A Case of Developing Community Terminology and Concepts for Reporting Dataset Quality Information

Ge Peng, PhD
Sr. Research Scientist, Earth System Science Center/NASA MSFC IMPACT
The University of Alabama in Huntsville, Huntsville, AL, USA
Co-Lead, International FAIR DQI Community Guidelines Working Group
Co-Chair, Information Quality Cluster of ESIP
Co-Chair, Expert Team on Information Management of WMO

gpeng93@gmail.com

Achieving a shared understanding of concepts Workshop, 13 April 2021, Virtual
What Is Dataset Quality Information (DQI)?

It includes information about quality or state of data, metadata, and documentation through the entire lifecycle of a dataset.

What we are trying to do:
- develop community guidelines to help stakeholders to curate and make dataset quality information findable, accessible, interoperable, and reusable (FAIR; Wilkinson et al. 2016; Scientific Data)

Need to define key terms:
- Data; Dataset
- Dataset lifecycle stages
- Quality attribute(s), perspectives, or dimensions
- Quality assessment types or models
- Metadata; Quality metadata or reports
Who We Are – Diverse Expertise/Experiences

- 22 International Interdisciplinary Domain Experts
- from 6 countries (USA, Spain, Australia, Germany, UK, France),
- with 22+ affiliations (government, academic, private sectors):
  - Data, science and service centers, institutional repositories, companies.
- with expert knowledge on data acquisition or production, data and information management, data publishing, services, and applications.

Terms may mean different things to different people!

- Accuracy – Quality attribute or assessment type?
  - Data value? Variable naming convention? Unit?
  - Metadata content? Data file naming convention?
  - Data services?
- Completeness, consistency, etc.
Definitions of Dataset

- **Dataset** is a type of managed data aggregation from multiple data elements which are considered as an aggregated unit for processing purposes.
  - Scholarly definition

- **Dataset** refers to an identifiable collection of data, and it can be processed, curated or published by a single agent.
  - Data catalogue

- **Dataset** is a digital rendition of factual materials, or a product of a given version of an algorithm/model, having the same geophysical variable(s) and product specification(s).
  - Earth science

- *Many other definitions* ([Renear et al. 2010](#))
Current State of Shared Terminology and Concepts in Reporting Dataset Quality Information

- **A term may be defined differently**
  - one organization from another,
  - one entity (division/project) from another within the same organization

- **Different terms may be defined for same thing**

- **Terms or concepts are needed but not yet defined**

- **Big challenges but also an opportunity to harmonize shared terminology and concepts for the global community**
  - **Foundation** for improved interoperability across domains, systems and tools!
  - **Minimizing gaps** between scholarly and domain/discipline-oriented definitions.
Typical Workflow For Curating and Reporting DQI

- **Define** and describe the purpose of the assessment and associate *quality attribute(s)* or *dimension(s)*
- **Identify** and describe the *assessment method* and framework
- Perform the assessment and **capture the results** in a structured, standard-based format
- **Make** the assessment *results readily available* and usable to stakeholders and collect feedback for improvement

(Based on: International FAIR-DQI Community Guidelines Working Group 2021, *in prep.*)
A Small-Scale Example - Defining Common Terminology

Dataset Lifecycle Stages and Dimensions

Digital Dataset Lifecycle

- Define
- Develop
- Validate
- Produce
- Evaluate
- Release
- Preserve
- Document
- Access
- Use
- Service
- Reuse/Improve

Dataset Lifecycle Stages

- Define
- Develop
- Validate
- Produce
- Evaluate
- Release
- Preserve
- Document
- Access
- Use
- Service
- Reuse/Improve

Dataset Quality Dimensions

- Define
- Develop
- Validate
- Produce
- Evaluate
- Release
- Preserve
- Document
- Access
- Use
- Service
- Reuse/Improve

(From: International FAIR-DQI Community Guidelines Working Group 2021, in prep.)
Dataset Lifecycle Stages
- Define
- Develop
- Produce
- Evaluate

Dataset Quality Dimensions
- Service
- Science
- Product

Quality Information Report Types
- Stewardship
- Access
- Use

Metadata
- Metadata-Tags
- Quality Dimensions
- Metadata - Intrinsic Identification
- Metadata - Provenance Access
- Metadata - Content Access

Digital Dataset Lifecycle

LEGEND
- Yellow: Dataset Lifecycle Stages
- Purple: Quality Dimensions
- Orange: Document Types
- Beige: Metadata Tags
- Pink: Metadata Entities

Version: v02r00 20210321
POC: gpeng93@gmail.com
CC-BY 4.0

(From: International FAIR-DQI Community Guidelines Working Group 2021, in prep.)
Achieving a shared understanding of concepts Workshop, 13 April 2021, Virtual

Quality Metadata Reporting Framework

Dataset Quality Dimension
- Dimension contains one or many quality attributes

DQ Attribute
- Is expressed by
- Is described by

DQ Measure Reference
- Measure Name
- Measure ID
- Measure Description

DQ Evaluation Method
- Evaluation Method Description
- Evaluation Procedure Reference

DQ Result
- Evaluation Quantitative Results
- Conformance Result Explanation

DQ Metadata
- Evaluation Date
- Evaluation Version
- Evaluation POC
- Evaluation Citation

(Source: International FAIR-DQI Community Guidelines Working Group 2021, in prep.)
Takeaway Messages

- Dataset quality information is **more than** just information about data quality.

- Common terminology and shared concepts are paramount for **improved** data and information sharing.
  - Should they be more scholarly or discipline-oriented?
  - How to best harmonize them across-disciplines?

- **A holistic approach** should be adopted.
  - Through the entire lifecycle of data and information,
  - Collaboration across domains (science, data management, technology, services, etc.)
Acknowledgement

International FAIR-DQI Community Guidelines Working Group

THANK YOU!

Contact Me

Email: gpeng93@gmail.com
OCRID: 0000-0002-1986-9115
LinkedIn: ge-peng-37543230
Twitter: @DrPengAtAVL

Additional Resources

- 2020 Workshop Summary Report: https://doi.org/10.31219/osf.io/75b92
- Preprint of Call-to-Action Statement Paper: https://doi.org/10.31219/osf.io/nwe5p
- ESIP wiki (to be updated): https://wiki.esipfed.org/FAIR_Dataset_Quality_Information
Backup Slide
Achieving a shared understanding of concepts Workshop, 13 April 2021, Virtual

(From: International FAIR-DQI Community Guidelines Working Group 2021, in prep.)