International Community Guidelines for Documenting and Sharing Dataset Quality Information

Ge Peng[1], Robert R. Downs[2], Hampapuram Ramapriyan[3], David F. Moroni[4], and Yaxing Wei[5]

[1] UA Huntsville; [2] Columbia University/SEDAC; [3] SSAI/GSFC;
[4] JPL/PO.DAAC; [5] ORNL

ESIP Information Quality Cluster, Co-chairs

ESIP IQC Monthly Telecon, 22 February 2022, Virtual

Definitions

Dataset: an identifiable collection of data - may contain one or many data files or records in a database in an identical format, having the same variable(s) and product specification(s).

Dataset Quality (Not Just Data Quality) includes:

- Quality of data (input and output),
- Quality of software and workflows,
- Quality of metadata and documentation,
- Quality of procedures, tools and systems.

Dataset Quality Information: Information about quality or the state of *data, metadata and documentation* through the entire lifecycle of a dataset:

- Data collection, acquisition or production, data and information management, data publishing and services, customer support and user engagement.

Needs, Challenges, and Benefits of Documenting and Sharing Dataset Quality Information

Not talk about today; References provided below for more details:

- **Case Statement and 2020 Workshop Summary Report:** https://doi.org/10.31219/osf.io/75b92
- **Call-to-Action Statement Paper:** http://doi.org/10.5334/dsj-2021-019
- **FAIR DQI White Paper:** https://doi.org/10.31219/osf.io/xsu4p

➔ Changing data and user community paradigm,
➔ Data and information quality dimensions,
➔ Across domain information and knowledge integration,
➔ Fitness for purpose.

Guidelines Are Developed By

International FAIR-DQI Community Guidelines Working Group

Ge Peng, Carlo Lacagnina, Ivana Ivánová,

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

(Many are IQC members!)

A Community of Practice

22 Global Domain Experts

- **from** government, academic, and private sectors
  - Data, science and service centers, institutional repositories, companies
- **with** expert knowledge on
  - data production, metadata curation, data publishing, services, standards, data systems, applications, etc.

Inspired by the FAIR Data Principles

Findable

The first step in (re)using data is to find them. Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an essential component of the **FAIRification process**.

F1. (Meta)data are assigned a globally unique and persistent identifier

F2. Data are described with rich metadata (defined by R1 below)

F3. Metadata clearly and explicitly include the identifier of the data they describe

F4. (Meta)data are registered or indexed in a searchable resource

Interoperable

The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.

I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

I2. (Meta)data use vocabularies that follow FAIR principles

I3. (Meta)data include qualified references to other (meta)data

Accessible

Once the user finds the required data, she/he/they need to know how they can be accessed, possibly including authentication and authorisation.

A1. (Meta)data are retrievable by their identifier using a standardised communications protocol

A1.1 The protocol is open, free, and universally implementable

A1.2 The protocol allows for an authentication and authorisation procedure, where necessary

A2. Metadata are accessible, even when the data are no longer available

Reusable

The ultimate goal of FAIR is to optimise the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.

R1. (Meta)data are richly described with a plurality of accurate and relevant attributes

R1.1. (Meta)data are released with a clear and accessible data usage license

R1.2. (Meta)data are associated with detailed provenance

R1.3. (Meta)data meet domain-relevant community standards

Source: https://www.go-fair.org/fair-principles/
Image Source: https://kidsfirstdrc.org/news/fair-data/

Current Status & Path Forward

Initial Discussion (ESIP IQC/BSC EQC)/later joined by AU/NZ DQIG

Pre-ESIP Workshop Announced to Prospective Collaborators

Virtual Pre-ESIP Workshop (July 13, 2020)

Pre-ESIP Workshop Summary and Case Statement
(DOI: 10.31219/osf.io/75b92)

Working Group and Guidelines Development

Public Call-to-Action Statement
(DOI: 10.5334/dsj-2021-019)

Community Review of the Guidelines Document

Guidelines Document First Baseline
(DOI: 10.31219/osf.io/xsu4p)

Guidelines Document Maintenance and Update

09/19 02/20 07/20 08/20 09/20 12/20 04/21 10/21

Planning → **Developing** → **Community Review** → **First Baseline** →

- Discipline Diversity
- Publication: under-review
- Use Cases

SciDataCon 2021 – Virtual Meeting – Oct. 18-28, 2021

- **Session 285**: The State of Documenting and Reporting Data and Information Quality for Supporting Open Science
- **Organizers**: Hampapuram Ramapriyan, Robert Downs, Ge Peng, Yaxing Wei
- **~40 Attendees**: International and Multidisciplinary
- **6 Invited presentations:**
  - Ge Peng (University of Alabama in Huntsville, USA) - Developing Guidelines for FAIR Dataset Quality Information
  - Yaxing Wei (Oak Ridge National Laboratory, USA) *et al* - **NASA** Data Quality Working Group (DQWG) Recommendations
  - Peng Yue and Boyi Shangguan (Wuhan U., China) - Quality Considerations for **TrainingDML-AI**
  - Steve McEachern (Australian Data Archive, Australian National University) - (Meta)Data Quality in the **Social Sciences**
  - Peter Elias (University of Lagos, Nigeria) - Understanding the Priorities and Principles of **Citizen Science** Data Quality
  - Mark Allen (Centre de Données astronomiques de Strasbourg, France) – Data Quality in **Astronomy**: The role of infrastructures, standards and data stewards
- **Session notes and presentation slides** are available at https://drive.google.com/drive/folders/1icPvDyII0UyZcKCoMm1Zz2bY7MdNTdUf

RDA 18th Plenary: 3-11 November 2021, Virtual

- **BoF (3 Nov 2021)**: Representing and Communicating Data Quality Information
- **Organizers**: Ge Peng, Lesley Wyborn, Robert Downs, Hampapuram Ramapriyan, Ivana Ivanova, Carlo Lacagnina, and Mingfang Wu
- **Affiliated groups**: ESIP IQC, AU/NZ DQIG, BSC EQC team, and OGC DQ DWG
- **Statistics**: ~47 Attendees; International and Multidisciplinary
- **6 Invited presentations:**
  - Optimizing stewardship of **genomic and related health data** in the cloud (Vasiliki Rahimzadeh, Stanford Center for Biomedical Ethics, Stanford University, USA)
  - Data quality in **astronomy** - Building trust (Francoise Genova, Strasbourg astronomical data centre, France)
  - (Meta)Data Quality in the **Social Sciences** (Steven McEachern, Australian Data Archive, Australian National University, Australia)
  - **Earth Science** community guidelines to improve the representation and communication of dataset quality information (Robert Downs, Center for International Earth Science Information Network, Columbia University, USA)
  - Development of geospatial data quality use cases:
    - Ivana Ivánová, **OGC** Data Quality Domain Working Group, Curtin University, AUS;
    - Christin Henzen, **GeoKur project team**, Geoinformatics/Technische Universität Dresden, Germany.

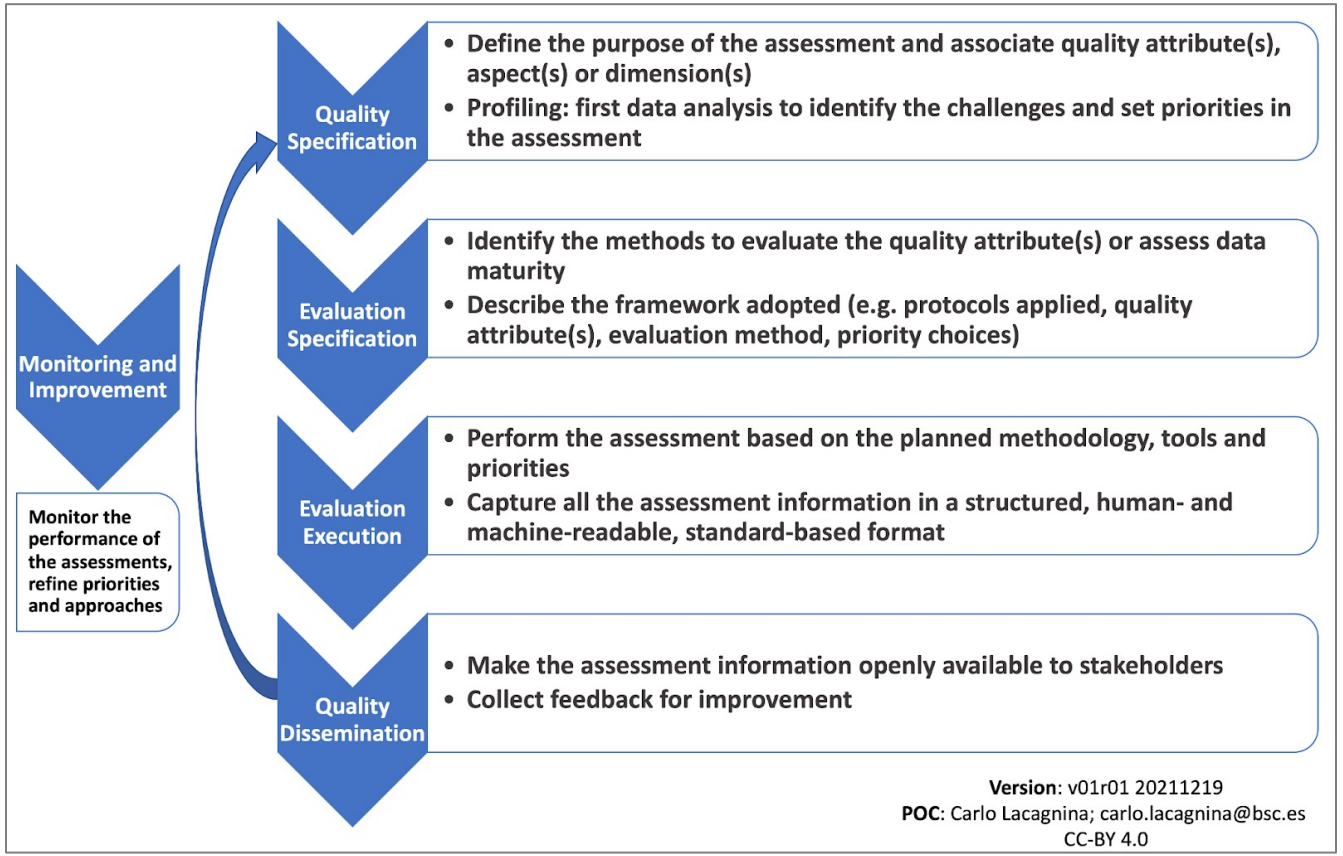
Session notes with links to presentations are available at: tinyurl.com/RDA18P-DQI

ESIP Winter Meeting 2022: Jan 20, 2022, Virtual

- **IQC Session**: Enhancing the Guidelines for Sharing and Reusing Dataset Information Quality
- **Organizers:** Robert Downs, Hampapuram Ramapriyan, Ge Peng, Yaxing Wei, and David F. Moroni
- **~40 Attendees**: International and Multidisciplinary
- **Presentations:**
  - Brief overview of ESIP Information Quality Cluster and motivation for developing the dataset quality guidelines. Yaxing Wei
  - Data Quality at Australia's **Integrated Marine Observing System** - Strengthening the foundations that underpin IMOS (IMOS). Dr. Natalia Atkins, University of Tasmania
  - Interdisciplinary Perspectives on Dataset Quality: Summary from SciDataCon. Hampapuram K. Ramapriyan
  - Interdisciplinary Perspectives on Dataset Quality: Summary from the RDA 18th Plenary RDA-18P. Ge Peng
  - Overview of the Dataset Quality Information Guidelines. Ge Peng
  - **Reviewing Elements** 3 and 4 of the Guidelines. Robert R. Downs
- **Session notes** with links to presentations are available at: https://docs.google.com/document/d/1mPoryW4MTsvWJIT-Hahw_gQGTBKwGzXfALbsvMYfg1I/edit

A Deeper Dive into the Guidelines

Basic Workflow for Curating and Disseminating DQI

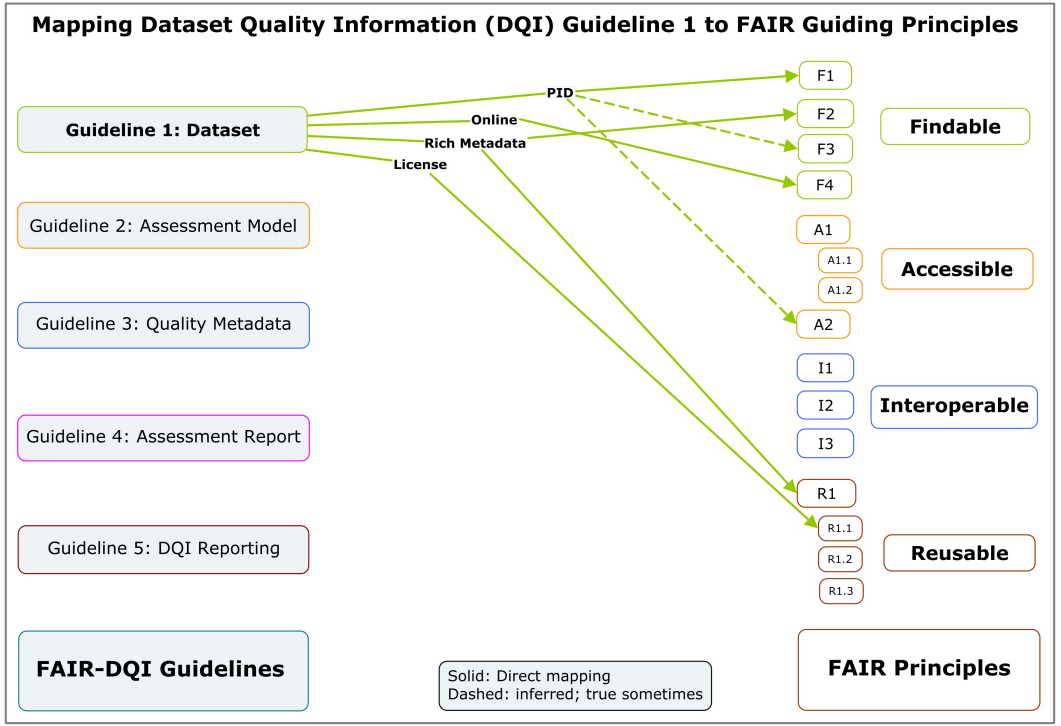


Source: Peng et al. 2021, OSF, https://doi.org/10.31219/osf.io/xsu4p

International Community Guidelines

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).



➢ Ensure the dataset is findable, accessible, and reusable

International Community Guidelines

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.

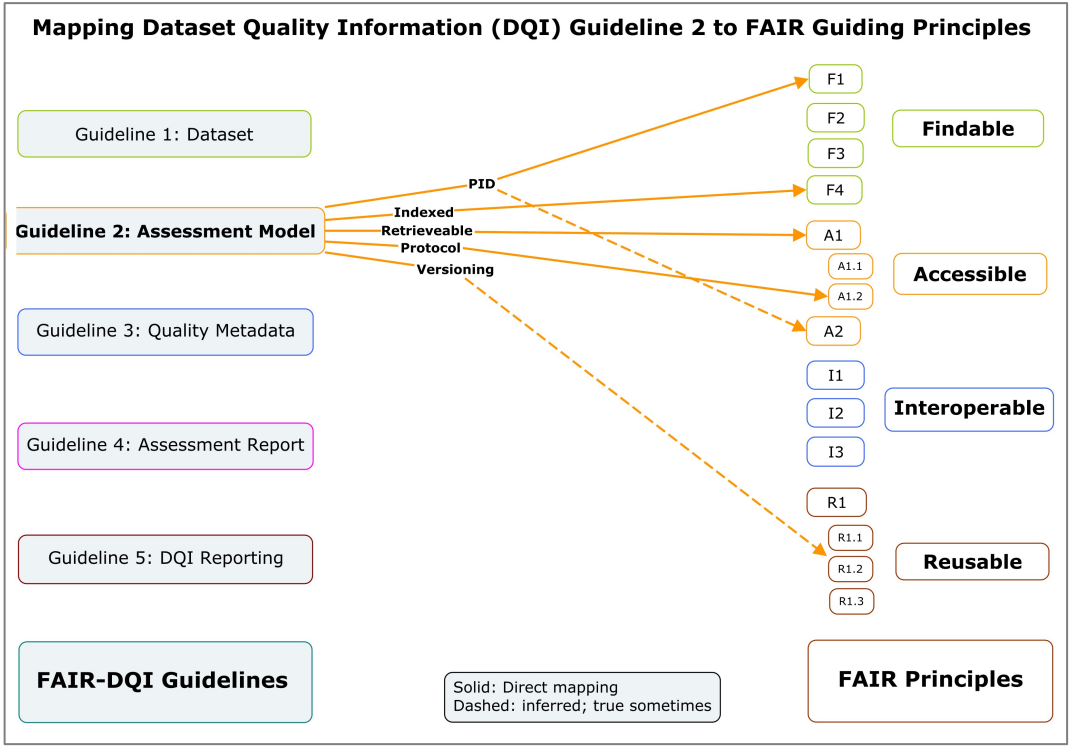
Examples of Describing Dataset → Citations

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

International Community Guidelines

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).



- Ensure the assessment model is findable and accessible

International Community Guidelines

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).

Example of Assessment Models

- NOAA Scientific Data Stewardship Maturity Matrix

SMM Document ID: NCDC-CICS-SMM_0001
Version: Rev. 1. 12/09/2014

NOAA/NSIDC Passive Microwave Sea Ice Concentration Climate Data Record
Version 2

Maturity Level as of 12/08/2016

Data Stewardship Maturity Scoreboard

Maturity Scale	Preservability	Accessibility	Usability	Production Sustainability	Data Quality Assurance	Data Quality Control/Monitoring	Data Quality Assessment	Transparency /Traceability	Data Integrity
Level 1 – Ad Hoc Not Managed	Any storage location Data only	Not publicly available Person-to-person	Extensive product-specific knowledge required No documentation online	Ad Hoc or Not applicable No obligation or deliverable requirement	Data quality assurance (DQA) procedure unknown or none	None or Sampling unknown or spotty Analysis unknown or random in time	Algorithm/method/model theoretical basis assessed (method and results online)	Limited product information available Person-to-person	Unknown or no data ingest integrity check
Level 2 - Minimal Managed Limited	Non-designated repository Redundancy Limited archiving metadata	Publicly available Direct file download (e.g., via anonymous FTP server) Collection/dataset level searchable	Non-standard data format Limited documentation (e.g., user's guide) online	Short-term Individual PI's commitment (grant obligations)	Ad Hoc and random DQA procedure not defined and documented	Sampling and analysis are regular in time and space Limited product-specific metrics defined & implemented	Level 1 + Research product assessed (method and results online)	Product information available in literature	Data ingest integrity verifiable (e.g., checksum technology)
Level 3 - Intermediate Managed Defined, Partially Implemented	Designated archive Redundancy Community-standard archiving metadata Conforming to limited archiving process standards	Level 2 + Non-standard data service Limited data server performance Granule/file level searchable Limited search metrics	Community Standard-based interoperable format & metadata Documentation (e.g., source code, product algorithm document, processing or/and data flow diagram) online	Medium-term Institutional commitment (contractual deliverables with specs and schedule defined)	DQA procedure defined and documented and partially implemented	Level 2 + Sampling and analysis are frequent and systematic but not automatic Community metrics defined and partially implemented Procedure documented and available online	Level 2 + Operational product assessed (method and results online)	Algorithm/method/model Theoretical Basis Document (ATBD) & source code online Dataset configuration managed (CM) Unique Object Identifier (OID) assigned (dataset, documentation, source code) Data citation tracked (e.g., utilizing Digital Object Identifier (DOI) system)	Level 2 + Data archive integrity verifiable
Level 4 - Advanced Managed Well-Defined, Fully Implemented	Level 3 + Conforming to community archiving standards	Level 3 + Community-standard data services Enhanced data server performance Conforming to community search metrics Dissemination report metrics defined and implemented internally	Level 3 + Basic capability (e.g., subsetting, aggregating) & data characterization (overall/global, e.g., climatology, error estimates) available online	Long-term Institutional commitment Product improvement process in place	DQA procedure well documented, fully implemented and available online with master reference data Limited data quality assurance metadata	Level 3 + Anomaly detection procedure well-documented and fully implemented using community metrics, automatic, tracked and reported Limited quality monitoring metadata	Level 3 + Quality metadata assessed (method and results online) Limited quality assessment metadata	Level 3 + Operational Algorithm Description (OAD) online, OID assigned, and under CM	Level 3 + Data access integrity verifiable Conforming to community data integrity technology standard
Level 5 - Optimal Level 4 + Measured , Controlled , Audit	Level 4 + Archiving process performance controlled, measured, and audited Future archiving standard changes planned	Level 4 + Dissemination reports available online Future technology and standard changes planned	Level 4 + Enhanced online capability (e.g., visualization, multiple data formats) Community metrics of data characterization (regional/cell) online External ranking	Level 4 + National or international commitment Changes for technology planned	Level 4 + DQA procedure monitored and reported Conforming to community quality metadata & standards External review	Level 4 + Cross-validation of temporal & spatial characteristics Physical consistency check Conforming to community quality metadata & standards Dynamic providers/users feedback in place	Level 4 + Assessment performed on a recurring basis Conforming to community quality metadata & standards External ranking	Level 4 + System information online Complete data provenance available online	Level 4 + Data authenticity verifiable (e.g., data signature technology) Performance of data integrity check monitored and reported

Dataset Information: https://dx.doi.org/10.7265/N55M63M1
Dataset POC: NOAA Climate Data Record Program Office, sea_ice_concentration_contacts@noaa.gov

SMM POC: Ge Peng; Ge.Peng@noaa.gov
SMM Assessment POC: Paul Lemieux, III; Paul.Lemieux@noaa.gov

Peng, G. 2014, DSMM Template. CC-BY-4.0. *Figshare*. https://doi.org/10.6084/m9.figshare.1211954

International Community Guidelines

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

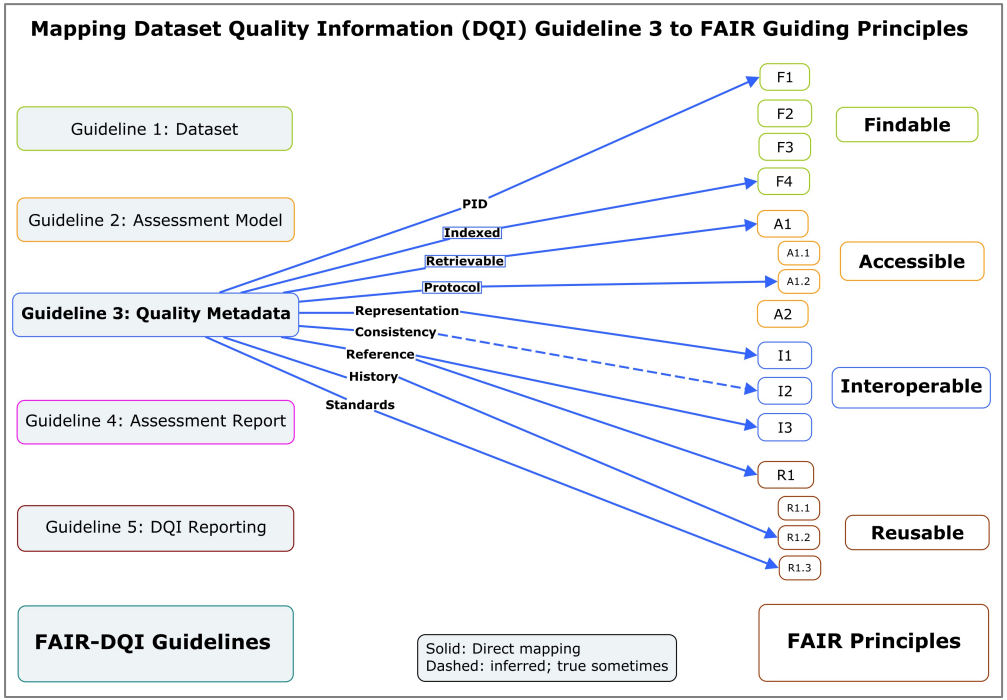
3.1. is semantically and structurally consistent and follows community standards - preferable to be compliant with national or international metadata standards that satisfy conditions 2.1-2.3,

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,

3.5. includes versioning and the history of the assessments.



- Ensure the quality information is findable, accessible, interoperable and reusable for *machine end-users*

Examples of Dataset Quality Metadata – 1 of 2

Measure Name	Data Stewardship Maturity Assessment
Measure ID	MM-Stew
Measure Description	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 Description	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.
Procedure Reference	Peng, Ge. The Scientific Data Stewardship Maturity Assessment Model Template. 2015-06-23. doi:10.6084/m9.figshare.1211954
Date of Measurement	2016-12-08
Quantitative Results	
Preservability	advanced
Accessibility	minimum
Usability	advanced
Production Sustainability	advanced
Data Quality Assurance	advanced
Data Quality Control/Monitoring	minimal
Data Quality Assessment	intermediate
Transparency/Traceability	intermediate
Data Integrity	advanced
Conformance Results Explanation	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.
Reference	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

ISO Quality Metadata

Source: Peng et al. 2019. DSJ. https://doi.org/10.5334/dsj-2019-041

Examples of Dataset Quality Metadata – 2 of 2

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

AtMoDat Maturity Indicator

Source: Heydebreck et al. 2020, Data Maturity Indicator Concept (v4.5)
https://doi.org/10.35095/WDCC/Data_MIC_v4.5

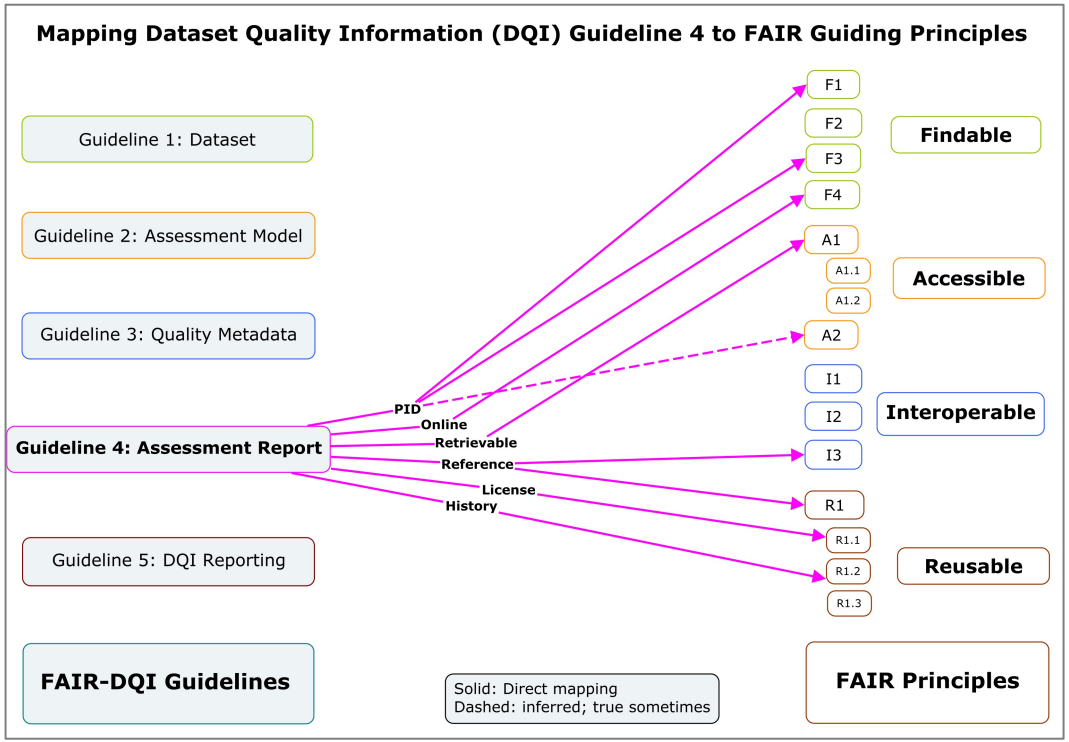
International Community Guidelines

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 conditions 2.1-2.3,

4.2. is published with an explicit open license and history of the report, satisfying the conditions 2.1-2.3, and

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



- Ensure the quality information is findable, accessible and reusable for *human end-users*

International Community Guidelines

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

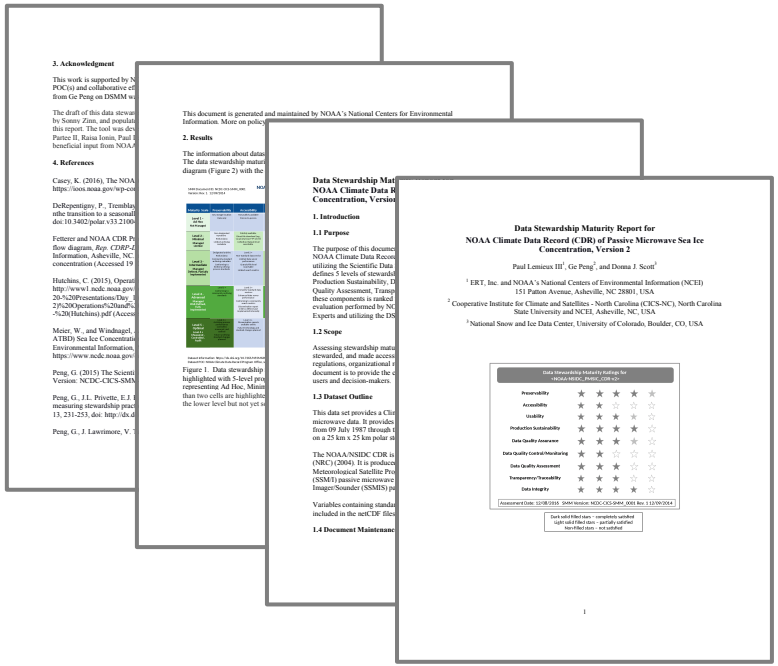
4.1. preferably follows a template that is published and satisfies conditions 2.1-2.3,

4.2. is published with an explicit open license and history of the report, satisfying the conditions 2.1-2.3, and

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

Example of Dataset Quality Report

- Data Stewardship Maturity Report



Lemieux et al. 2019. Figshare. CC-BY-4.0.
https://doi.org/10.6084/m9.figshare.5279932

International Community Guidelines

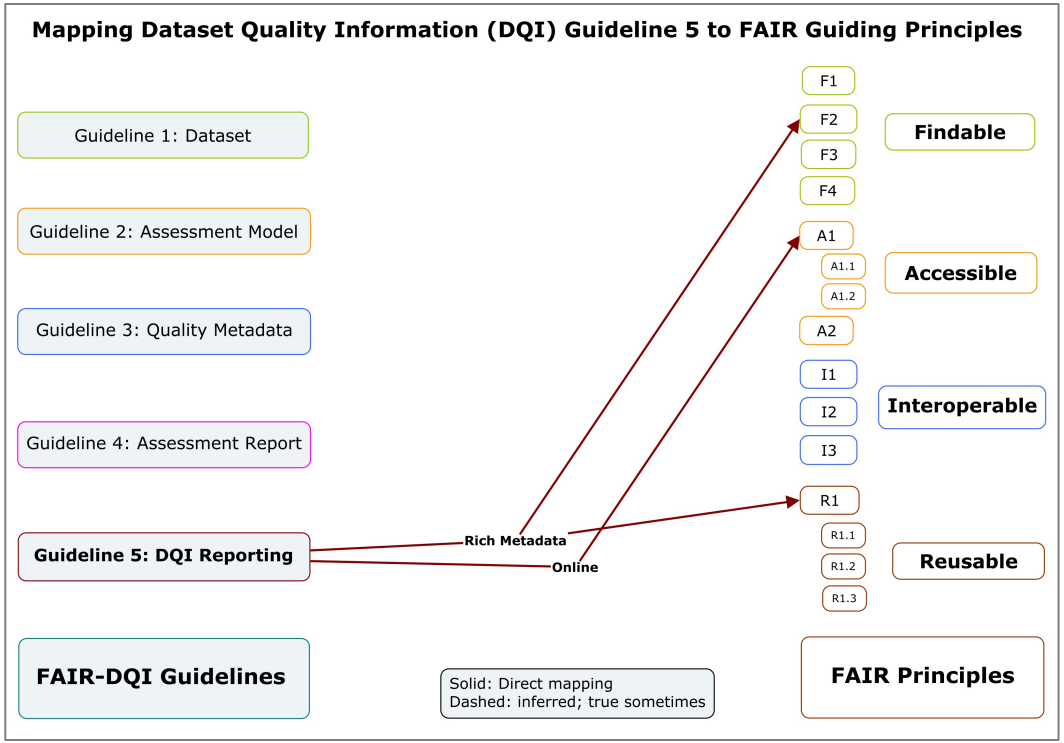
Guideline 5: Report/disseminate the dataset quality information in an organized way via a web interface with a comprehensive description of:

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,

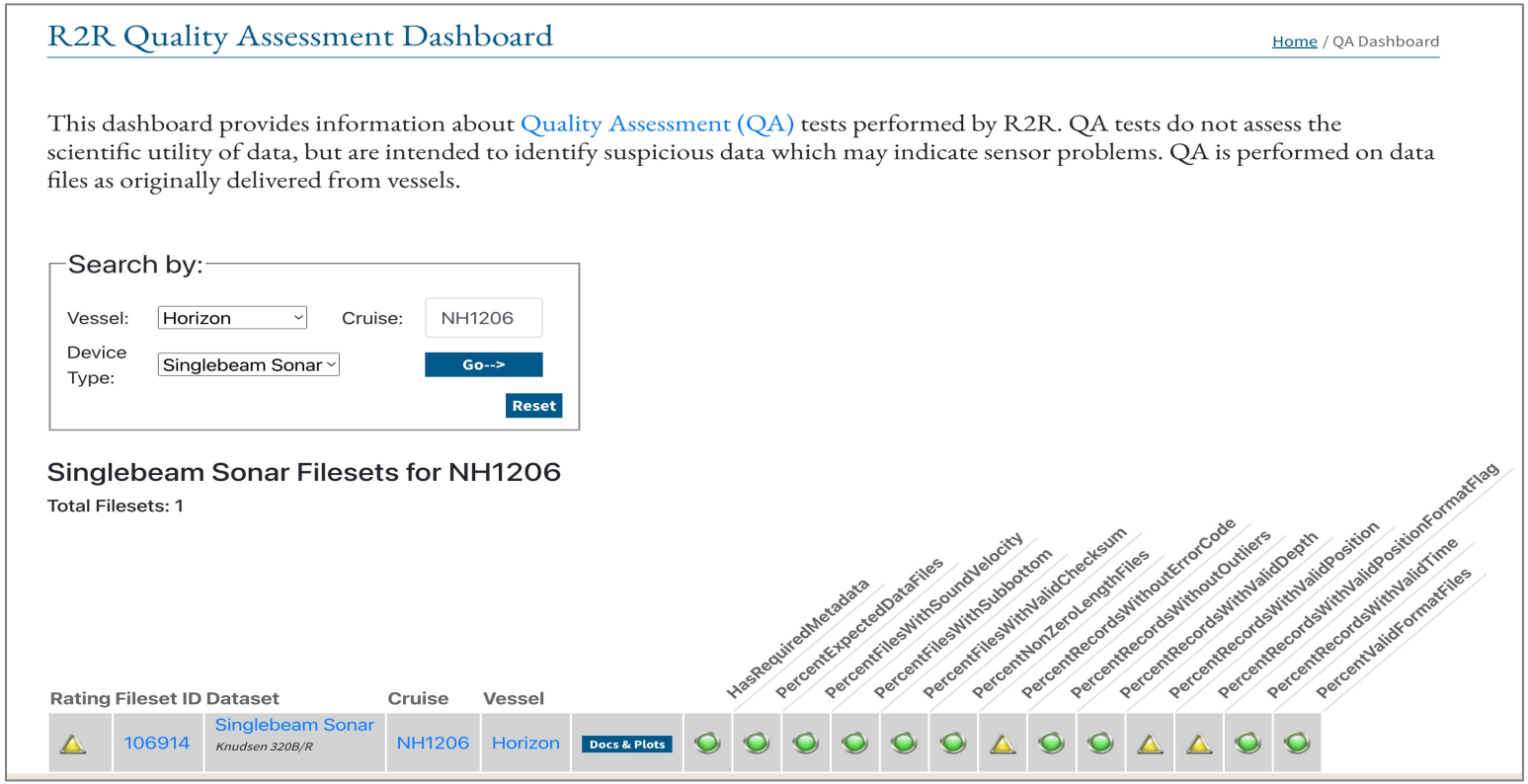
5.4. how to understand and use the information.



- Ensure the information is online, findable, understandable, and readily usable

Examples of Disseminating Quality Information 1 of 2

R2R Quality Assessment Dashboard



Source:https://www.rvdata.us/qa_info?vessel=string:Horizon&cruise=NH1206&device=string:Singlebeam%20Sonar

Examples of Disseminating Quality Information 2 of 2

C3S Climate Data Store

ERA5 monthly averaged data on pressure levels from 1979 to present

NOTIFICATION 2021-12-03: please be aware that for September 2021 the final release of ERA5 differs from the timely updates (ERA5T), and it will be for the months October-December 2021 as well. See the online ERA5 documentation for more information.

Overview | Download data | Quality assessment | Documentation

This is a new feature, work in progress. Should any inconsistency be found, ple

The CDS datasets are assessed by the Evaluation and Quality Control (EQC) funct
encompasses a framework of processes aimed to assure technical and scientific
through the CDS. During the EQC process, the documentation provided with the
reliability.

Variable:

Relative humidity

▾ Variable: Relative humidity — Last updated on 14/05/2021

INTRODUCTION	USER DOCUMENTATION	ACCESS	INDEPENDENT ASSESSMENT
Dataset overview	User guide	Toolbox compatibility	Data check
Temporal and spatial coverage and resolution	Scientific methodology	Archive	Expert evaluation
Providers	Uncertainty quantification		Dataset maturity
Dataset version	Validation		Key strengths and limitations
Data update	Inter-comparison		

Entries with the mark | display content that is specific for the variable selected

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

THANK YOU