WGISS-54

FAIR Dataset Quality Information Guidelines

Dr Ivana Ivánová, Curtin University
Agenda ID: 2022.10.03_12.15

WGISS-54
Tokyo, Japan (JAXA)
3-7 October 2022
Executive Summary

- FAIR DQI community guidelines provide specific advice on ensuring quality metadata compliant with the FAIR principles for the dataset
- FAIR DQI community guidelines are a living document developed by international community for international community
- Use-Cases on challenges with quality information are wanted!

- FAIR DQI guidelines support Priority 3: Support to CEOS Cal/Val Initiatives to increase CEOS Agency Cal/Val Collaboration
Why FAIR Quality Information?

❖ Increasingly the reuse of a dataset, particularly where multiple datasets are being merged, requires knowledge of the “quality” of the datasets to be merged.

❖ Particularly where datasets are repurposed for use cases beyond what the original creator intended: “quality” information becomes critical.

❖ With the rise of Artificial Intelligence (AI) and Machine Learning (ML), a new interpretation of FAIR is that it stands for “Fully AI Ready”: knowing the “quality” of data to be used is essential to avoid erroneous conclusions.

---

**Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud**

- **Article type:** Research Article
- **Authors:** Mos, Barendse, & Willem, Luc\n
**Affiliations:**[a] Leiden University, [b] Dutch Techcentre for Life Amsterdam, The Netherlands, [c] Australia | [d] Independent Laboratory | [e] Institute for Data Science, Maastricht, Amsterdam, The Netherlands, Madrid, Spain

**Correspondence:**[†] Correspondence: A. 9600, 2300 RC Leiden, The Netherlands

**Keywords:** FAIR Data, Open Science

**DOI:** 10.3233/IOS-170824

**Journal:** Information Services

**Published:** 7 March 2017

**Abstract**

The FAIR Data Principles provide a set of guiding principles to make data Findable, Accessible, Interoperable, and Reusable. However, many researchers have found that the principles are not understood in the way intended. This paper presents a set of recommendations for how these principles should be implemented in practice.

---

**eResearch’21 – BoF on FAIR Data Quality Information**
What actually is the ‘quality’ of a dataset? How does the user know which dataset to trust?

❖ Few common quotes:
   ▪ “We can’t use that dataset because it is of poor quality”;
   ▪ “Don’t trust data from sector, organisation or a person: it does not meet OUR quality requirements”
   ▪ “Don’t trust repository XXXX: their datasets are full of errors and of low quality”

❖ But when pressed, very few could provide concrete examples of:
   ▪ Exactly what and where the supposed errors were in the dataset;
   ▪ What they were benchmarking the supposedly “poor” quality dataset against
   ▪ None could provide a “community-agreed” reference/best practice document that specified what their expectations on quality were.

❖ “community-agreed” guidelines on quality, preferably at an international level are urgently needed
An International Effort Came Together...

- Co-organized by:
  - ESIP Information Quality Cluster (IQC),
  - Barcelona Supercomputer Centre Evaluation and Quality Control Team (EQC),
  - ARDC-supported AU/NZ Data Quality Interest Group (DQIG)

- 22 International Interdisciplinary Domain Experts:
  - Data producers (in situ, satellite, model),
  - Stewards (data/science/technology),
  - Services providers (data/information/infrastructure),
  - Data publishers and users

- from 7 countries (USA, Spain, AU, NZ, Germany, UK, France),
  - with 22+ affiliations (government, academic, private sectors):
    - Data, science, and service centres, institutional repositories
  - with expert knowledge from data acquisition or production, data and information management, data publishing, services, and applications.
Timelines and current status

- Initial Discussion (ESIP IQC/BSC EQC)
- Pre-ESIP Workshop Announced to Prospective Collaborators
  - Virtual Pre-ESIP Workshop (July 13, 2020)
  - Pre-ESIP Workshop Summary and Case Statement
    (DOI: 10.31219/osf.io/75b92)
  - Working Group and Guidelines Development
    - Public Call-to-Action Statement
      (DOI: 10.5334/dsj-2021-019)
    - Community Review of the Guidelines Document
  - Guidelines Document First Baseline
    (DOI: 10.31219/osf.io/xsu4p)
  - Guidelines Document Maintenance and Update

Timeline:
- 09/19
- 02/20
- 07/20
- 08/20
- 09/20
- 12/20
- 04/21
- 10/21
Four key outputs so far

August 2020 - https://osf.io/75b92/

May 2021 - http://doi.org/10.5334/dsj-2021-019

October 2021 - https://osf.io/xsu4p


Call to Action for Global Access to and Harmonization of Quality Information of Individual Earth Data Sets

Authors: De Smedt, Holger; Dowdeswell, John; Frewer, Russell; Larocca, Giuseppe; Meng, Qiaoli; Nandan, Anurag; Saluzzi, Andrea; Smith, Alexander; Tobin-Fletcher, Ian; Tse, Wai; Williams, Michael

Abstract

Laying the Groundwork for Developing International Community Guidelines to Effectively Share and Reuse Digital Quality Information: Case Studies, Workshop Summary Report, and Path Forward

Authors: Chen, Xing; Deng, Yuan; Li, Hang; Lu, Dehui; Pu, Yu; Wang, Lei; Yu, Lichun; Zhang, Fei; Zhang, Peng; Zhao, Wei

Research Papers

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

Authors: De Smedt, Holger; Dowdeswell, John; Frewer, Russell; Larocca, Giuseppe; Meng, Qiaoli; Nandan, Anurag; Saluzzi, Andrea; Smith, Alexander; Tobin-Fletcher, Ian; Tse, Wai; Williams, Michael

Abstract

Under the auspices of the Earth Science Information Partnership (ESIP) a working group on quality information in Earth science (ESIP QI) was established in 2015. The primary objective of the working group was to develop international community guidelines with practical recommendations on sharing and using quality information of the datasets. The ESIP QI group developed a set of principles, core elements, and recommended practices to improve the quality of information in the datasets. The working group reached out to a wide audience to solicit input on the quality of information in the datasets. The working group is currently in the process of finalizing the guidelines.
Guidelines development principles

❖ Adapting the FAIR guiding principles (Wilkinson et al. 2016);
❖ Taking a whole dataset-lifecycle approach;
❖ Being quality-attribute and assessment-type agnostic;
❖ Common terminology is essential for enabling interoperability;
❖ Developing for the community by the community:
  ▪ Through an iterative process, with continuous engagement with all stakeholders,
  ▪ Leveraging the experiences and expertise of a team of interdisciplinary domain experts and community best practices and standards.
Framework defined by 4 dimensions

Based on: https://doi.org/10.1045/july2017-ramapriyan
Basic workflow for curating and reporting DQI

- **Quality Specification**: Define and describe the scope of the assessment and associate quality attribute(s) or dimension(s)
- **Evaluation Specification**: Identify and describe the assessment method and framework
- **Evaluation Execution**: Perform the assessment and capture the results in a structured, human- and machine-readable, standard-based format
- **Quality Dissemination**: Make the assessment results readily available and usable to stakeholders and collect feedback for improvement

Guideline 1: Describing Dataset (e.g. version, producer)
  ▪ Ensure the dataset is findable and accessible.

Guideline 2: Utilizing a quality assessment model
  ▪ Ensure the assessment model is findable and accessible.

Guideline 3: Capturing the assessment method and results
  ▪ Ensure the quality information is interoperable and reusable (machine end-users).

Guideline 4: Describing the assessment method, workflow and results
  ▪ Ensure the quality information is findable, accessible, citable and reusable (human end-users)

Guideline 5: Reporting the dataset quality information
  ▪ Ensure the information is FAIR
Mapping Dataset Quality Information (DQI) Guidelines to FAIR Guiding Principles

Guideline 1: Dataset
- Online
- PID
- Rich Metadata
- Licence

Guideline 2: Assessment Model
- Indexed
- Retrieveable
- Protocol
- Versioning
- PID

Guideline 3: Quality Metadata
- Consistency
- Reference
- History
- Standards
- PID
- Online

Guideline 4: Assessment Report
- Referable
- Reference
- License
- PID

Guideline 5: DQI Reporting
- Rich Metadata
- Online

FAIR-DQI Guidelines
- Solid: Direct mapping
- Dashed: inferred; true sometimes

FAIR Principles
- Findable
- Accessible
- Interoperable
- Reusable

At the moment: we are collecting use cases to:

- Ensure that the guidelines are in line with the user communities and their applications;
- Justify the need for best practices in describing quality information to ensure and proper use data;
- Collect examples from multiple application domains on the use of FAIR quality information;
- Provide the community with implementation examples of the guidelines;
- Develop the guidelines for the community by the community;

Please contribute [here](#).
FAIR DQI guidelines: path forward

❖ Continue promotion through regular presence at: ESIP, OGC, RDA, SciDataCon, eResearch Australasia…

❖ FAIR DQI guidelines is a living document expected to evolve over time based on user feedback and emerging community best practice

❖ FAIR DQI guidelines are not only for Earth Science datasets – we are expanding the discipline diversity
Thank you!

ivana.ivanova@curtin.edu.au