, DOI:https://doi.org/10.3389/fmars.2019.00277

Peng, G 2014 NCDC-CICSNC Scientific Data Stewardship Maturity Assessment Model Template. Figshare. Updated: 2015. Version: v4.0-20150623. DOI:https://doi.org/10.6084/m9.figshare.1211954.v23

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

Peng, G, Privette, J L, Kearns, E J, Ritchey, N A, 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, Privette, J L, Tilmes, C, Bristol, S, Maycock, T, Bates, J J, Hausman, S, Brown, O, and Kearns, E J 2018 A Conceptual Enterprise Framework for Managing Scientific Data Stewardship. *Data Science Journal*, 17. DOI:https://doi.org/10.5334/dsj-2018-015

Peng, G, Milan, A, Ritchey, N, Partee II, R P, Zinn, S, McQuinn, Lemieux III, P E, Ionin, R, Collins, D, Jones, P, Jakositz, A, and Casey, K S 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, 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

Peng, G, Lacagnina, C, Downs, R R, Ivánová, I, Moroni, D F, Ramapriyan, H, Wei, Y, and Larnicol, G 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. *Open Science Framework,* https://doi.org/10.31219/osf.io/75b92

Peng, G, Downs, R R, Lacagnina, C, Ramapriyan, H, Ivánová, I, and others 2021 Call to Action for Global Access to and Harmonization of Quality Information of Individual Earth Science Datasets. *Data Science Journal*, 20. DOI: http://doi.org/10.5334/dsj-2021-019

Popp, T, Hegglin, M I, 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

Press, G 2016 Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says. Forbes. Version: March 23, 2016. Available at: https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/?sh=1ee368c06f63

Ramapriyan, H and Moses, J 2012 NASA Earth Science Data Preservation Content Specification. NASA GSFC. Document ID: 423- SPEC- 001. Available at: https://cdn.earthdata.nasa.gov/conduit/upload/10607/NASA_ESD_Preservation_Spec.pdf

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

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

Renear, A H, Sacchi, S, and Wickett, K M 2010 Definitions of Dataset in the Scientific and Technical Literature. ASIST 2010. *The American Society for Information Science and Technology*, Pittsburgh, PA. DOI:https://doi.org/10.1002/meet.14504701240

Redman, C T 1996 Data quality of the information age. Artech House, Boston. 303 pp.

Romain, D, Mabile, L, Mohamed, Y, Cambon-Thomsen, A, Archambeau, A-S, Bezuidenhout, L, and others 2018 How to operationalize and assess the inclusion of the 'FAIR' concept in data sharing: towards a simplified assessment grid for compliance with the FAIR criteria. (Version 1.0). *The National Open Science Day (JNSO 2018)*, Paris France: *Zenodo*. DOI:http://doi.org/10.5281/zenodo.1995646

Stockhause, M, Höck, H, Toussaint, F, and Lautenschlager, M 2012 Quality assessment concept of the World Data Center for Climate and its application to CMIP5 data. *Geoscientific Model Development*, 5(4), 1023–1032. DOI:https://doi.org/https://doi.org/10.5194/gmd-5-1023-2012

Tegmark, M 2013 Everything in the Universe Is Made of Math – Including You. discovermagazine.com. Version: 4 November 2013. Accessed 2/25/2017 at: http://discovermagazine.com/2013/dec/13-math-made-flesh

Tilmes, C, Privette, A P, Chen, J, Ramachandran, R, Bugbee, K M, and Wolfe, R E 2015a Linking from observations to data to actionable science in the climate data initiative. Proc. 2015 IEEE Geosci. and Remote Sensing Symposium, 26 - 31 July 2015, Milan, Italy.

Tilmes, C, Wolfe, R E, Duggan, B, Aulenbach, S, Goldstein, J C, Ma, X, and Zednik, S 2015b Supporting trust with provenance of the findings of the national climate assessment. METHOD 2015: The 4th Intl. Workshop on Methods for Establishing Trust of (Open) Data. 11 Oct. 2015, Bethlehem, PA, USA. Available at: http://www.few.vu.nl/~dceolin/method2015/papers/METHOD_2015_paper_2.pdf

U.S. Public Law 115-435 2019 Foundations for Evidence-Based Policymaking Act of 2018. Title II OPEN Government Data Act. Version: 14 January 2019. Available at: https://www.congress.gov/115/plaws/publ435/PLAW-115publ435.pdf

W3C 2016 Data on the Web Best Practices: Data Quality Vocabulary. Version: 15 December 2016. Available at: https://www.w3.org/TR/vocab-dqv/

Wagner, M, Henzen, C, 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, R Y and Strong, D M 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

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

WMO (World Meteorological Organization) 1992 International meteorological vocabulary. Document ID: WMO-No. 182. Available at: https://library.wmo.int/?lvl=notice_display&id=220#.YEjM5ndKiqc

WMO 2019 Manual on the High-quality Global Data Management Framework for Climate. Document ID: WMO-No. 1238. *World Meteorological Organization*. Available at: https://library.wmo.int/index.php?lvl=notice_display&id=21686

Woo, L M 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, & 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

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

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

APPENDICES

Appendix A. Terms and Definitions

There is a fair amount of work on terms and definitions related to data and information management. Many organizations, if not all, including domain and national data centers or repositories, maintain a glossary or a list of vocabulary, i.e., an alphabetical list of terms and brief explanations.[9] It often relates to a particular subject or domain. For example, the World Meteorological Organization (WMO) maintains an international meteorological vocabulary (WMO 1992), which has been continuously updated and is searchable online.[10] Some glossaries have been developed for the research data management community, such as the one maintained by the Committee on Data of the International Science Council (CODATA) [11], the Data Foundation and Terminology (DFT) Interest Group of Research Data Alliance (RDA) [12], and Consortia Advancing Standards in Research (CASRAI).[13] One can also obtain definitions of terms from online dictionaries such as Merriam-Webster [14] or wikipedia [15] as well as from community and international standards such as Mosely et al. (2009) and ISO/TC 211.[16] [17]

It is beyond the scope of this document to provide a comprehensive list of vocabulary pertaining to data and information quality management. It is, however, beneficial to our stakeholders to provide a list of key terms used in this document and their definitions and/or explanations. It is also helpful to better limit the scope of those terms as they may be used in various contexts in different disciplines.

Currently, there are no formal or unique definitions for many of those key terms. Definitions may be different within specific domains and disciplines. For example, Renear et al. (2010) examined the literature and found large variation among the definitions of dataset. Additional discussion on the definition of dataset can be found in OGC (2020).

In some instances, there appears to be a visible gap between a scholarly definition that tends to be highly abstract and that described with layman's terms which is more practical to follow. For example, definitions of dataset collected by the RDA DFT Interest Group include "a type of managed data aggregation from multiple data elements which are considered as an aggregated unit for processing purposes" as well as "A collection of data, published or curated by a single agent, and available for access or download in one or more formats." Again, it is beyond the scope of this document to discuss the variety of existing definitions of those key terms. We, therefore, have aimed to collect and integrate existing definitions from various sources to retain the generality of these definitions, but to adopt the ones that are easy to understand and also to align them with that of the Earth science community (Table A1). The goal is to provide a good starting point for

[9] https://science.nasa.gov/glossary

[10] https://public.wmo.int/en/resources/meteoterm

[11] https://codata.org/rdm-glossary/

[12] https://smw-rda.esc.rzg.mpg.de/dft-2.0.html

[13] https://casrai.org/rdm-glossary/

[14] https://www.merriam-webster.com/dictionary/data

[15] https://en.wikipedia.org/wiki/Data

[16] https://isotc211.geolexica.org/

[17] https://www.iso.org/obp/ui

reaching a community consensus. It is expected that definitions may evolve with feedback from the Earth science community.

Table A1: Definitions/description, Examples/Notes of Key Terms

Term Examples/Notes	**Definition/Description**
Data	*Data* are representations of observations, objects, or other phenomena and can refer to anything that is collected, observed, generated or derived, and used as a basis for reasoning, discussion, or calculation. Data can be either structured or unstructured, analog or digital, and can be represented in quantitative, qualitative, or physical forms.
Examples/Notes	Quantitative data can be either discrete or continuous numbers, such as in situ/ground, suborbital or satellite measurements. Qualitative data is descriptive text, such as description of weather stations and sensors used in meteorological observations. Tegmark (2013) has argued that the data in qualitative form can also be represented digitally. Data in a physical form can be deduced or interpreted from physical samples such as air, water, fish, or ice core samples. Physical data can also be represented in analog by images, hand-drawn nautical charts or physical meteorological records. Unless it is digitally described, a collection of data in physical form is not easy to share and is not machine interoperable. However, nowadays, almost all the images, nautical charts, or physical meteorological records are digital or digitized. Generated data can be results from a numerical model, e.g., a climate model, or a statistical model, e.g., a linear regression model. Derived data can be a data product with a well-thought out algorithm or approach that facilitates an end goal through the use of observations. 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 (e.g., FGDC 2002).
Dataset	*Dataset* refers to an identifiable collection of data (e.g., ISO 19115-1 2014). A dataset can be processed, curated or published by a single agent (e.g., W3C 2020).
Examples/Notes	A dataset is a type of managed data aggregation from multiple data elements which are considered as an aggregated unit for processing purposes. The general notion of datasets found in the literature currently is characterized by an interrelated family of more specific concepts: grouping, content, relatedness, and purpose (Renear et al 2010). Time series are good examples of a dataset. The 1790-1960 Decennial Censuses are described as Datasets by such repositories as the CISER Data Archive: Online Catalog. An overview of existing definitions of a dataset can be found in OGC (2020).
Data Collection	*Data Collection* refers to a grouping of digital data or products that share common characteristics, is represented by a single metadata record, and consists of one or more granules. A collection is often identified by a PID.

Examples/Notes	In this document, a data collection refers to a minimum citable unit of data, which oftentimes refers to a dataset. What consists of data collection can be very subjective. Contents in a data collection may be static or change over time. For example, a query issued to a database can be invariant, but the result may change each time. A sensor data stream always has new data from the most recent observation. The stream itself may be identified, but the contents are not static.
Data Granule	*Data Granule* refers to the smallest aggregation of data that can be independently managed (described, inventoried, and retrieved) in an archival and/or dissemination system.
Examples/Notes	Common examples of a granule can be an individual sample in a sample collection or an individual data file of a dataset. Like the concept of data collection, data granule can be quite subjective.
Data Quality	*Data quality* is a degree to which a set of inherent characteristics of data fulfills requirements (e.g., ISO 8000-2 2018)
Examples/Notes	Data quality can be expressed as a conformance with product specification (e.g., 95% conformant), or as detailed report including specifics of quality evaluation per each data characteristic (e.g., for 'positional accuracy': 25% of the nodes within data quality scope have an error distance greater than 1m).
Dataset Quality	*Dataset quality* includes quality of both data and associated information, examples of which are metadata, software, algorithms, and practices or procedures applied to the dataset throughout its entire life cycle. *Dataset quality* is a multi-dimensional construct perception and/or a judgment of data's fitness or trustworthiness to serve intended research uses in a given context.
Examples/Notes	The scope of quality of a dataset goes beyond that of data. See Figure C1 and Table C1 for lists of dataset quality aspects and attributes such as Completeness (no gaps in coverage) and standards (data and metadata). Not included is consistency (internal and external).
Dataset Quality Information	*Dataset quality information* includes both data quality descriptive information such as that captured in documents, e.g., papers, reports or user guides, and quality metadata that is captured in a metadata record, throughout the entire life cycle of a dataset.
Examples/Notes	
Information	*Information* is considered as data being processed, organized, structured, communicated or presented so as to be meaningful to the recipient in a given context.
Examples/Notes	"Information is data in context." Mosely et al. (2009). Examples include spatial distribution maps of topographic or bathymetric measurements.
Knowledge	*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.
Examples/Notes	*Knowledge* is gained from an understanding of the significance of information (Mosely et al. 2009). Knowledge represents the internalized or understood information that can be used to make decisions.[18] The relationship between information and knowledge includes prior knowledge about a specific information source and additional knowledge is added by the interpreted content of information which may be conveyed by data representing that information. One then says that data bears information (Dretske 1981).
Metadata	*Metadata* is literally data about data and provides information about the data. Metadata plays a role of documenting data and may be categorized into many types including: descriptive, structural, administrative, reference, and statistical metadata. Metadata entities may also be categorized as search and discovery, usability, access, usage, provenance, and quality metadata. Furthermore, metadata may be categorized as business and technical metadata.[19]
Examples/Notes	Examples include data about related datasets (including provenance metadata), software, publications, organisations, persons (such as producer of the data). Typically, descriptive metadata includes such things as source & time of creation. For data and report publication it may include administrative metadata such as authors & date of submission. A PID is an example of metadata used to reference data. An example is retention period metadata which defines the date when retention of the data object should be evaluated. A metadata record may be curated at dataset level (i.e., collection-level) or at file-level (i.e., granule-level). Metadata contained in a NetCDF data file is considered as a granule-level metadata.
Provenance	*Provenance* is a type of historical information or metadata about the origin and history of entities, locations, activities, and people involved in producing, modifying, and preserving a piece of data or object.
Examples/Notes	Examples include creation, attribution, or version history of managed data.
Quality Assessment	*Quality Assessment* is a totality of measures carried out consistently and systematically in order to assure that a product conforms with the requirements of a stated specification (ISO/IEC 2382-36 2019)
Examples/Notes	Assessment may be in the form of checks for evaluating accuracy or completeness of data values, variable names and units, content of metadata elements. Assessment is often guided by policy. These may include periodic data files integrity validation.
Quality Assurance	*Quality Assurance* is part of quality management focused on providing confidence that quality requirements will be fulfilled (ISO 9000: 2005)

[18] https://codata.org/rdm-glossary/knowledge/

[19] https://codata.org/rdm-glossary/metadata/

Examples/Notes	Data quality assurance example: A quality assurance framework for Earth observations (QA4EO) has been developed with a set of guidelines to provide guidance on EO quality assurance (Yang 2013). Additional documentation can be found at: http://qa4eo.org/documentation/. Metadata quality assurance example: A metadata quality assessment framework has been developed by the NASA the Analysis and Review of the Common Metadata Repository (ARC) team, focusing on the metadata quality dimensions of correctness, completeness, and consistency (Bugbee et al. 2021).
Data[set] Quality Attribute	*Data[set] Quality Attribute* is a characteristic describing a certain aspect of data[set].
Examples/Notes	Examples of quality attributes include (positional, thematic or temporal) accuracy, completeness, consistency, resolution, quality of service, homogeneity, provenance.
Quality Control	*Quality Control* is part of quality management focused on fulfilling quality requirements (ISO 9000:2005).
Examples/Notes	
Data[set] Quality Dimensions	*Quality Dimensions* represent the degree to which to which a set of inherent characteristics of data[set] fulfill requirements (ISO 8000-2:2020)
Examples/Notes	

Appendix B. FAIR Principles and Earth Science Implementation Examples

Table B1: Definitions of the FAIR Data Principles from Wilkinson et al. (2016), explanation based on the Swiss National Science Foundation, along with implementation examples of practices from the Earth sciences community.

	Definitions (From Wilkinson et al. 2016)	**Explanation** (Based on Swiss National Science Foundation)[1]	**Implementation Examples for Earth Sciences Datasets**
F	F1. (meta)data are assigned a globally unique and eternally persistent identifier.	Each dataset is assigned a globally unique and persistent identifier (PID), for example a DOI, ARK, RRID… These identifiers allow users to find, cite and track (meta)data.	Dataset is assigned a DOI (a globally unique and persistent digital object identifier), that is minted by DataCite (well-established minting service provider) with a defined metadata schema, resolved to a standardized layout dataset landing web page (e.g., NOAA OISST; NOAA/NSIDC Sea Ice Concentration), driven by a comprehensive dataset-level metadata record that conforms to ISO 19115-1 or WMO Integrated Global Observing System (WIGOS) metadata standard (domain, national, or international standards) and is integrated into geoportal or indexed in Google Data Search or catalog.data.gov (US); europeandataportal.eu (EU); ecat.ga.gov.au (GA); https://researchdata.edu.au/(AU) (a data registry and discovery platform with standard-based protocols of metadata and APIs.) The data can be accessed by direct download or via data servers such as Thematic Real-time Environmental Distributed Data Services (THREDDS), and Environmental Research Division's Data Access Program (ERDDAP) server (e.g., NCEI OISST; open, free data access protocol that allows for
	F2. data are described with rich metadata (defined by R1 below).	Each dataset is thoroughly (see below, in R1) described: the metadata document how the data was generated, under what terms (license) and how it can be (re)used, and provides the necessary context for proper interpretation. This information needs to be machine-readable.	
	F3. metadata clearly and explicitly include the identifier of the data it describes.	The metadata and the dataset they describe can be stored in separate files. The association between a metadata file and the dataset is obvious thanks to the mention of the dataset's PID in the metadata.	
	F4. (meta)data are registered or indexed in a searchable resource.	Metadata are used to build easily searchable indexes of datasets. These resources will allow us to search for existing datasets similar to searching for a book in a library.	
A	A1. (meta)data are retrievable by their identifier using a standardised communications protocol.	If one knows a dataset's identifier and the location of where it is archived, one can at least access the metadata. Furthermore, the user knows how to proceed to get access to the data.[2]	
	A1.1. the protocol is open, free, and universally implementable.	Anyone with a computer and an internet connection can access at least the metadata.	

	A1.2. the protocol allows for an authentication and authorization procedure, where necessary.	It often makes sense to request users to create a user account on a repository.[3] This allows authentication of the owner (or contributor) of each dataset, and to potentially set user specific rights.	authentication and authorization if needed.) ISO 19115-1 collection-level (dataset) metadata record explicitly includes dataset PID, the data access protocol, the data usage and access license, and provenance elements. It also includes vocabulary terms and their URI(s).
	A2. metadata are accessible, even when the data are no longer available.	Maintaining all datasets in a readily usable state eternally would require an enormous amount of curation work (adapting to new standards for formats, converting to different formats if specifically needed software is discontinued, etc.). Keeping the metadata describing each dataset accessible, however, can be done with much less resources. This allows building of comprehensive data indexes including all current, past and potentially arising datasets.	
I	I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.	Interoperability typically means that each computer system has at least knowledge of the other system's formats in which data is exchanged. If (meta)data are to be searchable and if compatible data sources should be combinable in a (semi)automatic way, computer systems need to be able to decide if the content of datasets are comparable. Ontologies help with constraining the meaning of terms (Cunningham et al. 2016). Obvious issues arise when different languages are used to describe the data or when spelling errors make the comparison of descriptions and variable names more difficult. It is critical to use controlled vocabularies and a well-defined framework to describe and structure (meta)data in order to ensure findability and interoperability of datasets.	Data in NetCDF (Network Common Data Format - interoperable; community standard) with global attributes for provenance and relevant data sources; compliant with NetCDF Climate and Forecast (CF) Metadata Conventions [4] (e.g., standard variable names, units, …), utilizing Global Change Master Directory (GCMD) keywords [5] or other domain specific vocabularies. (All GCMD keywords are publicly available and GCMD assigns a UUID (universally unique identifier) for each of the keywords.) NetCDF data files are self-describing and carry the data schema as they go so they are self-defining (Hey et al. 2009). So are HDF (Hierarchical Data Format) files and their variations.
	I2. (meta)data use vocabularies that follow FAIR principles.	The controlled vocabulary used to describe datasets needs to be documented. This documentation needs to be easily findable and accessible by anyone who uses the dataset.	

	I3. (meta)data include qualified references to other (meta)data.	If the dataset builds on another dataset, if additional datasets are needed to complete the data, or if complementary information is stored in a different dataset, this needs to be specified. In particular, the scientific link between the datasets needs to be described. Furthermore, all datasets need to be properly cited (i.e., including their persistent identifiers). For provenance, it is essential that each derivative dataset links to the version of the source dataset. Through identifiers it should be possible to link back to the original source version of the data and enhance reproducibility of the higher level data products (Klump et al. 2021).	Standards for capturing the data provenance: W3C PROV, DCAT2 or ISO 19115-1 Metadata is also given with a knowledge representation data model such as RDFS or OWL. Links to other related datasets are given, e.g., in the case of simulation results of a regional atmospheric model, the dataset with the boundary conditions of the simulation is mentioned in the metadata. Related publications to the dataset are cited with their respective PIDs, such as the publication for which the data was used, about data sources, error estimates, validation results, and/or auxiliary data (e.g., published boundary/initial conditions and forcing datasets for model data.)
R	R1. meta(data) are richly described with a plurality of accurate and relevant attributes,	Description of a dataset is required at two different levels: (1) metadata describing the dataset (intrinsic): what does the dataset contain, how was the data generated, how has it been processed, how can it be reused … (2) metadata describing the data submitter-defined: any needed information to properly use the data, such as definitions of the variable names.	See examples in F & A - may repeat relevant examples in this aspect if deemed necessary. Metadata management is core to support discovery and reuse of data products. Wagner et al. 2021 proposed a tool for automated metadata and data quality extraction to foster the provision of FAIR data by generating (complementing) metadata for collected data, and by providing structured machine-readable quality information
	R1.1. (meta)data are released with a clear and accessible data usage licence.	The conditions under which the data can be used should be clear to machines and humans. This has to be specified in the metadata describing a dataset.	
	R1.2. (meta)data are associated with detailed provenance.	Detailed information about the provenance of data is necessary for reuse: this will, for example, allow researchers to understand how the data was generated, in which context it can be reused, and how reliable it is. Provenance is a central component in	

		scientific databases to validate data.	
	R1.3. (meta)data meet domain-relevant community standards.	It is easier to reuse datasets if they are similar: same type of data, data organized in a standardized way, well-established and sustainable file formats, documentation (metadata) following a common template and using common vocabulary. If community standards or best practices for data archiving and sharing exist, they should be followed. Note that quality issues are not addressed by the FAIR principles. How reliable data is lies in the eye of the beholder and depends on the foreseen application.	

[1] Swiss National Science Foundation: Explanation of the FAIR data principles. Available at: http://www.snf.ch/SiteCollectionDocuments/FAIR_principles_translation_SNSF_logo.pdf

[2] This is not given - Additional information needs to be documented to inform users how to proceed to get access to data. For novice users who do not know anything about data access, but want to use the data for decision making, they will languish as frustration builds.

[3] This requirement may actually hinder data sharing when anonymity needs to be protected - should not be used widely and only for data that requires access constraints which can be specified in metadata along with information on the type of authentication and/or authorization. For machine access, authentication can be achieved via an access key in many APIs.

[4] Additional information on the NetCDF Climate and Forecast (CF) Metadata Conventions: http://cfconventions.org/cf-conventions/cf-conventions.html

[5] Additional information on the Global Change Master Directory (GCMD) keywords: https://earthdata.nasa.gov/earth-observation-data/find-data/gcmd/gcmd-keywords

Appendix C. Dataset Quality Attributes, Aspects, and Dimensions

A good number of distinctive quality attributes or characteristics may be associated with a dataset. For example, from a data consumer perspective, over 179 individual data quality attributes were identified by a survey highlighted in Wang and Strong (1996). Many of these quality attributes may be overlapping to a certain extent, for example, accuracy, correctness, free from bias, etc.

Multiple quality attributes of a dataset may be grouped together to emphasize a certain aspect of data and information quality such as findability, accessibility, interoperability, and reusability as FAIR for data sharing (Wilkinson et al. 2016) or accuracy, precision, and uncertainty for scientific quality aspect as defined by Ramapriyan et al. (2017).

Data quality attributes can be categorized into different dimensions and aspects. For instance, Wang and Strong (1996) prioritized 179 quality attributes down to 15 and categorized them into four dimensions that are important to data consumers:

- Intrinsic (accuracy, objectivity, believability, reputation);
- Contextual (relevance, value-added, timeliness, completeness, appropriate amount of data);
- Representational (interpretability, ease of understanding, concise representation and representational consistency);
- Accessibility (accessibility, access security).

Redman (1996) defined accuracy, completeness, consistency, and currency as four quality dimensions of data values. Garvin (1987) proposed eight dimensions of product quality management that can be used to analyze product quality characteristics for a company to deliver reliable products: performance, features, reliability, conformance, durability, serviceability, aesthetics, and perceived quality. Bruce and Hillmann (2004) identified completeness, correctness, provenance, consistency, timeliness and accessibility as common metadata quality dimensions.

ISO/IEC 25010 (2011) refers to dimension and attribute as “characteristic” and “sub-characteristic” and defines eight quality characteristics in its software product quality model: functional suitability, performance efficiency, compatibility, usability, reliability, security, maintainability, and portability. ISO 19157 (2013) defines quality along following dimensions: accuracy (spatial, thematic and temporal), completeness, consistency and metaquality.

Ramapriyan et al. (2017) categorized quality attributes into four aspects based on the full dataset life cycle outlined by the ring of circles in Figure 1; also shown in Figure C1.

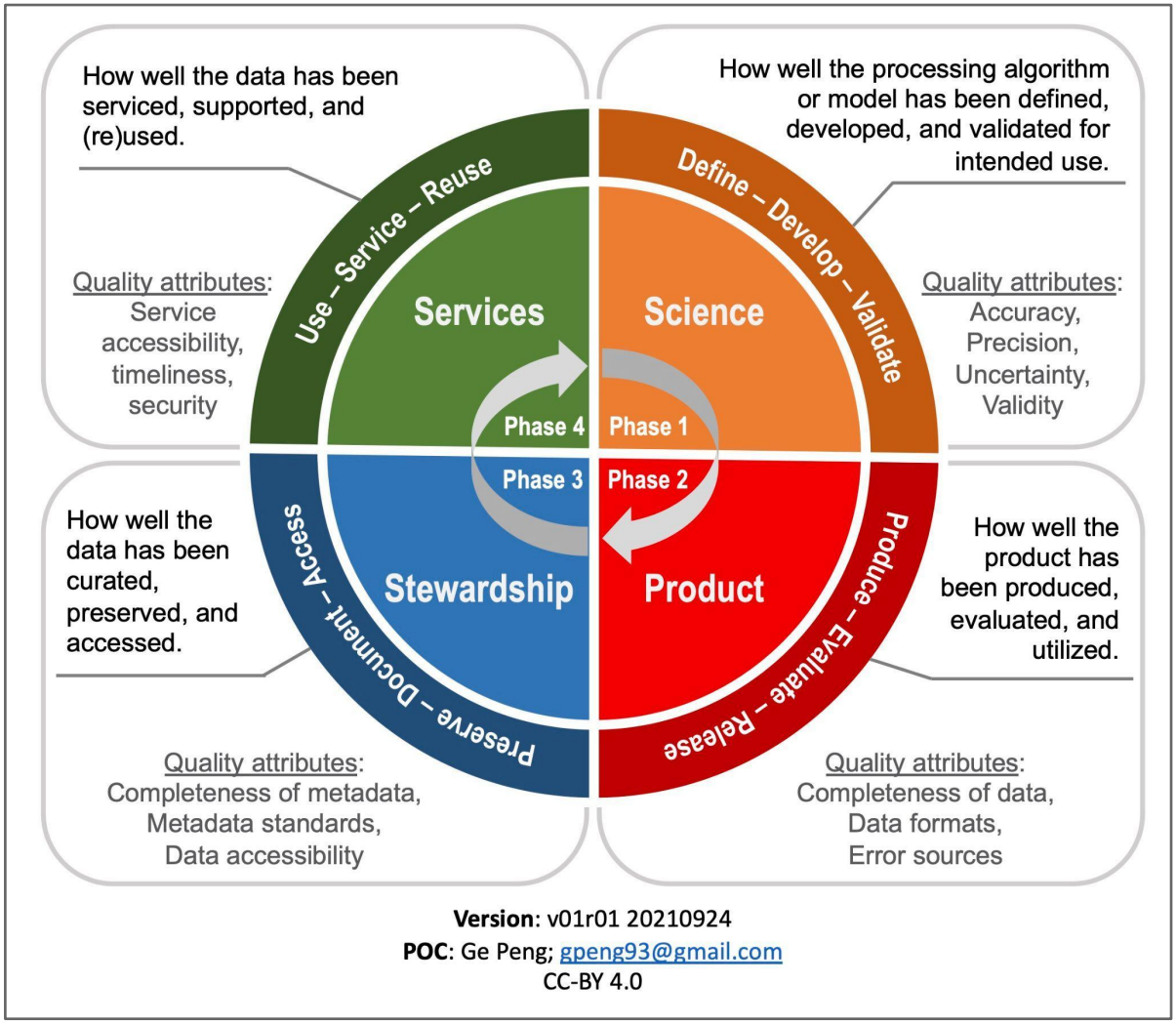

Figure C1: Description of quality aspects throughout a dataset lifecycle, three key stages and a few quality attributes associated with each quality aspect. The quality aspects and associated phases are based on Ramapriyan et al. (2017) with the following changes, based on feedback from the ESIP community and the International FAIR DQI Working Group: i) Replaced "Assess" by "Evaluate" in the Product aspect; ii) Replaced "Deliver" by "Release" in the Product aspect; and iii) Replaced "Maintain" by "Document" in the Stewardship aspect. Additionally, completeness of metadata is moved from the Product to Stewardship aspect. Creator: Ge Peng; Contributors to conceptual design: Lesley Wyborn and Robert Downs.

For each quality aspect, three key stages are identified. A description of these stages and associated document types is provided below in Table C1. The descriptions of the stages and document types are preliminary and tend to lean towards practices associated with managing satellite data, especially climate data, and may be modified according to feedback from other domains.

Table C1: Key stages of dataset lifecycle and associated quality aspect and document types

Key Stage	**Description**	**Quality Aspect /Document Type**
Define	Planning/designing/defining data accuracy and precision requirements for the intended use.	Science Quality Aspect/Science Report
Develop	Developing or implementing the software for the processing algorithm, quality assurance procedure for input data, quality control procedure for data generation, and data management plan.	
Validate	Assessing/validating the processing algorithm to ensure that the data product meets defined requirements. Generating the Science Report document including the data accuracy and precision requirements, intended purpose, and the timeline for data availability.	
Produce	Generating the data product according to the defined production working flow.	Product Quality Aspect/Product Report
Evaluate	Evaluating the quality of the data product against other similar data products; Initial analysis often carried out by scientist(s) who developed and produced the data product. The maturity of the software may also be evaluated. Generating Product Report document including description of actual production workflow, data processing flowchart, QA/QC procedures, data error sources and uncertainty.	
Release	Initial release to research community or delivery to a data center or repository; often limited in data access with minimum metadata and documentation. Users are required to have extensive knowledge about the data product and the subject the product was designed for.	
Preserve	Ingesting and archiving data; curating rich dataset-level metadata.	Stewardship Quality Aspect/Stewardship Report
Document	Maintaining and documenting: Ensuring data fixation; evaluating the stewardship maturity of the dataset and generating the Stewardship Report document.	
Access	Enabling data discovery and access	

Use	Enabling data utilization and application by the wide user community.	Use & Service Quality Aspect/Service Report
Service	Providing i) the secure and stable infrastructure for data users to find, obtain, and effectively use the data; ii) User support and engagement; and iii) venue for collecting user feedback.	
Reuse/Improve	Reused in purposes rather than initially designed for; Data users may not have in-depth knowledge of the dataset and science domain it was designed for. Improvement in data, metadata or service capability may be made in responding to users feedback. It may need to go to the planning phase for a new version of the data product.	
Document Type	**Description**	**Examples/Notes**
Science Report	Document that: i) describes the data product requirements or specifications including sensor or instrument characteristics if applicable, accuracy and precision and intended use, ii) describes the processing algorithm, how the algorithm is assessed and validated, iii) defines the process for developing and producing the data product, and iv) may include a data management plan for it to be accessed, stewarded and serviced.	NASA Mission/Program Document: Operation IceBridge Level-1 Science Requirements and Scientific Basis ISO 19131 compliant data product specification* One can simply view the Science Report as a description of the state of the scientific quality of the data product to be developed and plan for it to be accessed, stewarded and serviced. Once the product has been produced, the Science Report can also be used to benchmark the final product 'as is' against what it was intended to be.
Product Report	Document that describes: i) how the data product was generated (algorithm and process flow), ii) what input data (source and ancillary) were used including description of sensor or instrument characteristics if applicable, as well as iii) how the data product was evaluated and estimated product error sources or uncertainty.	One can simply view the Product Report as a description of the state of the product quality of the dataset that has been produced. An example of a product

		document is algorithm theoretical basis document (ATBD) that originated from NASA, commonly used for satellite data products. NOAA climate data record (CDR) program has adapted it and developed a template: NOAA C-ATBD Template; C-ATBD for OISST
Stewardship Report	Report documents the stewardship practices applied to the dataset and summarizes the current stage of stewardship maturity of those practices, including those for ensuring the scientific and product quality of the dataset and enabling data access and re(use).	One can simply view the stewardship maturity report as a description of the current state of the quality of stewardship practices applied to the dataset to enable data access and use. Data Stewardship Maturity Report
Service Report	Report summarizes a dataset use metric such as data download and citations, and impact metric. such as if the data has been used in national or international climate monitoring and assessment reports or by private sectors. User feedback for further improvement should also be described.	One can simply view the Service Report as a description of the current state of the quality of use and impact of the dataset. Website with dataset download metrics - to be added. NOAA Data Impact Reports: Retail and Manufacturing; Logistics and Transport; Reinsurance Or something similar to NASA EOSDIS annual data metrics report but for individual datasets.

* ISO 19131 can be used to describe specifications of the product, to benchmark the finished product, to potentially identify where the product did not meet specifications. In some areas this can be seen as a measure of 'quality' of the final product.

Appendix D. Dataset Quality Assessment Types

As outlined in section 4c, quality attributes and dimensions commonly assessed include accuracy, completeness, currency, relevancy, conformity, and consistency (e.g., Redman 1996; Austin et al. 2019). There are different ways to assess them through the entire data lifecycle and may be grouped as technical, scientific and stewardship assessments. The technical assessment regards the data and metadata files checks, while the scientific assessment consists of data content and cross-data content checks (Stockhause et al. 2012). As an example, when the evaluator looks for the metadata standard compliance such as compliance against the Attribute Convention for Data Discovery (ACDD),[20] here the evaluator is checking that the attributes describing the files are according to a set of community-recognized metadata characteristics (e.g., specific date-time format). There is no check of the file content, no scientific evaluation in this case, but a metadata conformity check, it is a purely technical assessment (Stockhause et al. 2012, Evans et al. 2017). One the other hand, when the evaluator plots the dataset variable and checks for reproducibility of El Niño events against skill metrics, here the scientific soundness of the data content is considered (e.g., Haiden et al. 2019).

In general, the technical assessment checks data files consistency among the distributed data and metadata repositories and conformance to formal standards. Checks can be as different as the specific needs of the service and may include compliance against community standards (conformity), temporal and spatial checks for unexpected gaps (completeness), identification of corrupted values (integrity). Lawrence et al. (2011) postulates a generic checklist for technical and scientific assessments. As far as the scientific assessment is concerned, this refers to scientific analyses of the physical content described by the dataset to check for its scientific soundness. Given the nature of this assessment, it is typically carried out by domain experts. Analyses may include uncertainty characterization, validation against reference datasets, reproducibility of temporal/spatial patterns. These are typical quality attributes that may be grouped as scientific (Ramapriyan et al. 2017) and are addressed in the scientific assessments. The two concepts overlap to some extent, but are distinct because the former refers to which quality dimension is considered, whereas the latter refers to how the quality attributes in the quality dimension can be evaluated. The technical and scientific assessments are often accompanied by the stewardship assessment to guarantee accessibility and understandability of the dataset distributed. Here we refer to stewardship as anything of relevance in dataset quality that is not associated with the data and metadata file content, e.g., documents accompanying the dataset describing how to use it. Typical examples regard the description of the algorithms or models used to produce and process the data, provision of the DOI and license of use, verified network address to access the data, and information about the archiving procedures. The goal is to ensure that the dataset is well documented, the processing chain is visible, the data readily obtainable and usable.

At times, the assessments described above are accompanied by maturity assessment models. These are formal approaches to support compliance verification, usually defined in discrete stages to evaluate practices applied in organizations, services or products. Maturity is meant as a desired or anticipated evolution from a more ad hoc approach to a more managed process (Peng 2018). Datasets associated with high maturity are produced following best practices of the community and in a more managed fashion, increasing user trust in the data record provided. It should be noted

[20] https://wiki.esipfed.org/Attribute_Convention_for_Data_Discovery_1-3

that a low maturity rating does not necessarily imply low scientific value for a dataset. It can happen especially for datasets managed by a single investigator that may be flagged to have low maturity due to poor quality in metadata, documentation and accessibility.

The types of assessments explained above indicate that there is considerable scope for subjective reviewer expertise, but some of the assessments are rather mechanical and amenable to automated checking. Automated checking is important for a number of reasons, the quality assessments need to be consistent (subjective variability in the analyses and human errors are minimized) and sustainable, especially when done on a routine basis and for a large number of datasets. This calls for efficiency and scalability, which can be supported by considering automated processes for evaluation. Furthermore, the quality information produced should be frequently updated to contain the most current details, which again requires automated processes in place. Automatization is possible if the assessments are machine-readable. A key element to produce dataset quality information that is both machine- and human-readable is to capture this information in standardized forms and distribute it digitized. To wrap up, given the need for scalability and machine readability, dataset quality assessments are recommended to be as much as possible automated, particularly in an operational environment. Operational environments are characterized by routine assessments that have to be “timely”, “frequently updated” and “scalable” (Leadbetter et al. 2020). Therefore, it is essential to evaluate quality attributes and represent the quality information in a systematic and consistent way. This implies that the data quality management (DQM) has to develop and maintain capabilities to assess and describe data quality.

Following the TRUST Principles for Data Repositories (Lin et al. 2020), the dataset quality assessment activities need to be supported by software, hardware, and technical services by implementing the relevant and appropriate standards, tools (e.g., Henzen et al. 2021) , and technologies for data management and curation. The technical developments underlying the automatization solution should not be underrated when budgeting the work for DQM. Pain is inefficiency and frustration of the domain experts who can better dedicate their knowledge in scientific added value analyses rather than in repetitive manual work. DQM should adopt clear definitions of the competencies, roles and responsibilities required for staff involved in the assessments, as well as develop plans for capacity building and training to ensure availability of the people with the competencies required (Leadbetter et al. 2020). Capacity building refers to human skills, organization tools and resources in general. It paves the way to an operational DQM function optimized in terms of cost-effective infrastructure and maintenance of tools.

All the above indicates that the production of comprehensive quality information requires cross-disciplinary expert knowledge and needs to be curated by domain experts ranging from science, data management, to computer engineering. This is extremely challenging for any single individual. Therefore, dataset quality practitioners are usually organized in well-established functions inside data dissemination services. Depending on the resources and maturity of the established function, the scope can extend from quality control only to comprehensive quality management activities. What guarantees that data is not corrupted and is accurate are the quality control procedures plus a series of protocols to avoid new quality issues in the system (e.g., database). This latter element is part of the quality assurance, which is a proactive process focused on "preventing defects" aiming at maintaining the desired level of quality in a data collection. In contrast, the quality control is a reactive process focused on "detecting defects" and encompasses

a set of procedures applied to identify and flag the errors to ensure that data available to users are sufficiently reliable to be used with confidence (ISO 9000 2005; WMO 2019).

Both quality control and quality assurance can consider documentation, 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, with the ultimate goal of increasing trustworthiness in data and information disseminated by the operational services. How to disseminate this quality-related information to end-users is an additional aspect of quality management. Dissemination is becoming increasingly important to guide users to understand the datasets and to promote data uptake. However, conveying dataset quality information in a manner that is understandable and usable to data users is often a challenge (Ramapriyan et al. 2017). At the current status, there is no consensus about a standardized way to disseminate this information and it also depends on the audience towards which this information is intended (Baker et al. 2016). The audience, i.e., the data users, may feedback which of the disseminated information is most relevant and what can be improved. User engagement activities are thus an additional component that may (or may not) be included in DQM.

The dataset quality information should evolve according to the user needs. Understanding "user satisfaction" or "meeting or exceeding user expectation" (Evans and Lindsay 2005) could prove helpful in characterizing data quality in a specific context. The user engagement has to establish several channels to collect the user requirements from the audience identified, which will benefit from the quality information disseminated. The requirements collected can be then analysed to reveal weaknesses, new needs and data issues.

Having a user engagement component in the DQM, including evaluating the understandability of the quality information with the users, may improve guidance in the strategic decisions for better dissemination of the dataset quality information and may help to identify issues with the data that might have escaped the quality control procedures, which in turn leads to improvement of the procedures applied. An additional benefit relates to the fact that involving users increases transparency in the DQM procedures and, more in general, increases trust about the data served. Co-designing with the users creates a strong sense of ownership of the product development process by the user (Hewitt et al. 2017). These aspects are well-aligned with the TRUST Principles for Data Repositories (Lin et al. 2020). Indeed, the TRUST principles highlight the need to make data quality assessed and disseminated for prospective users. This endeavour has to ensure that the expectations of target user communities are met in order to foster usability by enabling users to understand and assess dataset quality.

Appendix E. Additional Examples of Quality Assessment Models

This appendix provides additional examples of assessment models for the Earth science community:

- ATMODAT metadata assessment checklists (Ganske et al. 2020a; 2020b);
- Consistency of Satellite Climate Data Records for Earth System Monitoring (Popp et al. 2020);
- CORE-CLIMAX Product System Maturity Matrix (EUMETSAT 2013);
- Data Operational Readiness Levels (Moe et al. 2018; https://www.esipfed.org/orl);
- JPSS Data Product Maturity Matrix (Zhou et al. 2016);
- Metadata checklist by LTER (O'Brien et al. 2016);
- NASA IMPACT ARC Metadata Quality Framework (Bugbee et al. 2021);
- NASA Technical Readiness Level (Mankins 1995; 2009);
- NCEI/ESIP-DSC Data Use and Services Maturity Matrix (Serv-MM Working Group 2018);
- NOAA CDR Product Maturity Matrix (Bates and Privette 2012; Self-evaluation template available at: https://www.ncdc.noaa.gov/cdr/development-guidelines);
- Operational Readiness Levels (Moe et al. 2018; https://www.esipfed.org/orl); (ORL Ranking Tool - limited availability: https://survey123.arcgis.com/share/23ada947ba014cf19f651543c2ee8fb3)
- Quality assessment concept of the World Data Center for Climate and its application to CMIP5 data (Stockhause et al. 2012);
- Quality Assurance Templates (Nightingale et al. 2019);
- Quality Maturity Matrix used at DKRZ (Höck and Toussaint, 2019);
- RDA FAIR Data Maturity Indicators (RDA FAIR Data Maturity Model Working Group, 2020);
- RDA-SHARC Evaluation (Romain et al. 2018);
- WDS/RDA Assessment of Data Fitness for Use WG Checklist for Evaluation of Dataset Fitness for Use (Austin et al. 2019);
- Ocean Best Practices System (Pearlman et al. 2019). [21]

[21] https://www.oceanbestpractices.org/

Appendix F. Community Controlled Vocabularies and Content Standards

This Appendix provides some examples of controlled vocabularies and content standards in the Earth science community.

Climate and Forecast (CF) Metadata Convention (https://cfconventions.org) maintains and onboards standard variable names, standardized regional names and area types at the data file level. It is developed for the NetCDF data format but can be adopted by others.

Global Change Master Directory (GCMD) maintains and expands keywords for Earth science, services, data providers, projects, instruments/sensors, platforms/sources, locations, horizontal, vertical, and temporal data resolutions, URL content types, granule data formats, measurement types, and chronostratigraphic units – a total of thirteen sets of keywords currently (GCMD 2020). Each GCMD keyword is publicly available and assigned a universally unique identifier (UUID).

The British Oceanographic Data Centre (BODC) maintains a collection of controlled vocabularies developed by the marine science communities:
https://vocab.nerc.ac.uk/collection/

The United States Geological Survey maintains a Landsat glossary at:
https://www.usgs.gov/core-science-systems/nli/landsat/landsat-glossary

Global Climate Observing System (GCOS) maintains a list of essential climate variables that are critical to the characterization of Earth's climate (https://public.wmo.int/en/programmes/global-climate-observing-system/essential-climate-variables). The definitions can be found at:
https://www.ncdc.noaa.gov/gosic/gcos-essential-climate-variable-ecv-data-access-matrix.

Content specifications for preserving Earth sciences data products can be found in (Ramapriyan and Moses 2012), which is now a part of ISO standards (ISO 19165-2 2020).

The World Wide Web Consortium (W3C) maintains a controlled data catalog vocabulary (DCAT) and the current version can be found at:
https://www.w3.org/TR/vocab-dcat-2/

ESIP Attribute Convention for Data Discovery (ACDD) is maintained at:
https://wiki.esipfed.org/Attribute_Convention_for_Data_Discovery_1-3

Appendix G. List of Author Names, Affiliations, Roles and/or Subject Areas and ORCIDs

Name (1st Last)	Affiliation(s), Country	Sector(s), Roles and/or Subject Areas	ORCID
Ge Peng	The University of Alabama in Huntsville (UAH)/NASA MSFC IMPACT (Previously NCSU/NOAA NCEI), USA	Academic, Science Center; Data Producer, Data User, Scientific Steward, Scientific Data Stewardship	0000-0002-1986-9115
Carlo Lacagnina	Barcelona Supercomputing Center (BSC), Spain	Service Provider, Data Quality Management	0000-0001-9434-9809
Robert R. Downs	Columbia University, USA	Academic, Domain Data Center, Digital Archivist	0000-0002-8595-5134
Hampapuram Ramapriyan	Science Systems and Applications, Inc./NASA GSFC, USA	Government Contractor, Data Systems (Data Stewardship, Information Quality, Provenance)	0000-0002-8425-8943
Ivana Ivánová	Curtin University, AUS	Academic, Spatial Data Quality Research, Standardisation Expert	0000-0001-6836-3463
Anette Ganske	Technische Informationsbibliothek (TIB), Germany	Infrastructure Provider, Data Publication and Metadata Standards, Scientist	0000-0003-1043-4964
Dave Jones	StormCenter Communications | GeoCollaborate, USA	Private Sector, Data User/Applications, Service Provider	0000-0003-4573-2400
Lucy Bastin	Aston University, UK	Data Scientist	0000-0003-1321-0800
Lesley Wyborn	Australian National University, AUS	Academic, Domain Expert (geoinformatics)	0000-0001-5976-4943
Irina Bastrakova	Geoscience Australia, AUS	Spatial Data Architect	0000-0002-4643-7289
Mingfang Wu	Australian Research Data Commons (ARDC), AUS	Research Data Specialist, Information Retrieval, Metadata	0000-0003-1206-3431
Chung-Lin Shie	University of Maryland at Baltimore County/NASA GSFC	Academic, Scientist, Data Provider, Domain Expert	0000-0002-1115-1029
David Moroni	NASA Jet Propulsion Laboratory/California Institute of Technology, USA	NASA PO.DAAC Data Center, Data Manager	0000-0003-2994-557X
Gilles Larnicol	BSC/Magellium, Spain/France		

Yaxing Wei	NASA Oak Ridge National Laboratory (ORNL), USA	NASA ORNL DAAC Data Center, Center Scientist	0000-0001-6924-0078
Nancy Ritchey	NOAA's National Centers for Environmental Information (NCEI), USA	NOAA NCEI Data Center, Archive Manager, Domain Expert (Data Management and Stewardship)	0000-0003-3939-6287
C. Sophie Hou	Apogee Engineering/USGS; Ronin Institute, USA	Governmental Contractor, Data Usability Specialist	0000-0002-8087-1775
Ted Habermann	Metadata Game Changers, USA	Domain Expert (Metadata, Metadata standards)	0000-0003-3585-6733
Sarah Champion	North Carolina State University, USA	Data Architect, Information Quality Management and Metadata Specialist	0000-0002-5080-6286
Gary Berg-Cross	Ontolog Forum, USA	Knowledge Engineer, Domain Expert (Ontology/Semantics)	
Kaylin Bugbee	NASA Marshall Space Flight Center (MSFC), USA	Informatics	0000-0001-6733-5698
Jeanné le Roux	UAH/NASA MSFC, USA	Informatics	0000-0002-8274-987X

Appendix H. Acronyms

ACDD	Attribute Convention for Data Discovery, ESIP
AHC	All Hazards Consortium
APT	Algorithm Publication Tool, NASA IMPACT
ARC	Analysis and Review of the Common Metadata Repository (CMR), NASA IMPACT
ARDC	Australian Research Data Commons
ATBD	Algorithm Theoretical Basis Document
ATMODAT	Atmospheric Model Data
AU/NZ DQIG	Australia/New Zealand Data Quality Interest Group
BSC	Barcelona Supercomputing Center
CASRAI	Consortia Advancing Standards in Research
CC BY	Creative Commons Attribution License
CC0	Creative Commons Public Domain Dedication
CDR	NOAA Climate Data Record
CEOS	Committee on Earth Observation Satellites
CF	NetCDF Climate and Forecast
CICS	Cooperative Institute for Climate and Satellites, NOAA
CODATA	Committee on Data of the International Science Council (ISC)
CORE-CLIMAX	COordinating Earth observation data validation for RE-analysis for CLIMAte ServiceS
CREWS	Consortium for Research on Environmental Water Systems
DAAC	Distributed Active Archive Centers, NASA
DataONE	Data Observation Network for Earth, NSF
DCAT	Data Catalog Vocabulary
DFT	RDA Data Foundation and Terminology
DHS	US Department of Homeland Security
DKRZ	German Climate Computing Center
DMS MM	WGISS Data Management and Stewardship Maturity Matrix
DOI	Digital Object Identifier
DQM	Data Quality Management
DSMM	Data Stewardship Maturity Matrix
DSMR	Data Stewardship Maturity Report
ECMWF	European Center for Medium-Range Weather Forecasts
EEI	Edison Electric Institute
EOSDIS	Earth Observing System Data and Information System, NASA
EPSCoR	Established Program to Stimulate Competitive Research, NSF
ERDDAP	Environmental Research Division's Data Access Program, NOAA
ESA	European Space Agency
ESIP	Earth Science Information Partners
ESRI	Environmental Systems Research Institute
ET-DRC	Expert Team on Data Requirements for Climate, WMO
EUMETSAT	European Organisation for the Exploitation of Meteorological Satellites
FAIR	Guiding principles for Findable, Accessible, Interoperable, and Reusable
FEMA	US Federal Emergency Management Agency
FGDC	Federal Geographic Data Committee
GCMD	Global Change Master Directory, NASA
GSFC	Goddard Space Flight Center, NASA
HTTPS	Hypertext Transfer Protocol Secure
IMOS	Integrated Marine Observing System
IMPACT	Interagency Implementation and Advanced Concepts Team, NASA MSFC

IQC	Information Quality Cluster, ESIP
IS-ENES	Infrastructure for the European Network for Earth System Modelling
ISO	International Standards Organization
JPSS	Joint Polar Satellite System, a collaborative program between NOAA and its acquisition agent, NASA
LTER	Long-Term Ecological Research, NSF
MSFC	Marshall Space Flight Center, NASA
NASA	National Aeronautics and Space Administration
NCA	US National Climate Assessment
NCDC	National Climatic Data Center, NOAA
NCEI	National Centers for Environmental Information, NOAA
NCSU	North Carolina State University
NESDIS	National Environmental Satellite, Data, and Information Service, NOAA
NetCDF	Network Common Data Form
NICC	DHS National Infrastructure Coordinating Center
NOAA	National Oceanic and Atmospheric Administration
NREs	National Response Events
NSF	National Science Foundation
NSIDC	National Snow and Ice Data Center, NASA
OAI-PMH	Open Archives Initiative - Protocol for Media Harvesting
OGC	Open Geospatial Consortium
OISST	NOAA daily Optimum Interpolation Sea Surface Temperature
ORCID	Open Researcher and Contributor ID
ORL	ESIP Operational Readiness Levels
ORNL	NASA Oak Ridge National Laboratory
OWL	Ontology Web Language
PAMARCMIP	Pan-Arctic Measurements and Arctic Regional climate model simulations
PDCA	Plan-Do-Check-Act
PID	Persistent Identifier
QA/QC	Quality Assurance and Control
QMM	Quality Maturity Matrix
R2R QA	Rollingdeck to Repository Quality Assessment dashboard
RDA	Research Data Alliance
RDFS	Resource Description Framework Schema
RMAGs	EEI Regional Management Assistance Groups
SHARC	RDA-Sharing Rewards and Credit (SHARC) interest group
SISE	Sensitive Information Sharing Environment
SMM	Stewardship Maturity Matrix
SMM-CD	WMO Stewardship Maturity Matrix for Climate Data
SNSF	Swiss National Science Foundation
THREDDS	Thematic Real-time Environmental Distributed Data Services
TIB	Technische Informationsbibliothek, Germany
TRUST	Principles for Digital Repositories: Transparency, Responsibility, User Focus, Sustainability, Technology
UAH	University of Alabama in Huntsville
URI	Uniform Resource Identifier
USGCRP	US Global Change Research Program
USGS	US Geological Survey
UUID	Universally Unique Identifier
W3C	World Wide Web Consortium

WDCC	World Data Center for Climate
WDS	World Data System
WGISS	Working Group on Information Systems and Services, CEOS
WIGOS	Integrated Global Observing System, WMO
WMO	World Meteorological Organization

Appendix I. Community Comments and Responses

Starting Line #No	Content Type (phase/ Paragraph/ Figure/Table)	Comment Type [22]	Comments	Proposed Change	Observations of Project leaders
464	Phase	ge	"*Appendix C*" Text highlighted in red.	Change "C" to "C".	As a demo. Agreed. Implemented in the revised version.
716	Sentence	ed	"We thanks .." Grammar error.	Replace "thanks" by "thank".	As a demo. Agreed. Implemented in the revised version.
0	Filename	ed	"FARI_": Typo	Change it to "FAIR_".	Agreed. Filename modified and uploaded to OSF (4/22/2021)
53	TOC	te	Missing subsections 4f-h in the table of contents	Include subsections 4f-h in TOC	Agreed. Implemented.
294	Subsection Title	ge	Subsection title indent not consistent	Adjust indent of subsection title 3e.	Agreed. Implemented.
630	Subsection Title	ge	Inconsistent subsection title	Bold subsection title 4g.	Agreed. Implemented.
All	Whole document	ge		file size can be compressed a lot, as it seems without loss of information.	Yes, the document contains a lot of information for background and examples for practical purposes. We did, however, removed Appendix G as it is already included in Peng et al. (2021)
0		ge	Are you recommending people should use ISO 19157, or ISO 19115-1, as the structure for documenting the data quality?		ISO 19157 or 19115-1 is one of the methods that people can use to report quality assessment reports but we are not endorsing any particular method.
0		ge	Are the guidelines only pertaining to Earth science datasets?		The guidelines are primarily developed for the Earth science community. However, they are general enough to be readily adopted to other disciplines.
All	Whole document	ed	Both "dataset" and "data set" are used	Consistent use of either one.	Agreed. All "data set" are replaced by "dataset" in the revised version.
399	Sentence	ed	possible typo?	Replace "document" with "documented".	*"document" is the object of "captured". Added "a" in front of "document" to be clear*
453	Phase	ed	should this be referring to Appendix F (rather than Appendix E)?	Replace "Appendix E" by "Appendix F"	Agreed. Implemented in the revised version.

[22] ge-general, te-technical, ed-editorial

459; 475	Word	ed		replace “exhausted” with “exhaustive”	Agreed. Implemented in the revised version.
562		te	reference to “schema.com”. Looks like the wrong link? Or this domain has been taken over at least because this redirects to a private company. I wondered if this was supposed to reference schema.org but that doesn’t seem right within this context since you can’t mint PIDs with them.		Good catch. It should be “schema.org”. Implemented in the revised version.
1258	Appendix F	ge	NVS would be a worthy mention. It is a large collection of vocabularies that we use for marine science and other applications. http://vocab.nerc.ac.uk/collection/ (side note: many of these vocabularies are related and combined to for complex descriptions: https://github.com/nvs-vocabs/P01/blob/master/P01_wheel.pdf)		Thanks for the suggestion. It is included in the revised version.
715	Sentence	ed	“helping laying …”: grammar error	Replace “laying” by “lay”	Agreed. Implemented in the revised version.
681	Sentence	ed	“which in turns”: grammar error	Replace “turns” by “turn”	Agreed. Implemented in the revised version.
		ge	What I highly appreciate is an increased emphasis on open science: it is regularly referenced, and extends as well to software and methodology. In that sense, a good step forward is being made. Scientists and applications of spatial data will need more tools and information to explore the quality of the data. I further notice that you use the terms accuracy, precision and fitness for use. But I am missing the error (or uncertainty) propagation. That is for me always the key issue of SDQ: how to quantify the quality of the product that is obtained with publicly available data, modeling and software, that possibly was collected/created with other intentions than those for which it will be used. I could imagine that here open science can make a huge step. An interesting and somewhat unexplored component in open science are its dangers and restrictions – maybe that leads to a nice and interesting new quality criterion: the risk of abuse, i.e. second use or unethical use, of data (and models and software).		Appreciate the positive comments and the great points on i) uncertainty propagation and ii) how to ensure the data, software, and information are being used correctly and ethically, i.e., the risk of abuse. Both are big challenges for the community to address, which is beyond the scope of this document. However, providing quality information including uncertainty in various stages/aspects should support efforts aiming to address either one of those challenges – whether and how useful will be a great use case for the community to develop. A discussion has been included in the revised document.
80	Word/Phase	ge	It would be good to see FAIR described briefly when it is first introduced in the exec summary and further in the background section		Agreed partly - the FAIR guiding principles were already described in the background session, hence left unchanged. ‘FAIR’ abbreviation expanded in the Executive Summary.
1	ge	ge	General comment that further information on communication of results to non specialists would be		It is an excellent point but it may be better considered in an implementation guidance document.

			useful i.e. what is a minimum set of reporting requirements for the decision maker and how does one arrive at this set of requirements		
Not a specific location set.	Paragraph - Technical and Citizen Science	te	I note two key things I would value seeing added: 1) that the role of technology design in influencing data quality, which specific discussion around different approaches to capturing knowledge; and 2) the discussion of intersections between classic geospatial data and citizen science can augment each other, empowering scientists and communities alike, if information is provided via FAIR principles.		These are great points, however, at this version of the guidelines they are out of the scope. We revised the introduction and conclusion by highlighting that these should be considered long-term.
92-93	Sentence	ge	We fully agree on the statement about quality assessment regarding representation and integration across systems and tools and kindly suggest to reference our paper about recommendations for data management plans (DMP) (section 2.2 recommends to include structured quality information in DMPs; section 3, in particular 3.1 discusses linking and sharing across tools, in our case data management system, DMP tool, and knowledge hub). Further, we suggest to reference our paper about a research data infrastructure component to extract quality information (section 3.1/figure 1 indicate that (automatically generated) quality information can be used in several phases of the research data lifecycle, section 3.4/figure 2 stresses the importance linking of tools).	Quality assessments also need to be consistently represented and readily integrated across systems and tools to allow for improved sharing of quality information at the dataset level **(cp. Henzen et al. 2021, Wagner et al. 2021)** for individual quality attributes or dimensions such as those defined in Wang and Strong (1996) or Ramapriyan et al. (2017).	Content integrated as suggested.
98-99	Sentence	ge	We agree on that and suggest to reference our paper about the Geo-Dashboard concept, which includes findings on current data quality challenges (see for instance section 2.2; examples in table 1), e.g. lacking guidance; several sources for quality information; quality information on different levels of details; requirements for user-friendly presentations.	Although the need for assessing the quality of data and associated information at the individual dataset level is well recognized, methodologies for an evaluation framework and presentation of resultant quality information to end users (cp. Figgemeier et al. 2021) may not have been comprehensively addressed within and across disciplines.	Content integrated as suggested.

138	Heading	ge	We use the term scope slightly different. In this section, some framing concepts like dataset, data life cycle are presented. Thus, we would recommend to clarify what is meant with scope at the beginning of the paragraph or change the heading to (conceptual) framework.		The “scope” is a standard term to use in the literature for the extent of the area or subject matter that something deals with or to which it is relevant. Therefore, scope seems more appropriate to identify the target of a document. No change is made.
154ff		ge	Considering the term quality dimension, we’d like to remark: In some specifications, the term dimension is used differently. In the Data Quality Vocabulary, for instance, quality dimension covers criteria for assessing quality (https://www.w3.org/TR/vocab-dqv/#dqv:Dimension) and links to ISO/IEC 25012 dimensions (https://www.w3.org/TR/vocab-dqv/#DimensionsOfISOIEC25012)		We agree that the term dimension was lacking consistency particularly regarding figure 1 compared to the use of this term in the other parts of the document, thus we changed the term “dimension” inside figure 1 in favour of “aspect”.
159-160	Sentence	ge	Dataset quality information and quality elements seems to be used synonymously? QE is used only two times in the text.	We suggest harmonizing terminology eventually. Change quality element to dataset quality information	ISO 19115-1 and 19157 use quality elements. The two occurrences mentioned: One occurrence was to describe ISO standards and another was to state that quality elements used in ISO standards are equivalent to “quality attributes”. We have modified the sentence for the first occurrence but kept the second. Aiming to use common terminology was one of our main goals but unfortunately we have realized quickly that this issue is bigger than us – even ISO standards may use the same term differently and we have tried to use consistent terminology/concepts for our examples as a way forward.
163-166	Sentence	ge	Such a statement would fit as well the rationale of data management plans or at least the rationale of how a data management plan should be. Several communities/working groups (like the RDA WG Discipline-specific Guidance DMP) are currently discussing data management plans, e.g. how to provide discipline-specific information or how to guide researchers to provide useful information about data quality, etc..	There should be a link to modern living data management plan, which aim exactly at fulfilling such a statement.	The references to the community activities have been added, including those associated with data management plans such as the recommended RDA WG on Discipline-specific Guidance DMP.

172			This is an important statement and includes several aspects: (1) quality information should be gathered/tracked on a suitable level of detail and (2) should be presented in user-friendly visualizations (e.g. a quality dashboard) to support evaluating fitness for purpose/use. Therefore, we suggest to use our Geo-Dashboard concept as an example and to stress the relevance of linked provenance information (as mentioned in the FAIR principles).	Describing the quality of a data product and providing access to such quality information can support potential users of a particular dataset to determine whether it is appropriate for their planned usage, i.e., fitness for purpose. Figgemeier et al 2021 propose a Geo-dashboard to visualize quality information and related provenance information.	Instead of the place mentioned, reference to this work has been included in 4f guideline 5.
176	Sentence	ge	In our opinion, in appendix G, there is a real-life example of why quality information is critical. However, in we don't see any particular detailed description or practical example of what was improved in terms of quality. Which quality aspect were improved? Quality assurance? Quality controls? Quality information?	We suggest to remove "detailed" or, even better, add much more practical details of what was done in term of improved data quality.	Word “detail” removed.
209-211	Sentence	ge	This sentences includes lots of highly relevant aspects. However, we would add the availability / providing quality information via human- and machine-readable interfaces. Within our project GeoKur, we do not only see cross-discipline challenges, but even within a discipline certain quality information are reported differently, e.g. uncertainty/thematic accuracy for remote sensing data are described with different metrics, using different ground truth, etc.. Thus, we suggest to add challenges in sharing/using quality information within a discipline. Does “fully traceable” include the automatic extraction of quality information? We addressed some aspects related to the automatic extraction of quality information for earth system science data in our paper, which could be of interest for your guidance document. We suggest to revew Wagner et al. 2021, which also covers technical aspects	For dataset quality information to be effectively (re)used, it needs to be consistently curated, fully traceable, made available via adequately human- and machine-readable interfaces, adequately documented, updated timely, able to support users to address their specific needs. This is, however, a daunting objective because it necessitates both a wide range of data quality attributes and heuristic information to ascertain fitness for purpose, while facing challenges in cross-disciplinary and even in discipline-specific knowledge integration (Peng et al. 2020).	Part of the text was added, not the first part because this is an outcome of the guidelines not introduction. The paper Wagner et al. 2021 has been added in other parts of the document to refer the reader to the details mentioned here.

			on how to include extraction tools in research data management infrastructures – sharing quality information across tools (Data Management tools, DMP tools, etc.		
219f	Sentence	ge	What is about vocabularies, such as the data quality vocabulary (DQV) (https://www.w3.org/TR/vocab-dqv/)? GeoDCAT, a well-known and well-used metadata schema for geodata similar to ISO19115, includes extension points (via DCAT) to use the DQV. Thus, we recommend to add DQV.	despite the fact that international standards or vocabularies for describing the quality of geographic data have been in place since 2003 (e.g., ISO 19157: 2013; ISO 19115-1:2014, DQV).	Text changed accordingly as W3C 2016.
219-220	Sentence	ge	We fully agree on that statement. However, we would appreciate, if you review our papers on automated quality information extraction (Wagner et al. 2021) and our Geo-Dashboard concept for visualizing quality and provenance information (Figgemeier et al. 2021) to be linked here.	Add on Line 219: "Despite some efforts (Figgemeier et al. 2021, Wagner et al. 2021), dataset quality information is not routinely…"	Text added accordingly.
229			We do not fully understand this. Do you include provenance in quality information? We see a strong linkage and recommend to visualize provenance and quality, but we would not use the term provenance as part of quality.		We agree that the statement sounds ambiguous, reference to provenance was removed.
231-233	Sentence	ge	We agree on that, but we would add that at least some quality information can be extracted automatically from geospatial files, e.g. comission/omission. Our metadata extraction tool, for instance, can be used to automatically extract quality information from geodata files (Geopackage, GeoTiff, CSV, Shp). Users just need to run the tool, expert knowledge on quality information is not needed for that. However, the set of extracted measures is still limited.	A frequently cited barrier against documenting the quality of spatial data is that it mostly requires special domain-expert technical knowledge, while documenting general metadata can be done automatically or by non-specialists (Coetzee 2018, Wagner et al. 2021). Or you could add: However, a limited set of quality information can be automatically extracted from files (Wagner et al. 2021).	The first option has been integrated in the document.
245			As stated before, FAIR principles address provenance and we recommend linking provenance and quality information, getting insights		It is a good point to take into account in a later version of these guidelines. For the general purpose of this first version of the guidelines, we consider that highlighting

			for quality changes based on provenance information. Thus, starting with quality information on dataset level is fine. However, we suggest indicating the importance of quality information for data series or parts of datasets (e.g. quality for certain regions differs).		quality information at dataset level is already a good starting point.
243	Sentence	ge	it is not strictly necessary, but the FAIR work "didn´t end" in 2016. You might add some more up-to-date references to highlight that the FAIR work is in constant progress.	Add eventually a short paragraph highlighting the FAIR work in progress and cite some more up-to-date references https://doi.org/10.15497/RDA00035 (2019). https://doi.org/10.3233/ISU-170824 (2017) https://doi.org/10.1038/sdata.2018.118 (2018)	Original text is already full of other references for years after 2016. No more added.
326		te	We are wondering, if Coryea et al 2006 is a relevant reference? In any case, it has the wrong citation key.	We suggest removing or fixing the citation key (Cordy et al…).	Citation fixed.
371-372	figure	ge	The workflow can be simplified and improved in its design.	We suggest a circular flow with a central iterative concept (monitoring and improvement). The specification might be an overarching title/concept having underneath quality and evaluation. The same for the subsequent two phases can have an overarching title/concept (assessment) with underneath execution and dissemination.	We prefer to keep the current workflow picture. Indeed, a circular flow would imply that quality dissemination feeds quality specification, which is not true. What feeds quality specification is the monitoring/improvement, which is partially input by the last step but it is also input from all the other steps. Moreover, quality specification is input from steps that are not in the figure, such as vision management and user requirements. The input is not specified on purpose, while a circular plot would indicate that all inputs are considered by quality dissemination.
418-419		te	We suggest keeping the terminology simple and avoiding some complex terms, which were never used and defined in the text, e.g. "preservation process" or "stewardship workflow". This would help the reader to focus on the content.	Adopt terms defined in figure 1 (Dataset lifecycle). For instance: The preservation it is defined as a stage within the stewardship quality dimension, not as a process. Instead of stewardship workflows, could be used the stewardship dimension?.	Fixed with having just "stewardship".
421-423	sentence	ge	The same concept was stated on lines 398-400	You might remove redundancies	Removed.
479ff	sentence	ge	It might be useful to include some examples for proper metadata schemas or at least to address that		Examples are already provided below the text referenced.

			metadata schemas already include (some of) the listed elements as mandatory fields		
453		te	Typo/incorrect appendix reference	Change Appendix E with Appendix F	Already fixed because of another similar comment.
500-501	sentence	ge	Is license optional?	If the license is optional, it should be clarified that without license usage of public available data is critical. Basic rule of thumb is that without license you are not allowed to do anything without risking copyright infringement.	In Guideline 4 we encourage data quality information providers to state the license related to quality information. The information about the license indicates its obligation, limitations and repercussions
527-539	sentence	ge	We consider the sentence from line 525 (“if no suitable…”) to 527 (”…and accessible.”) to be sufficient in the case no assessment model exists. The rest of the text from line 527 to 539 is an somehow a repetition of what is already stated in guideline 2. However, if you like to keep the sentences, adding GitHub as an example for publishing the model might makes sense. It could be used to share implementation solution for the quality model or for community support (issue tracking for further developments, etc.). Moreover, a GitHub repo is typically still available, when projects are finished.	Remove/Shorten the paragraph from 527 to 539.	Removed 527 to 529, the rest is not a repetition.
530		te	Two links land on the same page.	Remove one web link.	One link removed.
543	sentence	ge	We recommend to make more explicit suggestions here, e.g. for comparing quality attributes / assessment results for several datasets and for the visualization of quality assessment results (e.g. as chart; or along a provenance graph), we need (named/meaningful) links. In earth system science (e.g. in remote sensing), analysing the results and evaluating the fitness for use based on quality information is somehow related to ground truth/training data. Therefore, we suggest to add "include descriptions/links to training data.”		Not sure it is wise to be more explicit, it is easier to be biased by prescribed quality dimensions and it is what we want to avoid. Indeed, even the comment is biased, “quality information is somehow related to ground truth”, this is specific to satellite data and specific to scientific quality validation, not even all the scientific quality range.

555	sentence	ge	We recommend to suggest adapting an existing framework and if not possible to develop a new quality framework. Adapting an existing schema or model includes providing information on the adaption to enable linking to the original schema or model.		Sentence has been rephrased.
541-562	sentence		We developed a metadata profile to include detailed quality information including further project requirements. We suggest to review our best practice document (describing the profile; and the process how to develop such a profile) and probably include it as a reference		Reference to their work is included as additional examples in the guideline.
630	heading	ed	Heading should be written in bold	Add bold formatting	Already fixed because of another similar comment.
634	table	te	Table 1, second row “guideline 1” F, R1 - Isn´t missing a number beside the F?	Add the proper number corresponding to the Findable principle F (F1? F2?)	The letter F, A, I or R in the right column denotes that the guideline from the left column can crosswalk to all criteria of being findable, accessible, interoperable, or reproducible, respectively, while the number (*n*) after the letter of F, A, I, or R refers to the *n*th criterion in that aspect of the FAIR. A footnote has been added to the table improve the clarity
654	table	ge	As presented in the latest working group meeting, the usability element will be removed from the ISO19157 in the new version.	If available, you might choose another example not referring to the usability element.	Replaced “Usability” by “Accessibility”
672		ed	These are conclusions.	You might rename the chapter in conclusions.	Chapter renamed to “Summary and Conclusions”.
1054		te	Are perspective and dimensions synonymous?	The term perspective is used in the title of appendix C but is never defined. The single time is seems to be defined is on line 405. Please, define what perspective is and check carefully this term throughout the text. Perspective is a synonymous of stakeholder (see line 1045).	“perspective” is replaced by “aspect” to be consistent with that of Ramapriyan et al. (2017).
1097	table	te	In table C1, links are not working or hyperlinks are missing	We suggest to provide a working web link example for each single document type.	The links available work.

1167-1173			We fully agree on that and see a strong relation to our ideas/recommendations on future DMPs and related tools (Henzen et al. 2021). Thus, we suggest to slightly modify the sentence and reference our paper.	We suggest modifying the sentence of line 1167-1169 as follows: “The dataset quality assessment activities can be greatly improved by adopting state of the art DMP tools (Henzen et al. 2021) implementing the relevant and appropriate standards, tools,…”	Reference added.
1252		te	The weblink is not working	Remove the Link. Reference is enough	Link works for us, but removed because indeed reference is enough.
1254		ge	Weblink is unnecessary	Remove the Link. Reference is enough	Link works, but removed because indeed reference is enough.
		ge	What I highly appreciate is an increased emphasis on open science: it is regularly referenced, and extends as well to software and methodology. In that sense, a good step forward is being made. Scientists and applications of spatial data will need more tools and information to explore the quality of the data. I further notice that you use the terms accuracy, precision and fitness for use. But I am missing the error (or uncertainty) propagation. That is for me always the key issue of SDQ: how to quantify the quality of the product that is obtained with publicly available data, modeling and software, that possibly was collected/created with other intentions than those for which it will be used. I could imagine that here open science can make a huge step. An interesting and somewhat unexplored component in open science are its dangers and restrictions – maybe that leads to a nice and interesting new quality criterion: the risk of abuse, i.e. second use or unethical use, of data (and models and software).		Appreciate the positive comments and the great points on i) uncertainty propagation and ii) how to ensure the data, software, and information are being used correctly and ethically, i.e., the risk of abuse. Both are big challenges for the community to address, which may be beyond the scope of this document. However, providing quality information including uncertainty in various stages/aspects should support efforts aiming to address either one of those challenges – whether and how useful will be a great use case for the community to develop. We could include this further research in the next version of the guidelines.
142	Phrase	ge	“starts at the planning and designing stage of a data product after data collection” is not full data lifecycle	Good QI has to be planned for, starting before data collection, e.g. capturing sensor calibration information	We will not treat this part specifically, but we expanded further this sentence to highlight this limitation of the guidelines and the importance of the pre-data collection stage for QI.
470	Section Heading	ge	Section 4f not in ToC		Added in ToC.

479	Phrase	ge	Guideline 1	This is simply an elaboration on FAIR principles, which are already specified in several FAIR implementation guidelines/metrics	Yes, this guideline is about how to make a dataset findable, accessible and potentially reusable based on the existing FAIR implementation guidelines.
503	Phrase	Ge	Guideline 2	At a high level it should be possible to create "a structured quality assessment model" that is domain agnostic. It seems like that should be a goal of this publication	This guideline recommends using a structured quality assessment model and what can be done to ensure the assessment is searchable and retrievable. The goal of this document is not about how to create a quality assessment model.
512	Phrase	ge	Examples	If these are good examples, they should have a DOI	Yes they all have a DOI, as reported in the reference section.
609	Notes	ge	Guideline 5	More should be done to summarize/synthesize the best practices for Guideline 5	Revised and improved.
637	Phrase	te	19115 (2014)	19115-1 (2014)	Fixed.
687	Phrase	gr	Community guidelines … is one step closer	Community guidelines … are one step closer	Implemented in the sentence which is also rephrased.
305	Phrase	ge	"Journal editors and reviewers may refer to the guidelines when assessing data that are associated with manuscripts under evaluation for potential publication"	They would not refer to these guidelines, but rather to a community-specific implementation of DQI as guided by this doc.	There is no reason why journal editors and reviewers could not refer to the guidelines, hence we keep this option on the list. We also believe that editors and reviewers would refer to the guidelines first and then, for more context, to any available community implementation of these
0	Filenam e	ed	"FARI_": Typo	Change it to "FAIR_".	Agreed. Filename modified and uploaded to OSF (4/22/2021)
0		ge	Are you recommending people should use ISO 19157, or ISO 19115-1, as the structure for documenting the data quality?		ISO 19157 (soon to be superseded by 19115-1) is one of the methods that people can use to report quality assessment reports but we are not endorsing any particular method.
0		ge	Are the guidelines only pertaining to Earth science datasets?		The guidelines are primarily developed for the Earth science community. However, they are general enough to be readily adopted to other disciplines.
53	TOC	te	Missing subsections 4f-h in the table of contents	Include subsections 4f-h in TOC	Agreed. They are added to the TOC in the revised version.

294	Subsection Title	ge	Subsection title indent not consistent	Adjust indent of subsection title 3e.	Agreed. Implemented.
399	Sentence	ed	possible typo?	Replace “document” with “documented”.	*“document” is the object of “captured”. Added “a” in front of “document” to be clear*
453	Phase	ed	should this be referring to Appendix F (rather than Appendix E)?	Replace “Appendix E” by “Appendix F”	Agreed. Implemented in the revised version.
464	Phase	ge	“*Appendix C*” Text highlighted in red.	Change “C” to “C”.	Agreed. Implemented in the revised version.
562		te	reference to “schema.com”. Looks like the wrong link? Or this domain has been taken over at least because this redirects to a private company. I wondered if this was supposed to reference schema.org but that doesn’t seem right within this context since you can’t mint PIDs with them.		Good catch. It should be “schema.org”. Implemented in the revised version.
630	Subsection Title	ge	Inconsistent subsection title	Bold subsection title 4g.	Agreed. Implemented.
681	Sentence	ed	“which in turns”: grammar error	Replace “turns” by “turn”	Agreed. Implemented in the revised version.
715	Sentence	ed	“helping laying …”: grammar error	Replace “laying” by “lay”	Agreed. Implemented in the revised version.
716	Sentence	ed	“We thanks ..” Grammar error.	Replace “thanks” by “thank”.	Agreed. Implemented in the revised version.
459; 475	Word	ed		replace “exhausted” with “exhaustive”	Agreed. Implemented in the revised version.
1258	Appendix F	ge	NVS would be a worthy mention. It is a large collection of vocabularies that we use for marine science and other applications. http://vocab.nerc.ac.uk/collection/ (side note: many of these vocabularies are related and combined to for complex descriptions: https://github.com/nvs-vocabs/P01/blob/master/P01_wheel.pdf)		Thanks for the suggestion. It has been included in the revised version.

1004	paragrap h	ed	"… may different …"	Replace "… may be different …"	Thank you for picking this up - the omission has been corrected.
1019	Table	ed	"Data are … or other entities ..."	Replace "entities" by "phenomena"	Agreed. Replaced as suggested.
1019	Table	te	"Data can be either structured or unstructured ..."	Add: "analog or digital"	Agreed. Added as suggested.
1019	Table	te	"Data in its physical form can be air, fish, or ice core samples..." This would mean that any (physical) object (or any phenomena?) is "data". To me this makes no sense as it detaches the term "data" from being a "representation" (of an object or phenomena). In this way data and objects become indistinguishable!	Delete sentence and, if no better examples are found, abandon the category of "physical data".	Agreed, partly - in Earth Science, the term 'physical data' is well established. We revised the sentence to differentiate between data and object.
1019	Table	te	Definition of "data set" uses the word "collection" which is used also for another separately defined term. This is confusing.	Revise definition(s) and clearly distinguish "data collection" from "data set" (or declare them synonymous). Use consistent spelling "data set" or "dataset", don't mix!	Agreed, partly. We revised the guidelines for consistent use of 'dataset'- there are now no occurrences of 'data set'. We use the definition of a dataset from ISO 19115-1, which says 'a dataset is a collection…' We reviewed the guidelines for the use of 'collection' and we believe that it is consistent with the definition of both, 'dataset' and 'data collection', hence no confusion is expected.