
Introducing Community Guidelines for Sharing Dataset Quality Information

Ge Peng, PhD

Sr. Principal Research Scientist

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

International FAIR-DQI Community Guidelines Working Group, Lead

Information Quality Cluster of the Earth Science Information Partners (ESIP), Co-Chair

FAIR Points Webinar, 30 Nov 2022

CC-BY-4.0

DOI: <https://doi.org/10.5281/zenodo.7314004>



In Collaboration With

International FAIR-DQI Community Guidelines Working Group

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



About Peng



 [ge-peng-37543230](https://www.linkedin.com/in/ge-peng-37543230)

- Numerical modeler
 - Observations – “data”
 - Initial/boundary conditions; Truth
- Data scientist
 - Observational data; Model output
 - Analysis: Trend and variability
- Scientific steward
 - Data products
 - Evaluation: Product quality; Stewardship maturity

About IMPACT



- Interagency project
 - under NASA Earth Science Data Systems (ESDS) program,
 - managed by Marshall Space Flight Center (MSFC)
- Interdisciplinary team
 - Domain experts: informatics, science/management/technology, ML, data systems, etc.
- Identifying and addressing Earth Science data needs
 - data acquisition and processing (e.g., CSDA, DCD, HLS-2, SNW/G),
 - data management and systems (e.g., ADMG, APT, ARC, MAAP),
 - data discovery and visualization (e.g., SDE, VEDA)

Additional information: <https://impact.earthdata.nasa.gov/>



About ESIP IQC



- ESIP Information Quality Cluster:
 - Cross-domain/agency data professionals
 - Come together to address common data/information quality challenge(s)
- Become authoritative and responsive information resource
 - Data quality standards and best practices of the Earth Science community
- Facilitate the sharing of experiences and best practices
 - Collaboration - Nationally and internationally
 - Invited speakers at monthly telecons – looking for speakers for next year
 - Sessions and/or presentations (AGU, EGU, ESIP, RDA, CEOS, OGC, etc.)

Join us: https://wiki.esipfed.org/index.php/Information_Quality

Dataset Quality

Quality of

- data (input and output),
- metadata and documentation,
- software and workflows,
- procedures and processes,
- infrastructure and tools.

Dataset refers to 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 Information (DQI)

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

- Data acquisition or production,
- Data and information management,
- Data publishing and services,
- Customer support and user engagement.



We Need (Consistently Curated) Quality Information

- **Decision-making support**
 - **Data use:** Informed decision (e.g., 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**
 - **Improved** productivity;
 - **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!

Quality information – needed but hard to find and/or integrate. **Why?**

- “Good” quality means different things to different applications
 - Weather forecast vs climate analysis
 - Same quality dimension means different things to different **quality aspects**
 - Data completeness vs metadata completeness
 - Quality information curation requires cross-domain knowledge integration
 - Data uncertainty estimates by data producer to metadata specialist
-
-

Use Cases for Documenting DQI – Quality Aspects

- When developing a data product, a scientist documents algorithm validation information
 - Scientific Quality
- When generating a data product, a data producer captures product evaluation (data uncertainty) information
 - Product Quality
- When curating dataset-level metadata, a data steward conveys the data uncertainty information
 - Stewardship Quality
- A service provider reports the quality information to users
 - Service Quality



To address the community needs,

We Have Developed

Quality-Attribute Agnostic Guidelines

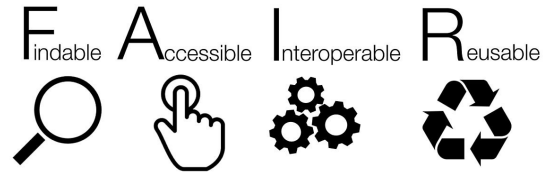
for making consistently documented quality information readily available to both machine and human end users!

➤ **For Improved Sharing and (Re)use**

How To Improve Sharing?

Adopting FAIR Guiding Principles

(Wilkinson et al. 2016)



(Image by SandyaPundir. CC BY-SA 4.0)

FAIR Data Guiding Principles

(In a Nutshell)

➤ Uniquely Identifiable
and Discoverable



Findable Principle

F1 -> PID

F2 -> Rich Metadata

F3 -> Cross-Reference PID

F4 -> Searchable Resource



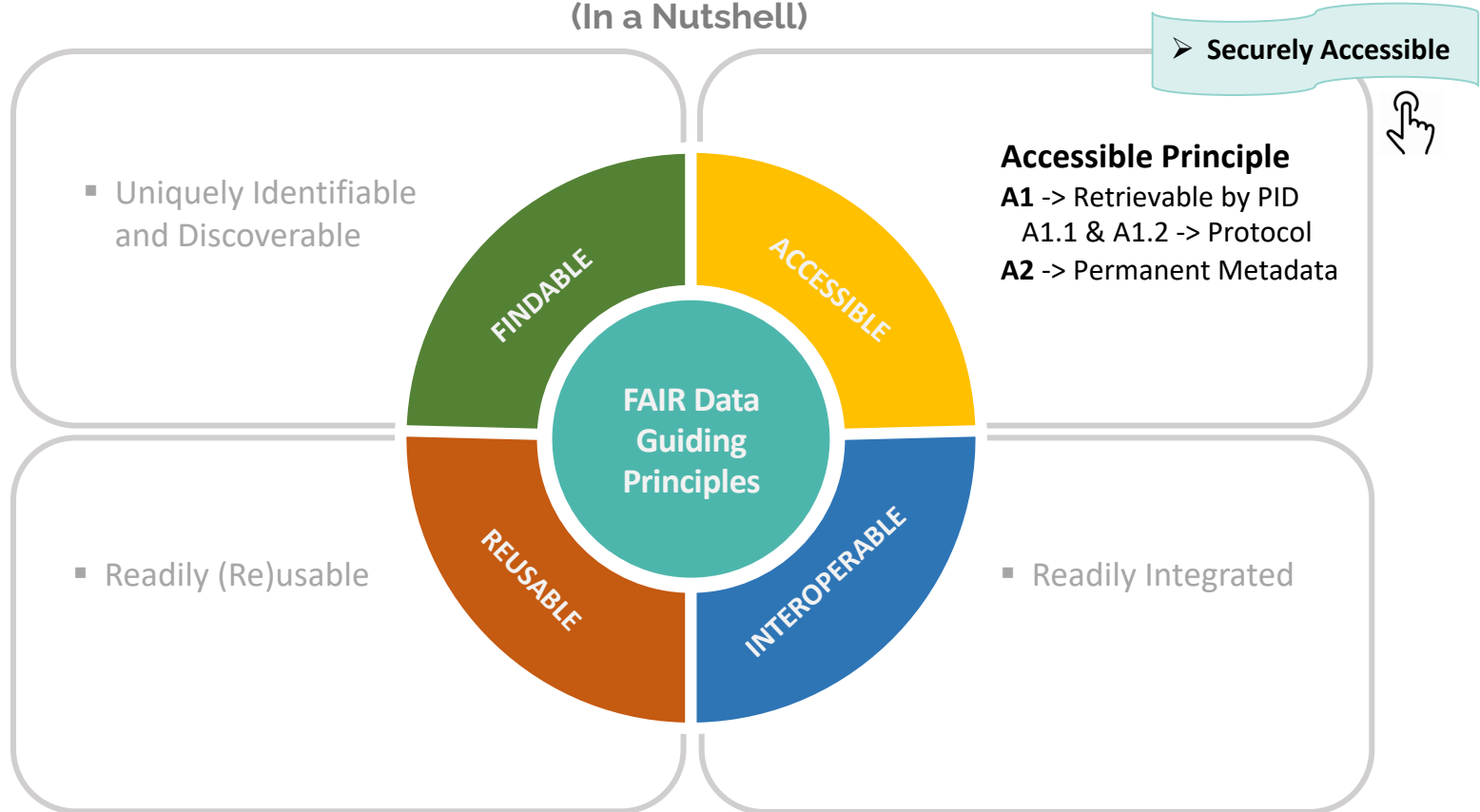
▪ Securely Accessible

▪ Readily (Re)usable

▪ Readily Integrated

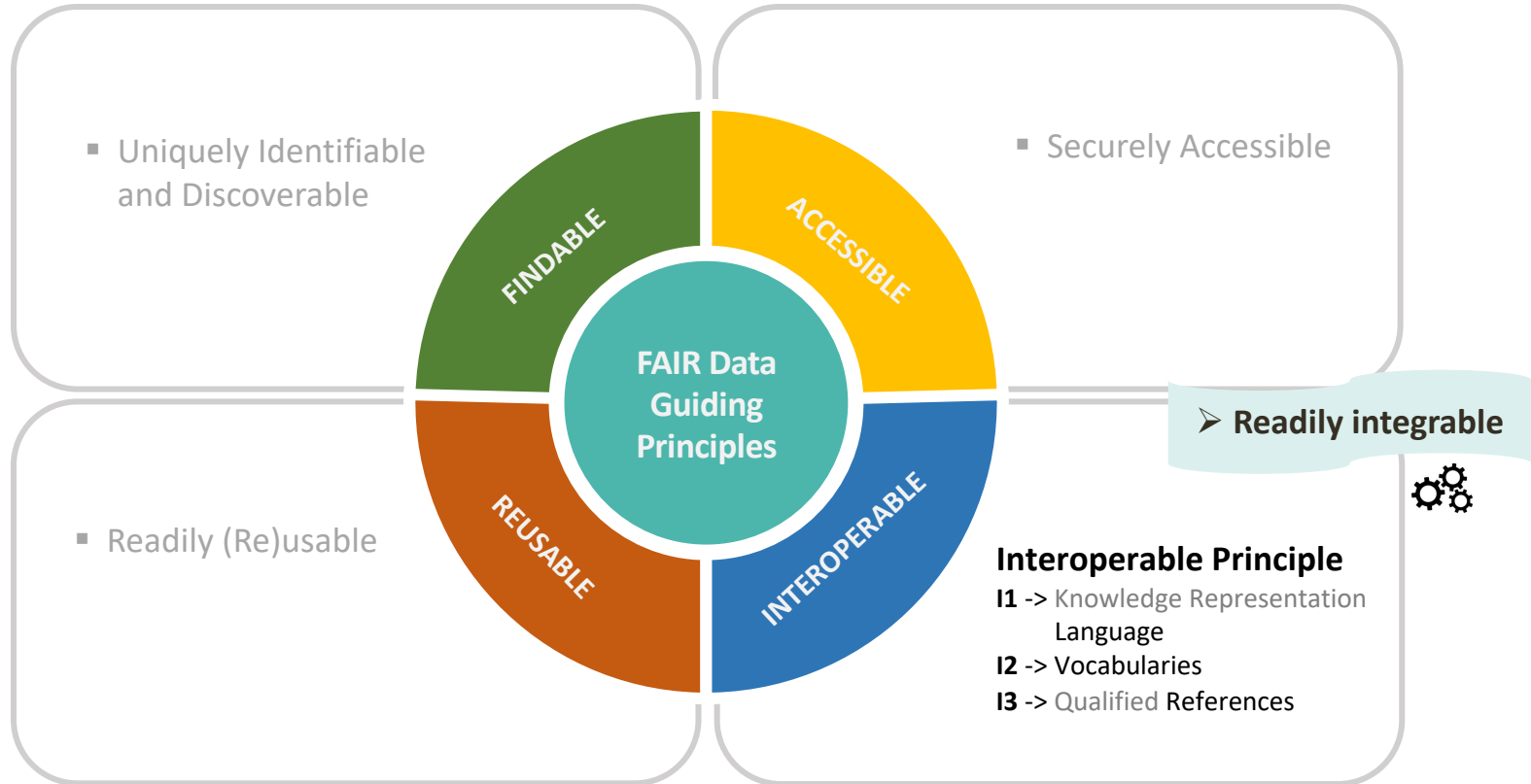
FAIR Data Guiding Principles

(In a Nutshell)



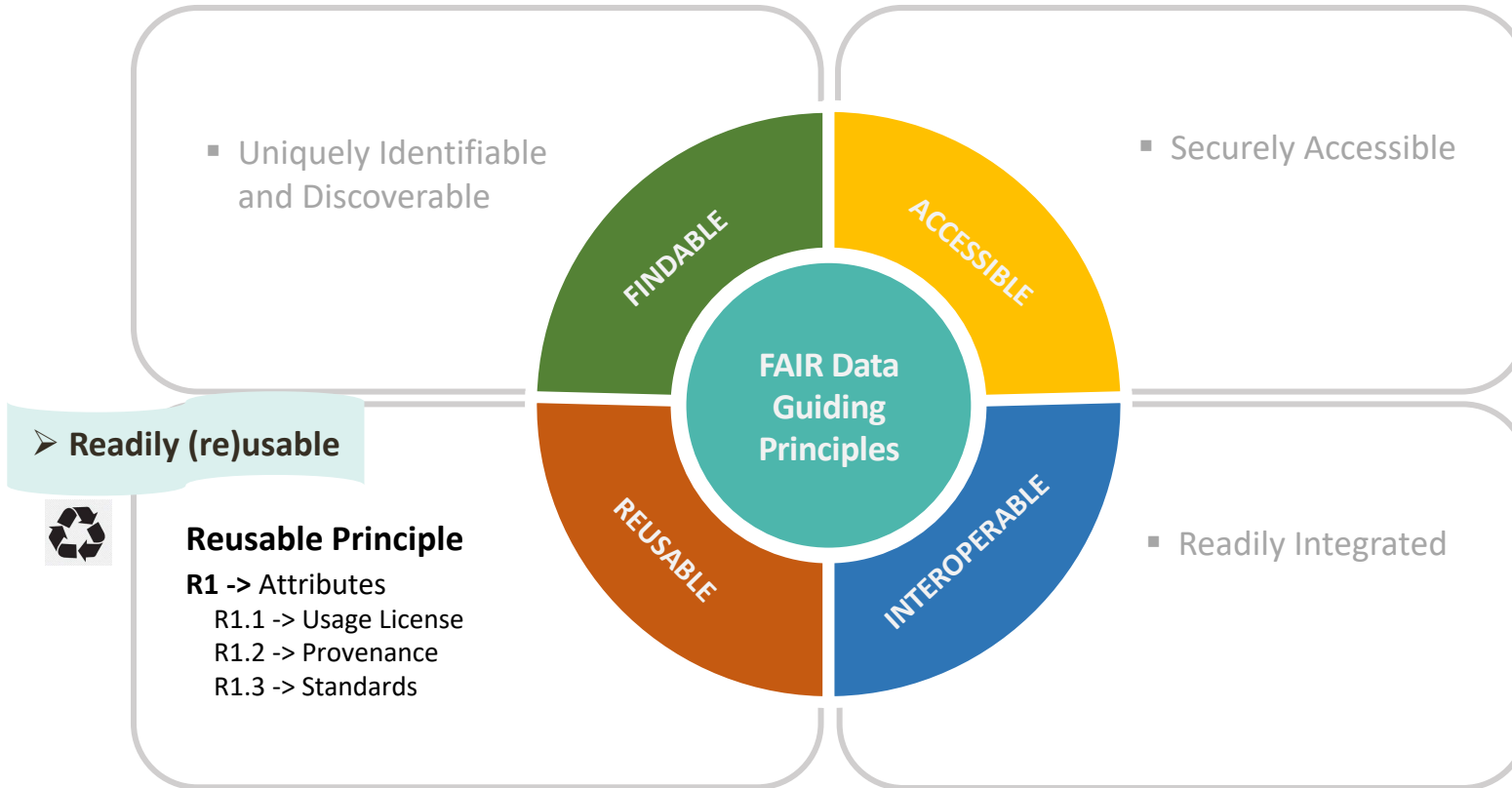
FAIR Data Guiding Principles

(In a Nutshell)



FAIR Data Guiding Principles

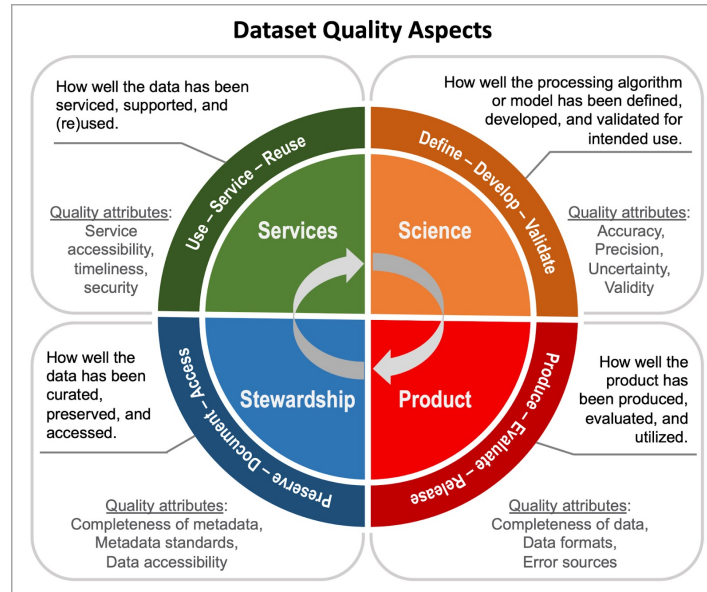
(In a Nutshell)



Practical Guidance on Consistently Reporting Quality Information

International FAIR-DQI Community Guidelines

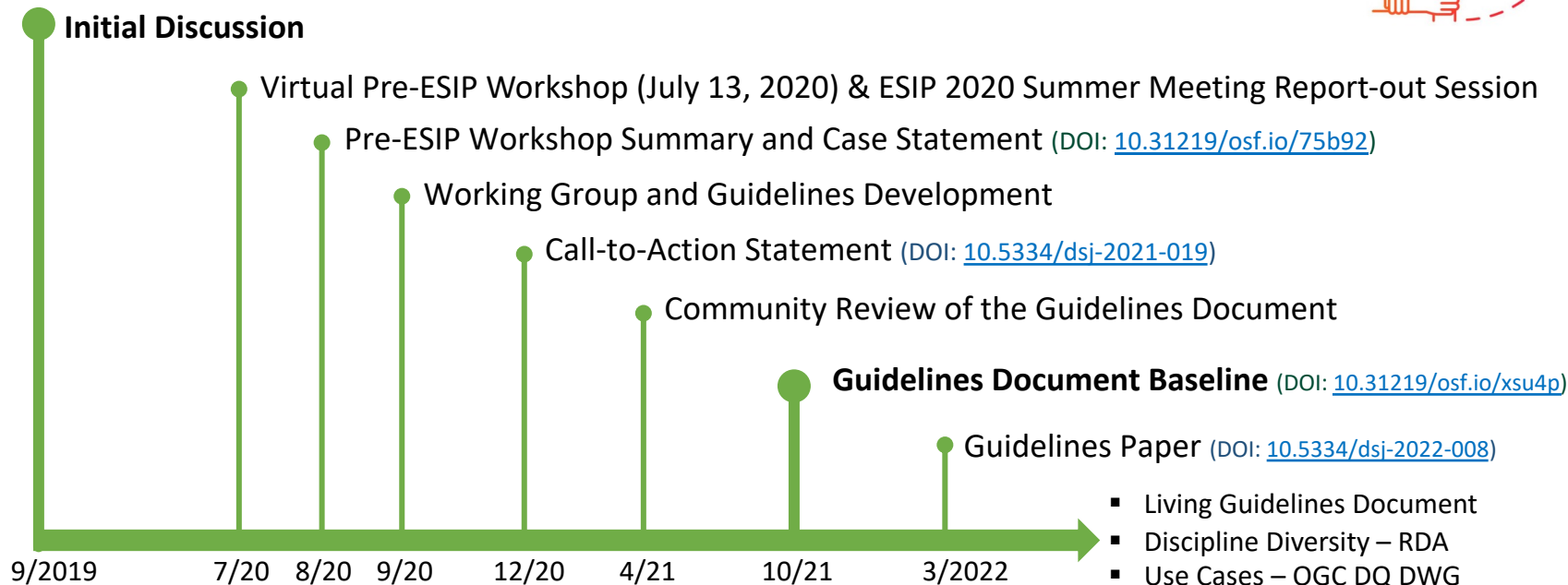
(Peng et al. 2022. *DSJ*. DOI: [10.5334/dsj-2022-008](https://doi.org/10.5334/dsj-2022-008))



Guidelines Development

Co-organized by

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



FAIR-DQI Guidelines

(At a Glance)

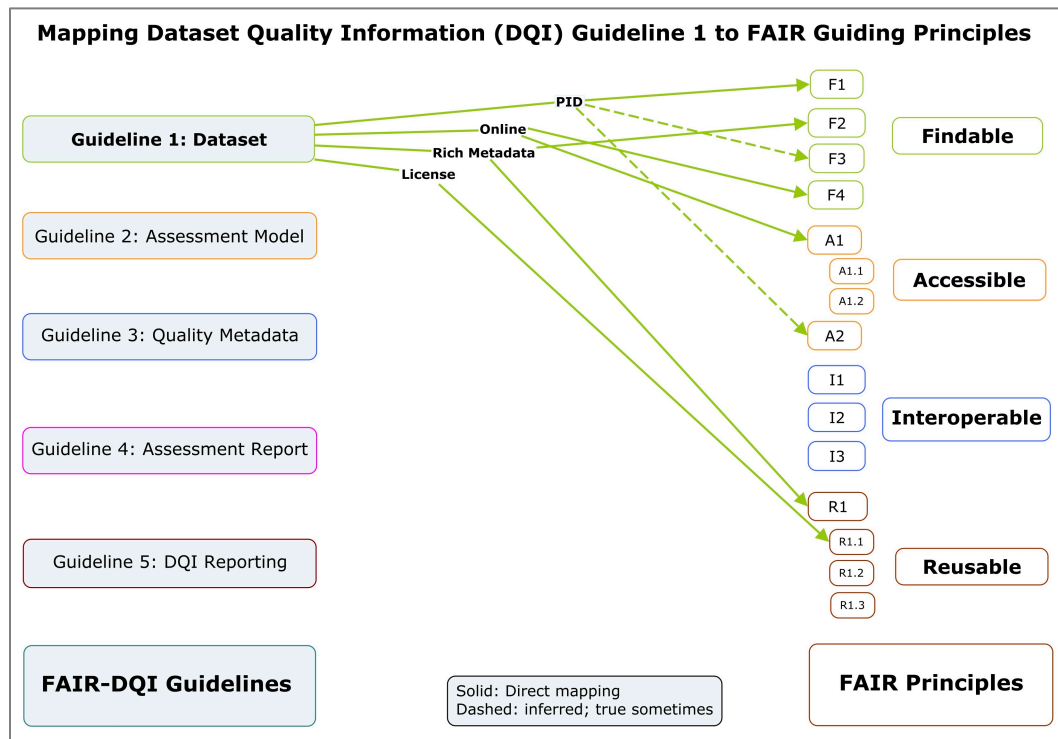
- **Guideline 1: Describing Dataset**
 - Ensure the dataset is findable, accessible and reusable
- **Guideline 2: Utilizing a quality assessment model**
 - Ensure the assessment model is structured, findable, accessible and reusable
- **Guideline 3: Documenting the assessment method and results (dataset metadata)**
 - Ensure the quality information is findable, 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

FAIR-DQI Guidelines

(In More Detail)

Guideline 1: Describe Dataset

- Title,
 - Persistent identifier (PID) resolvable to a comprehensive landing page,
 - Version,
 - Data producer,
 - Publication/update date,
 - Publisher,
 - Date accessed,
 - Usage license.
- **Ensure** the dataset is findable, accessible, and reusable

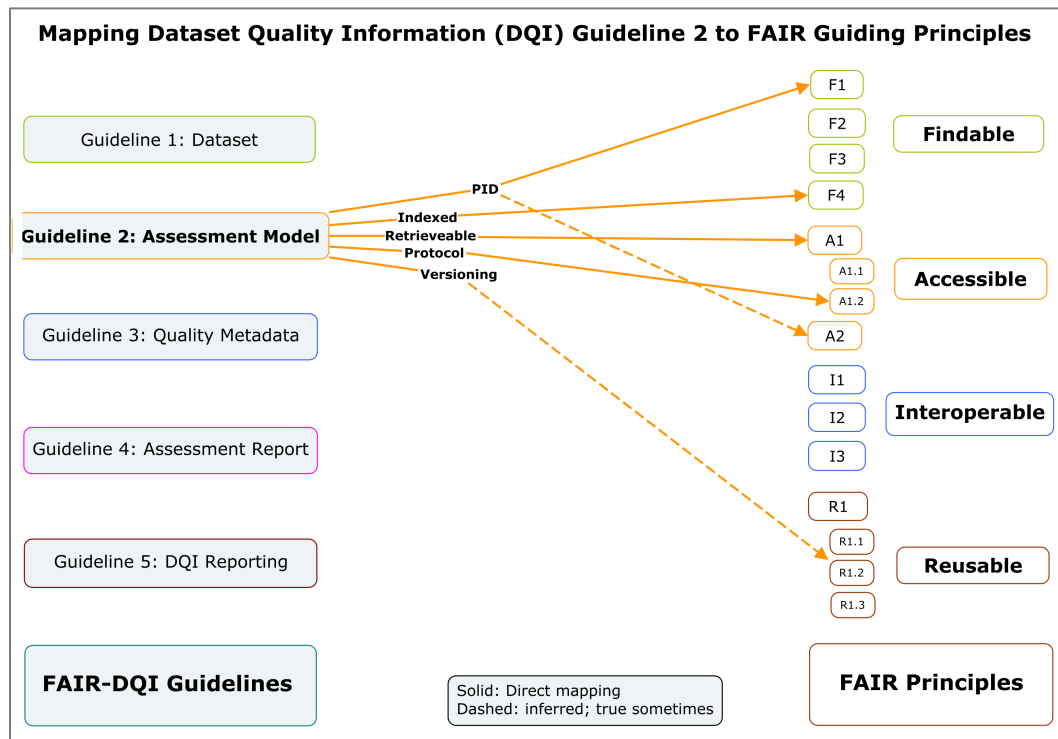


FAIR-DQI Guidelines

(In More Detail)

Guideline 2: Utilize a quality assessment model

- Structured (1, ..., N dimensions),
 - Versioned,
 - Publicly available with a unique, resolvable PID,
 - Registered or indexed in a searchable resource,
 - Retrievable using a standardized protocol.
- Ensure the assessment model is structured, findable, accessible and reusable



FAIR-DQI Guidelines

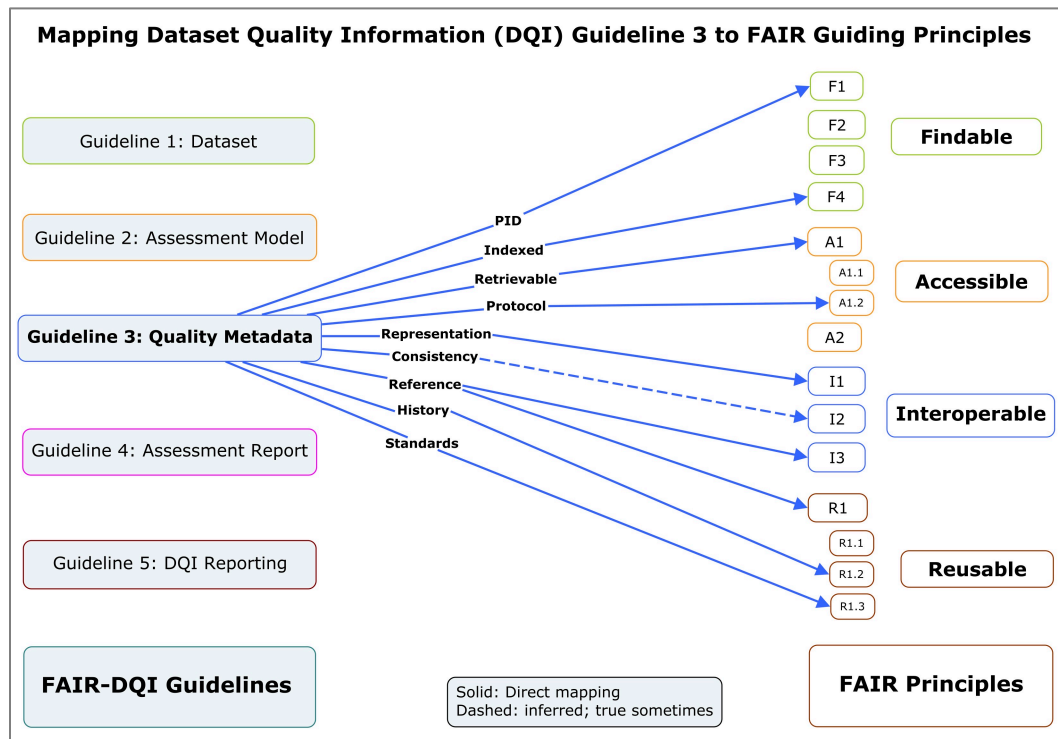
(In More Detail)

Guideline 3: Capture the quality attribute, assessment method and results in dataset-level metadata record

- Including versioning and history of the assessments

using a framework or schema

- Semantically and structurally consistent,
- Following community standards:
 - Compliant with Guideline 2
- Ensure the quality information is findable, accessible, interoperable and reusable by machine end-users



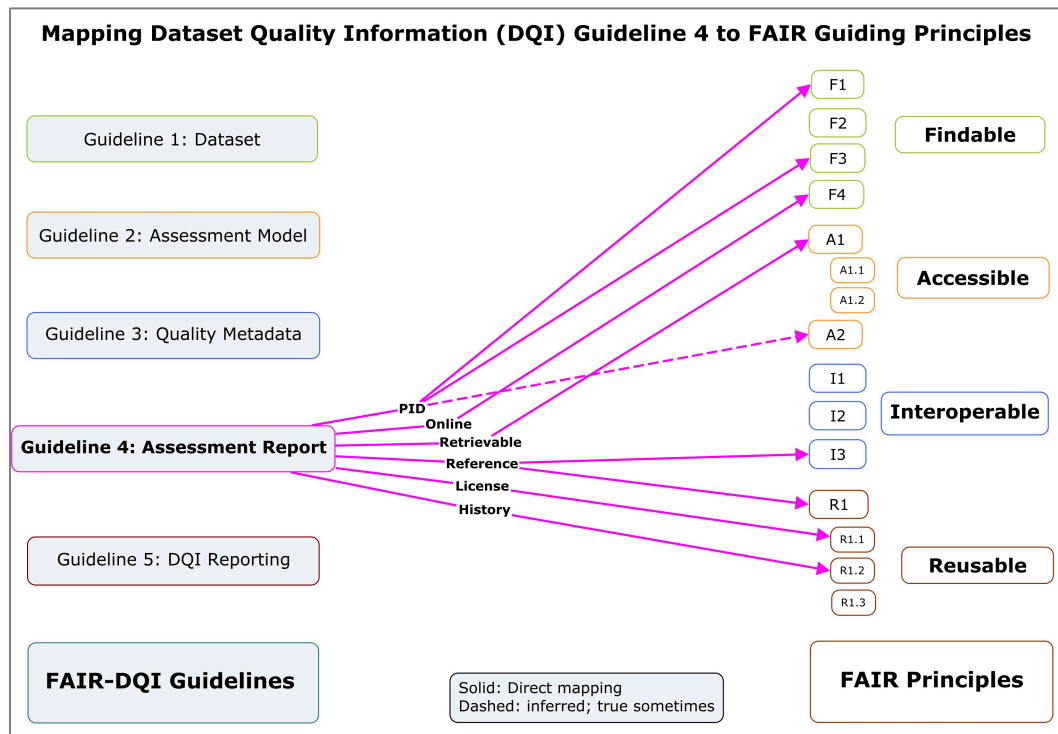
Based on: Peng et al. (2022). DOI: [10.5334/dsj-2022-008](https://doi.org/10.5334/dsj-2022-008)

FAIR-DQI Guidelines

(In More Detail)

Guideline 4: Describe the assessment method, workflow, and results in a human-readable quality report

- Using a template – published with PID, findable & accessible,
 - Published with PID, open license & the report history,
 - Linking the report PID to the dataset-level metadata record.
- **Ensure** the quality information is findable, accessible, citable and reusable by human end-users

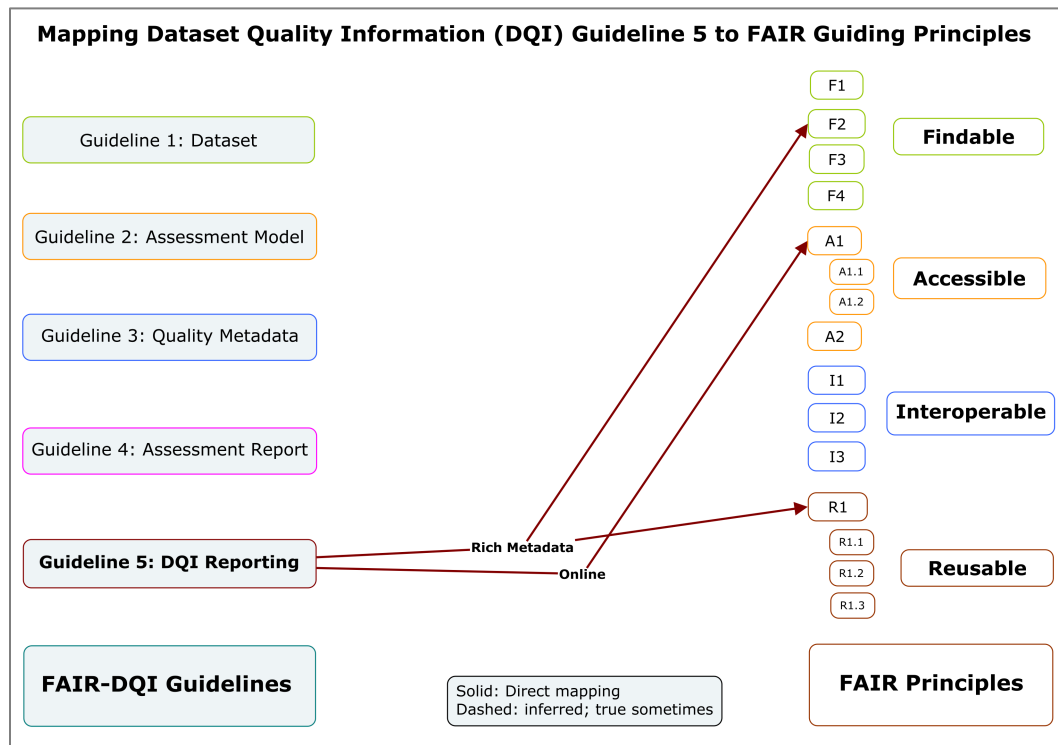


FAIR-DQI Guidelines

(In More Detail)

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

- Dataset (Guideline 1),
 - Assessed quality attribute /dimension,
 - Assessment method and process,
 - How to understand and use the information.
- **Ensure** the information is online, described, understandable and readily usable



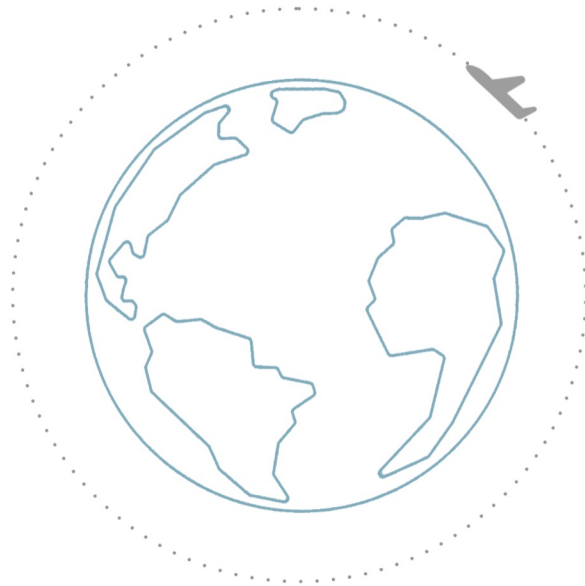
Takeaways

- Dataset quality is **more than** just data quality
 - Quality information **should be documented** throughout the entire dataset lifecycle
 - FAIR Principles **can help with** enhancing the sharing of dataset quality information (DQI)
 - FAIR-DQI guidelines **can help get** started on documenting and reporting DQI
-
-

Call-to-action statement (Peng et al. 2020): [10.5334/dsj-2021-019](https://doi.org/10.5334/dsj-2021-019)

Guidelines document (Peng et al. 2021): [10.31219/osf.io/xsu4p](https://doi.org/10.31219/osf.io/xsu4p)

Guidelines paper (Peng et al. 2022): [10.5334/dsj-2022-008](https://doi.org/10.5334/dsj-2022-008)



Thank You!

**Let me know your quality use case
and/or feedback on the guidelines**

 ge.peng@uah.edu

 @DrPengAtAVL

 [ge-peng-37543230](https://www.linkedin.com/in/ge-peng-37543230)

Additional Slides

FAIR Questions

You've assessed the FAIRness of your data holding!


THEN WHAT?

How would your end-users:

- **Know** what you have assessed?
 - **Understand and (re)use** your methods/approaches?
 - **Get** your assessment results?
 - **Readily integrate** the FAIRness information across tools?
-
-

Utilizing the Guidelines

From the FAIRness Assessment Perspective

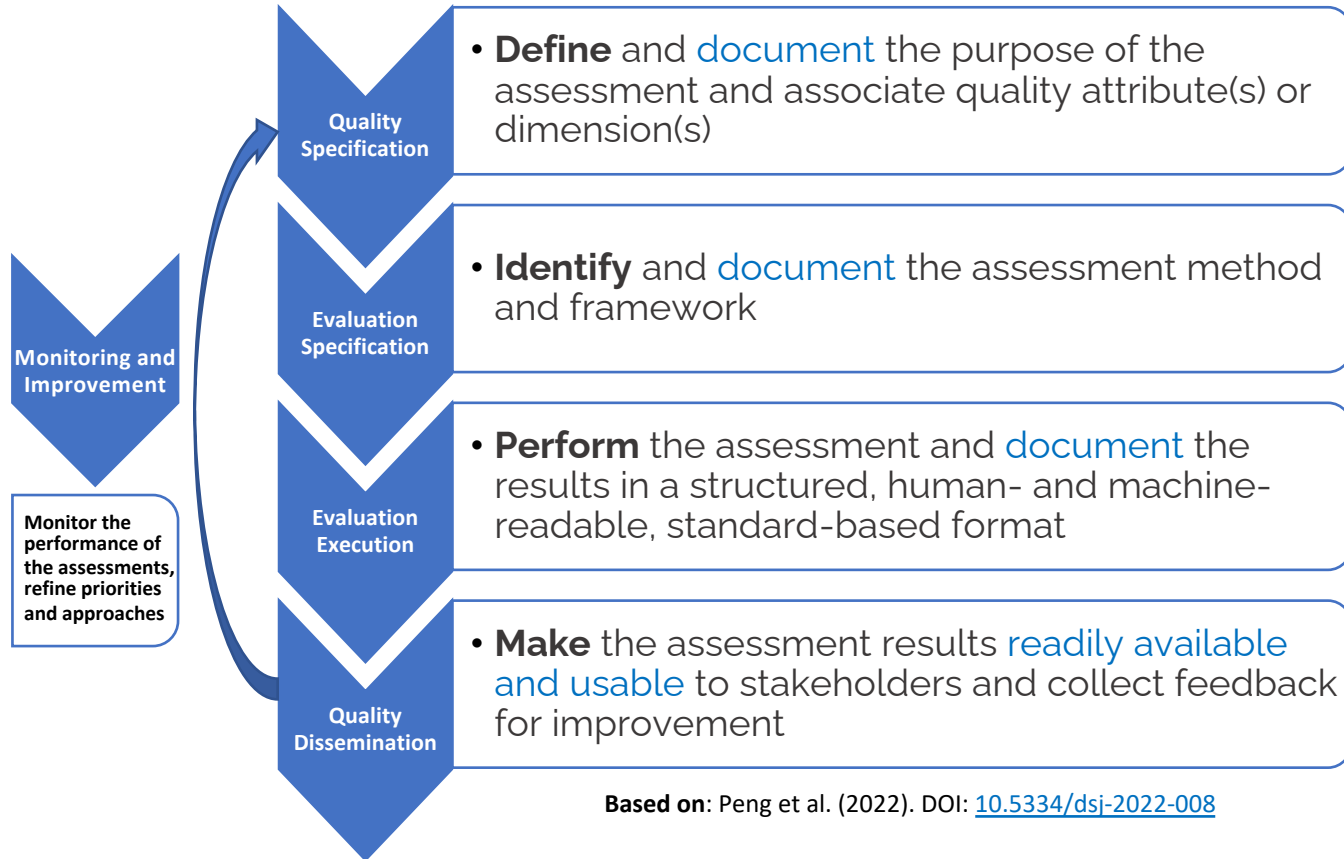
- 
- **Describe** the dataset/data holding to be assessed -> [Guideline 1](#)
 - **Select** a FAIRness assessment model -> [Guideline 2](#)
 - **Document** the quality attribute(s), method and results in a structured, searchable, machine-actionable metadata record -> [Guideline 3](#)
 - **Document** the quality attribute(s), method, process and results in a structured, findable, citable, human-readable report -> [Guideline 4](#)
 - **Describe** and **disseminate** the FAIRness information online -> [Guideline 5](#)

➤ **Ensure:**

- The FAIRness assessment processing is transparent;
- The quality attributes, assessment method, process and results are
 - Documented, findable, machine-actionable, human-readable, and reusable;
 - Available online, comprehensively described, easily understandable and readily (re)usable.

Backup Slides

Basic Workflow of Curating and Disseminating DQI



Based on: Peng et al. (2022). DOI: [10.5334/dsj-2022-008](https://doi.org/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 - Unique 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

Source: Peng et al. (2022). DOI: [10.5334/dsj-2022-008](https://doi.org/10.5334/dsj-2022-008)

Examples of representing quality entities, assessment models and assessment results in machine-readable quality metadata and their compliance with Guideline 3

Quality Metadata Framework	NOAA <i>OneStop</i> DSMM Quality Metadata (Peng et al. 2019a)	AtMoDat Maturity Indicator (Heydebreck et al. 2020)	MetadataFromGeodata (Wagner et al. 2021)
Quality Entity	Stewardship	Any Quality Entity	Data and Metadata
3.1 - Semantically and Structurally Consistent	Yes	Yes	Yes
3.1 - Metadata Framework/Schema	International	Domain	Domain
3.2 - Quality Entity Description	Yes	Yes	Yes
3.3 - Assessment Method/structure Description	Yes	Yes	Partly (contains evaluation of quality description and not description of quality assessment)
3.4 - Assessment Results Description	Yes	Yes	Yes
3.5 - Versioning and the history of the assessments	Yes	Versioning	Creation/Last Update Dates

Source: Peng et al. (2022). DOI: [10.5334/dsj-2022-008](https://doi.org/10.5334/dsj-2022-008)



LEGEND

- Dataset Lifecycle Stages
- Dataset Quality Aspects
- Document Types
- Metadata Tags
- Metadata Entities

Schematic diagram of dataset lifecycle stages, quality aspects and associated documentation types and metadata tags (MM-*), and metadata entities.

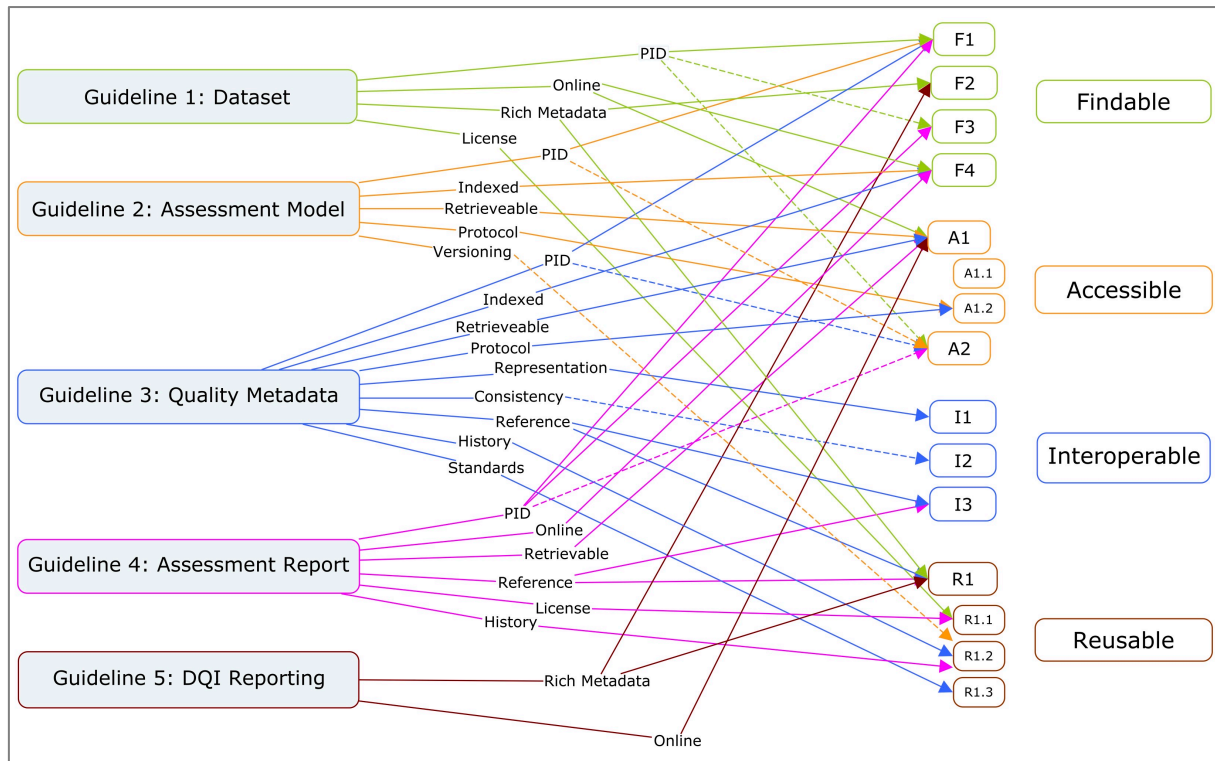
Source: Peng et al. (2021).
 DOI: [10.31219/osf.io/xsu4p](https://doi.org/10.31219/osf.io/xsu4p)

Version: v02r01 202109719
 POC: gpeng93@gmail.com
 CC-BY 4.0

Mapping the FAIR-DQI Guidelines to the FAIR Principles

FAIR-DQI Guidelines

FAIR Principles



Based on: Peng et al. (2022). DOI: [10.5334/dsj-2022-008](https://doi.org/10.5334/dsj-2022-008)

Solid lines: direct mapping
Dashed lines: inferred; true sometimes