Community Guidelines on Documenting and Reporting Dataset Quality

Ge Peng, Sr. Principal Research Scientist

Earth System Science Center/MSFC IMPACT
The University of Alabama in Huntsville

ESIP Information Quality Cluster, co-chair

NASA GES DISC Seminar, 19 April 2022

DOI: 10.5281/zenodo.6472568

Done By

International FAIR-DQI Community Guidelines Working Group

Core Writing Team, Ge Peng, Carlo Lacagnina, Ivana Ivánová, Robert R. Downs, Hampapuram K. Ramapriyan, Anette Ganske, Lesley Wyborn, Dave Jones, Lucy Bastin, Chung-lin Shie, David F. Moroni

Presentation Outline

- Dataset Quality (Information)
- FAIR Principles in a nutshell
- Impact and costs of not sharing
- Needs and challenges
- Community guidelines development
- High-level view of the guidelines

Service Quality Attributes

Data Quality Attributes

Metadata Quality Attributes

Software Quality Attributes

Workflow/Procedure Quality Attributes

Service Quality Attributes

Data Quality Attributes

- Accuracy
- Completeness
- Consistency
- Timeliness
- Integrity
- Validity

Metadata Quality Attributes

Software Quality Attributes

Workflow/Procedure Quality Attributes

Service Quality Attributes

Data Quality Attributes

- Accuracy
- Completeness
- Consistency
- Timeliness
- Integrity
- Validity

Metadata Quality Attributes

- Accuracy
- Completeness
- Consistency
- Integrity
- Conformance

Software Quality Attributes

Workflow/Procedure Quality Attributes

Service Quality Attributes

Data Quality Attributes

- Accuracy
- **Completeness**
- Consistency
- Timeliness
- Integrity
- Validity

Metadata Quality Attributes

- Accuracy
- **Completeness**
- Consistency
- Integrity
- Conformance

Software Quality Attributes

Workflow/Procedure Quality Attributes

Dataset Quality Information

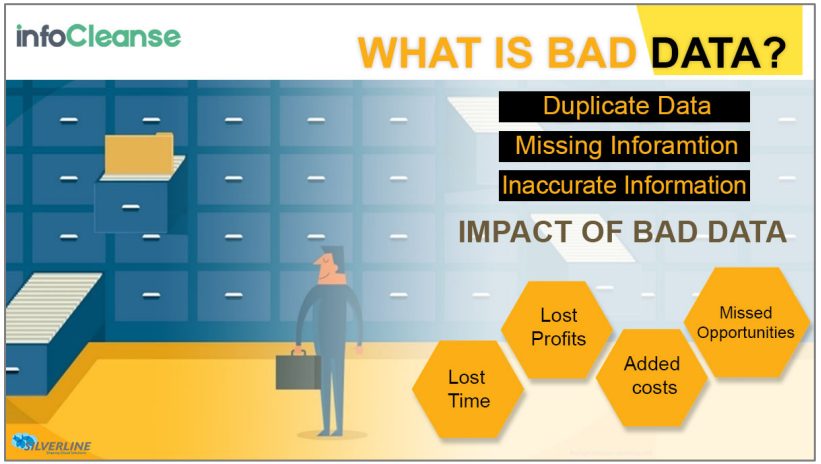
Information about quality or the state of
data, metadata, software, workflow, procedure, etc.
through the entire lifecycle of a dataset

- **Not just information about data quality**
- **Dataset lifecycle approach**

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

Why Should We Care?

Impact of 'Bad' Data Quality

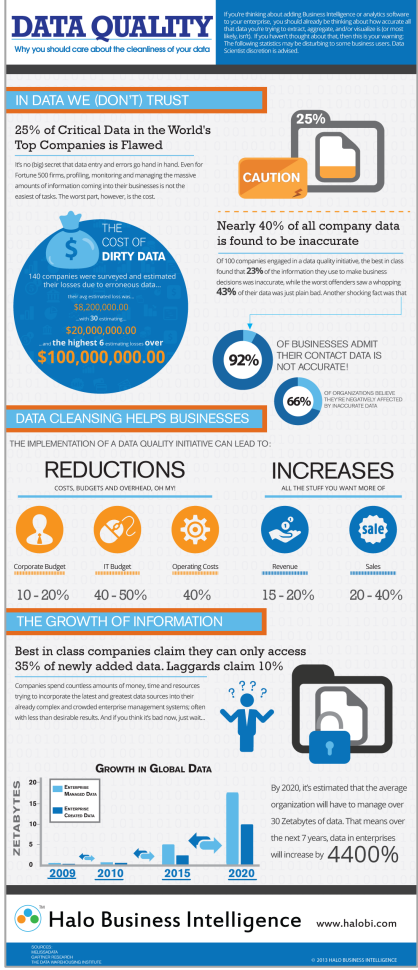


Source: https://www.infocleanse.com/impact-of-poor-data-quality-in-business

Source: https://basecapanalytics.com/the-impact-of-bad-data/

For Us: Data Producers, Data/Service Providers, Stewards, Data Centers

- Negative in reputation/Trust
- Bad user experiences
- Lost of productivity



Source: https://www.pinterest.com/ileana9cruz/data-quality

Why Should We Care?

Costs of Not Sharing Data

(Collecting Datasets: 10-19% [CrowdFlower 2016])

For EU: Minimum of €10.2bn per year

(**Source:** European Commission and PwC EU Services 2018)

Costs of Not Sharing Information about Data Quality

(Cleaning and Organizing Data: 60-70% [CrowdFlower 2016])

Globally

- Productivity lost - redundancy in assessing data quality,
- Million $ decisions - disaster responses.

We Need (Consistently Curated) **Quality Information**

- **Decision-making**
  - Data use: Informing the reliability and usability of the dataset;
  - Data trust: Establishing the trust between data providers and consumers, policy-makers.
- **Compliance reporting and open science support**
  - Consistently curated;
  - Readily available and understood by humans and machines.
- **Support data and information sharing and reuse**
  - Support new technologies: Interoperable dataset quality information for utilizing Cloud and Machine Learning technologies;
  - Reduce access barrier: global access and harmonization of quality information.

Quality Is Complicated!

- Multi-dimensionality
- Cross-domain knowledge integration
- Fitness for purpose

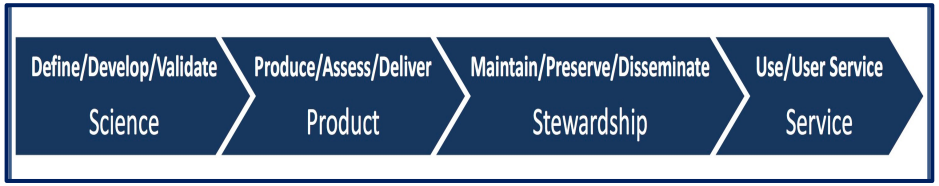
Quality Is Complicated!

- Multi-dimensionality
- Cross-domain knowledge integration
- Fitness for purpose

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

(Wang and Strong 1996, *J. Management Info. Sys.)*

Dataset lifecycle Stages and Quality Aspects



(Ramapriyan et al. 2017, *D.-Lib Magazine*)

Quality Is Complicated!

- Multi-dimensionality
- Cross-domain knowledge integration
- Fitness for purpose

Scientists

Metadata Specialists

Systems/Tools

Quality Is Complicated!

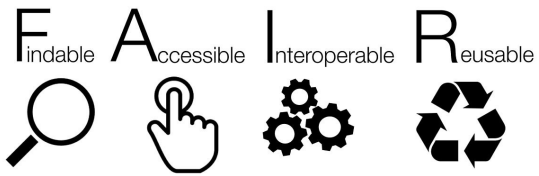
- Multi-dimensionality
- Cross-domain knowledge integration
- Fitness for purpose
  - Data user paradigm shift

➢ Making consistently documented quality information readily available

How To Improve Sharing?

Adopting FAIR Guiding Principles

(Wilkinson et al. 2016)



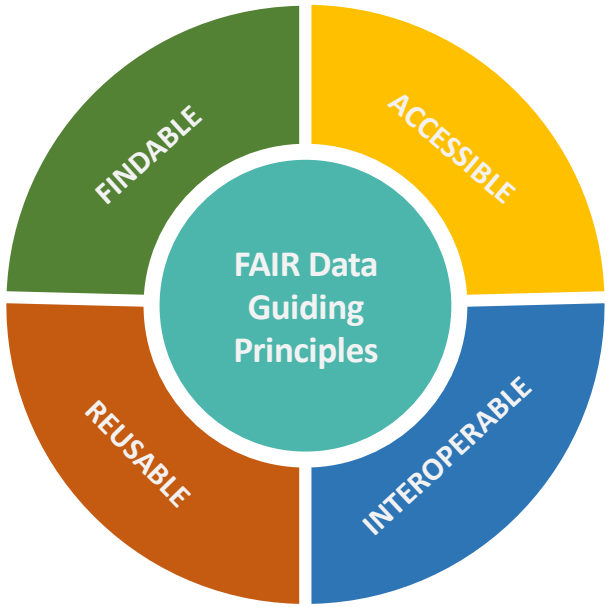
FAIR Data Guiding Principles

(In a Nutshell)

- **Uniquely Identifiable and Discoverable**

Findable Principle

F1 -> PID
F2 -> Rich Metadata
F3 -> Cross-Reference PIDs
F4 -> Catalog



- Easily Obtainable

- Readily (Re)usable

- Readily Integrated

FAIR Data Guiding Principles

(In a Nutshell)

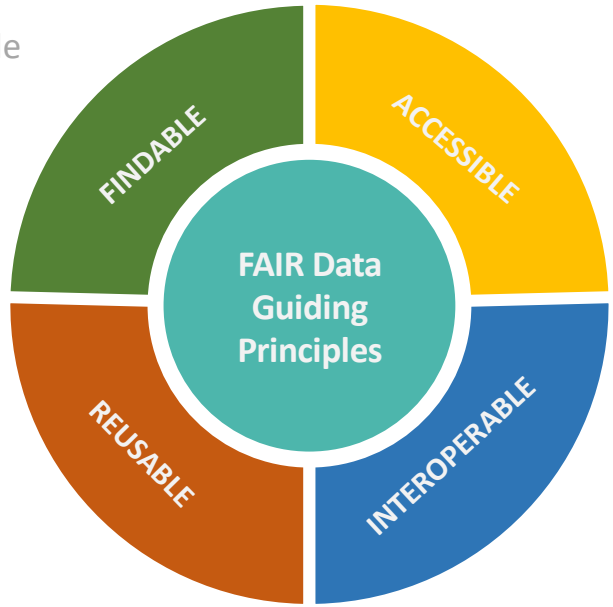
- Uniquely Identifiable and Discoverable

➢ **Easily Obtainable**

Accessible Principle

A1 -> Securely Retrievable

A2 -> Permanent Metadata



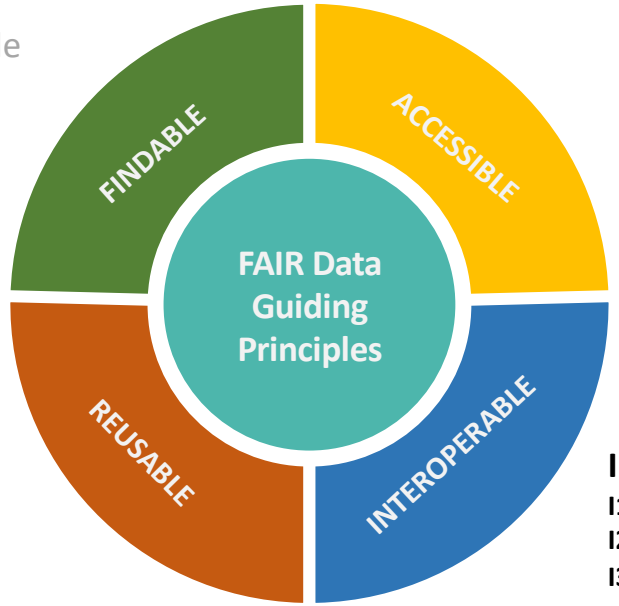
- Readily (Re)usable

- Readily Integrated

FAIR Data Guiding Principles

(In a Nutshell)

- Uniquely Identifiable and Discoverable
- Easily Obtainable



- Readily (Re)usable

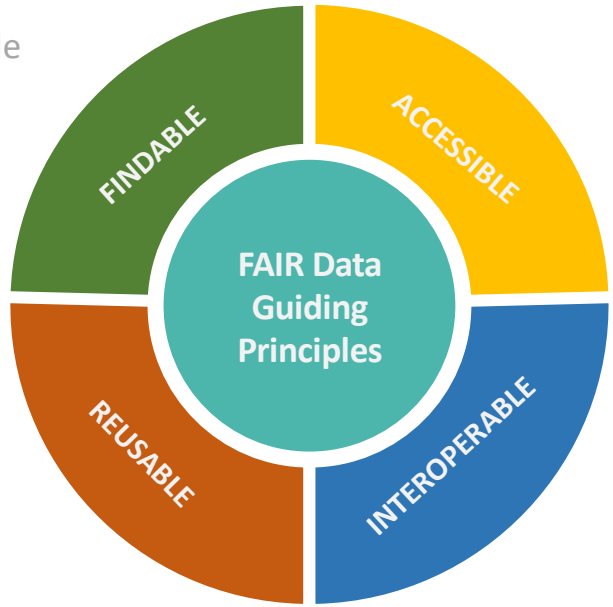
➢ **Readily integrable**

Interoperable Principle

I1 -> Knowledge Representation
I2 -> Vocabularies
I3 -> Cross-Referenced Resources

FAIR Data Guiding Principles

(In a Nutshell)



- Uniquely Identifiable and Discoverable

- Easily Obtainable

➢ **Readily (re)usable**

Reusable Principle

R1 -> Rich Meta(data)

R1.1 -> Usage License

R1.2 -> Provenance

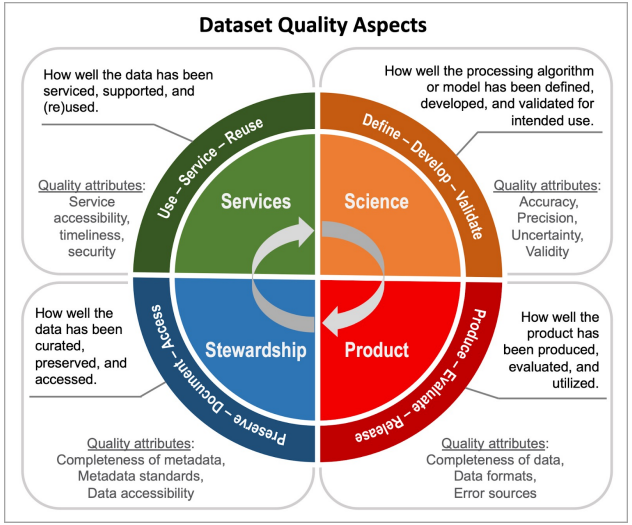
R1.3 -> Standards

- Readily Integrated

How To Consistently Document?

International FAIR-DQI Community Guidelines

(Peng et al. 2022. DOI: 10.5334/dsj-2022-008)



Guidelines Development

Co-organized by

- ESIP Information Quality Cluster (IQC);
- BSC Evaluation and Quality Control (EQC) Team;
- AU/NZ Data Quality Interest Group (DQIG).

Initial Discussion (ESIP IQC/BSC EQC)

Virtual Pre-ESIP Workshop (July 13, 2020) & ESIP SM20 Report-out Session

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

Working Group and Guidelines Development

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

Community Review of the Guidelines Document

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

Guidelines Paper (DOI: 10.5334/dsj-2022-008)

- Guidelines Document Maint. and Update
- Discipline Diversity

9/19 7/20 8/20 9/20 12/20 4/21 10/21 3/22

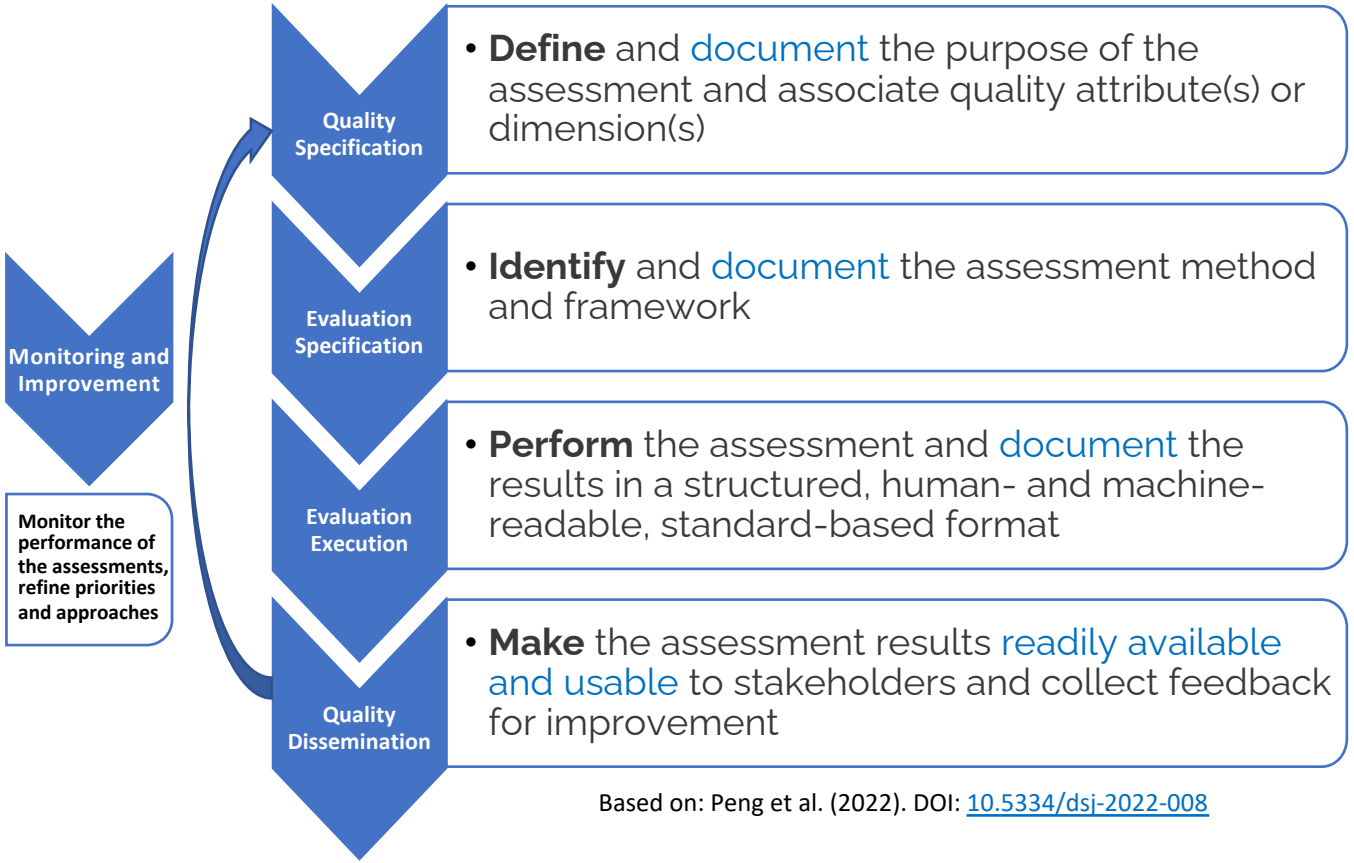
Who We Are

Community Of Practice

Participated by

- 22 International Interdisciplinary Domain Experts
  - Data producers (in situ, satellite, model),
  - Stewards (data/science/technology),
  - Services providers (data/information/infrastructure),
  - Data publishers and users.
- 7 Countries (US, ES, AU, NZ, DE, UK, GB)
- 22+ Affiliations (government, academia, private sectors)
  - Data, science and service centers, institutional repositories, companies.
- Expert Knowledge
  - Data acquisition or production,
  - Data and information management,
  - Data publishing, services, and
  - Data applications.

Basic Workflow of Curating and Disseminating DQI



Based on: Peng et al. (2022). DOI: 10.5334/dsj-2022-008

FAIR-DQI Guidelines
(At a Glance)

- **Guideline 1: Describing Dataset**
  - ➢ Ensure the dataset is findable and accessible
- **Guideline 2: Utilizing a structured quality assessment model**
  - ➢ Ensure the assessment model is findable and accessible
- **Guideline 3: Documenting the assessment method and results (dataset metadata)**
  - ➢ Ensure the quality information is interoperable and reusable (*machine end users)*
- **Guideline 4: Documenting the assessment method and results (human-readable document)**
  - ➢ Ensure the quality information is findable, accessible, citable and reusable (*human end users)*
- **Guideline 5: Reporting the dataset quality information**
  - ➢ Ensure the information is online, findable and readily (re)usable

Based on: Peng et al. (2022). DOI: 10.5334/dsj-2022-008

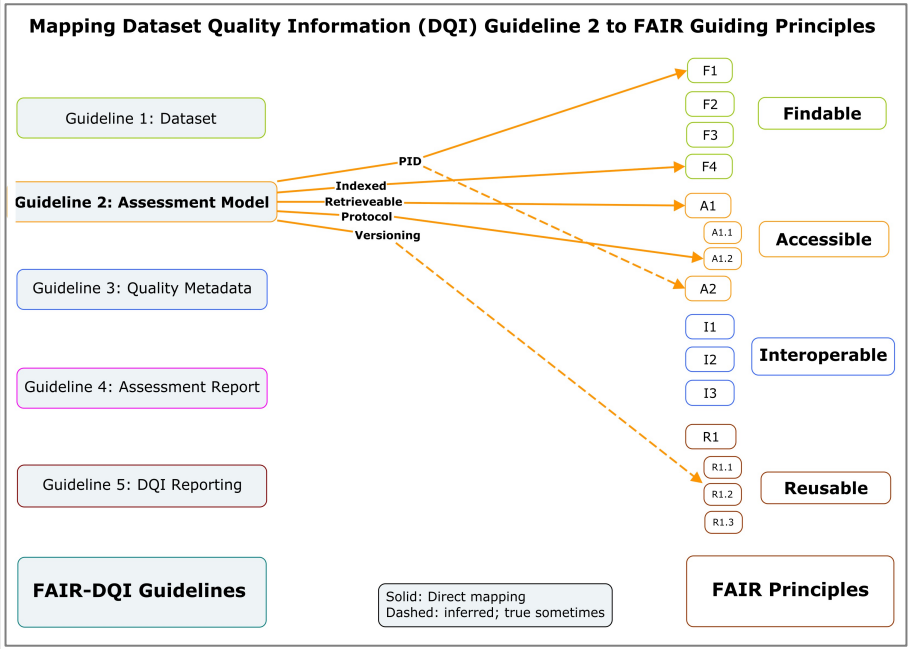
Guideline 2 In Detail

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

Based on: Peng et al. (2022). DOI: 10.5334/dsj-2022-008

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

Assessment Model	**Scientific Data Stewardship Maturity Matrix** (Peng et al. 2015)	**Stewardship Maturity Matrix for Climate Data** (Peng et al. 2019b)	**FAIR Data Maturity Model** (RDA FAIR Data Maturity Model Working Group 2020)	**Metadata Quality Framework** (Bugbee et al. 2021)	**Data Quality Analyses and Quality Control Framework** (Woo and Gourcuff 2021)
Quality Entity (i.e., attribute, aspect, or dimension)	Stewardship	Stewardship	FAIRness	Metadata	Data
2.1 - Publicly Available	Yes	Yes	Yes	Yes	Yes
2.1 - PID	DOI	DOI	DOI	DOI	DOI
2.2 - Indexed	Data Science Journal	Figshare	Zenodo	Data Science Journal	Integrated Marine Observing System Catalog
2.3 - Retrievable Using free, open, standard-based Protocol	Yes	Yes	Yes	Yes	Yes

From: Peng et al. (2022). DOI: 10.5334/dsj-2022-008

Takeaways

- Dataset quality is more than just data quality
- Quality should be considered throughout dataset lifecycle
- FAIR Principles can help with enhancing sharing of dataset quality information (DQI)
- FAIR-DQI guidelines can help get started on DQI documentation and sharing

Call-to-action statement (Peng et al. 2020): 10.5334/dsj-2021-019
Guidelines document (Peng et al. 2021): 10.31219/osf.io/xsu4p
Guidelines paper (Peng et al. 2022)**:** 10.5334/dsj-2022-008

Join ESIP IQC: https://wiki.esipfed.org/Information_Quality

Thank You!

Contact Info: ge.peng@uah.edu

International FAIR-DQI Community Guidelines Working Group

Ge Peng, Carlo Lacagnina, Ivana Ivánová,

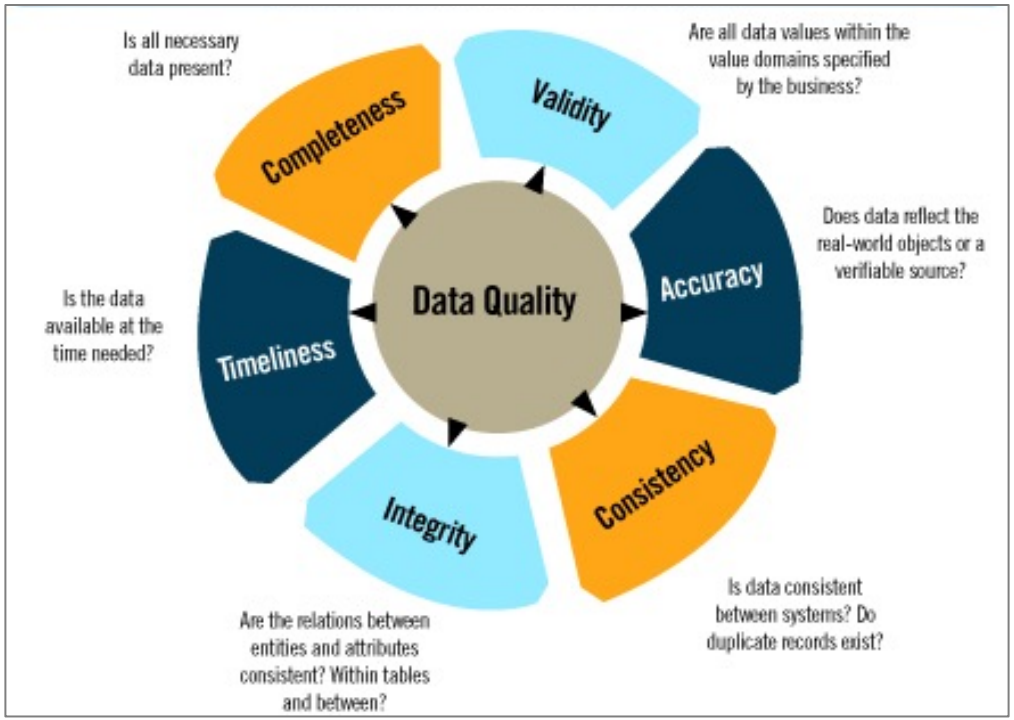
Robert R. Downs, Hampapuram Ramapriyan, Anette Ganske, Lesley Wyborn, Dave Jones, Lucy Bastin, Chung-Lin Shie, David Moroni, Irina Bastrakova, Nancy Ritchey, Mingfang Wu, Yaxing Wei, Gilles Larnicol, Kaylin Bugbee, C. Sophie Hou, Ted Habermann, Sarah Champion, Gary Berg-Cross, Jeanné le Roux

Ge Peng is supported by NASA Grant **NNM11AA01A** between UAH and MSFC Interagency Implementation and Advanced Concepts Team (IMPACT) project.

THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

IMPACT Interagency Implementation and Advanced Concepts Team

Backup Slides



Source: https://www.kovair.com/blog/data-quality-the-fundamental-of-any-data-migration-project/

Schematic diagram of dataset lifecycle stages, quality aspects and associated documentation types and metadata tags (MM-*), and metadata entities.
From: Peng et al. (2021).
DOI: 10.31219/osf.io/xsu4p