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

CEOS
Committee on
Earth Observation Satellites

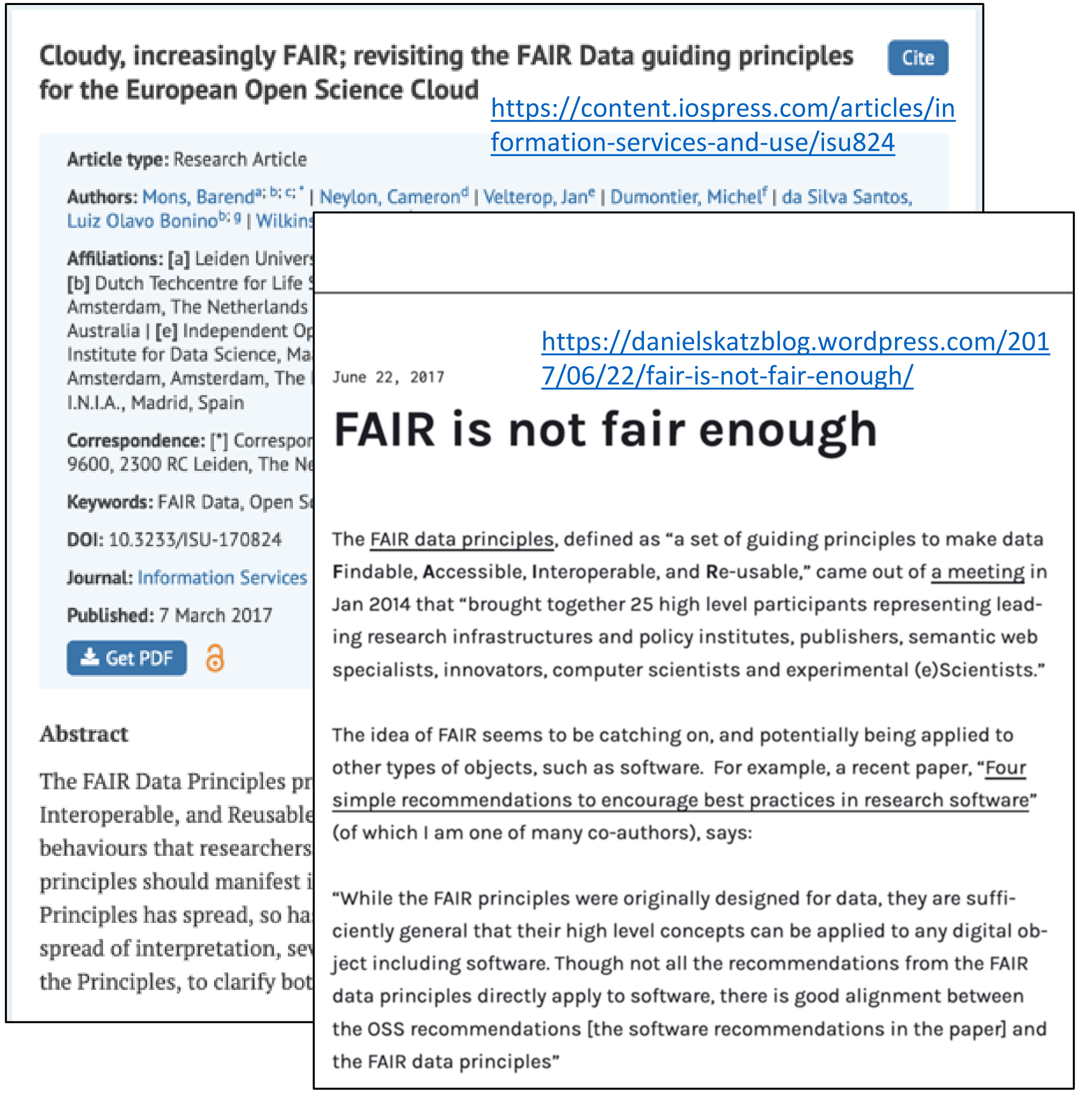
Executive Summary

- 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

Cite

https://content.iospress.com/articles/information-services-and-use/isu824

Article type: Research Article

Authors: Mons, Barend[a; b; c; *] | Neylon, Cameron[d] | Velterop, Jan[e] | Dumontier, Michel[f] | da Silva Santos, Luiz Olavo Bonino[b; g] | Wilkins

Affiliations: [a] Leiden Univers
[b] Dutch Techcentre for Life S
Amsterdam, The Netherlands
Australia | [e] Independent Op
Institute for Data Science, Ma
Amsterdam, Amsterdam, The
I.N.I.A., Madrid, Spain

Correspondence: [*] Correspon
9600, 2300 RC Leiden, The Ne

Keywords: FAIR Data, Open S

DOI: 10.3233/ISU-170824

Journal: Information Services

Published: 7 March 2017

Get PDF

Abstract

The FAIR Data Principles pr
Interoperable, and Reusable
behaviours that researchers
principles should manifest i
Principles has spread, so ha
spread of interpretation, se
the Principles, to clarify bot

June 22, 2017

https://danielskatzblog.wordpress.com/2017/06/22/fair-is-not-fair-enough/

FAIR is not fair enough

The FAIR data principles, defined as "a set of guiding principles to make data Findable, Accessible, Interoperable, and Re-usable," came out of a meeting in Jan 2014 that "brought together 25 high level participants representing leading research infrastructures and policy institutes, publishers, semantic web specialists, innovators, computer scientists and experimental (e)Scientists."

The idea of FAIR seems to be catching on, and potentially being applied to other types of objects, such as software. For example, a recent paper, "Four simple recommendations to encourage best practices in research software" (of which I am one of many co-authors), says:

"While the FAIR principles were originally designed for data, they are sufficiently general that their high level concepts can be applied to any digital object including software. Though not all the recommendations from the FAIR data principles directly apply to software, there is good alignment between the OSS recommendations [the software recommendations in the paper] and the FAIR data principles"

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.

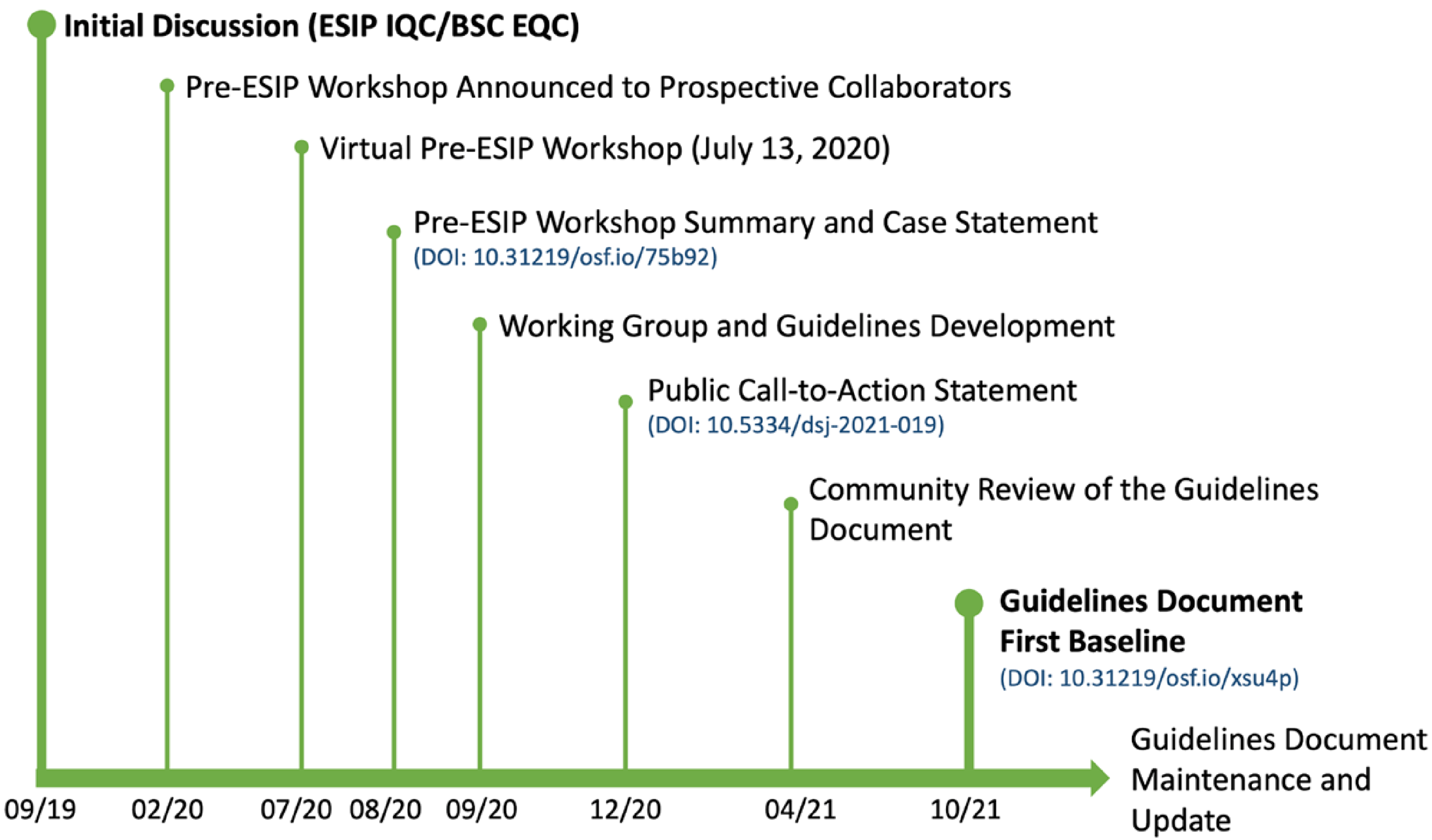
TOGETHER

WE GO FAR

NOAA
NASA
USGS
JPSS
e
cnrs
EUMETSAT
ARDC
AGU
WMO
NSIDC
NIH NIEHSSRP
LDEO
The University of Iowa
Landcare Research Manaaki Whenua
PANGAEA.
OGC
JPL
Dartmouth
Michigan State University
UNCW
CASE
GNS Science
Mercator Ocean
Australian Government Geoscience Australia
Curtin University
Oak Ridge National Laboratory
SDSC
The University of New Mexico
Metadata Game Changers
Riverside
TIB
Aston University
Universität Bremen
AERIS
NC State University
Goddard Space Flight Center
KBR
StormCenter
Wilson Center
NPL
Apogee
DKRZ
NCICS
CIESIN Columbia University
SSAI
United States Global Change Research Program
NCI
RHEA Group
BSC
ESIP

13 July 2020 CC-BY 4.0 bit.ly/Pre-ESIP-FAIR-DQI-Wkshp

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

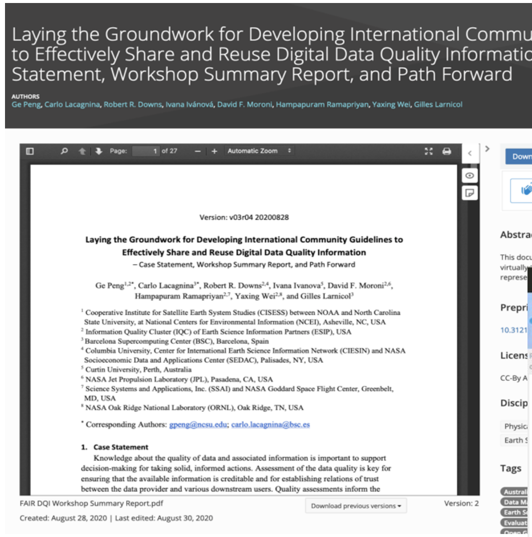
Public Call-to-Action Statement
(DOI: 10.5334/dsj-2021-019)

Community Review of the Guidelines Document

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

Guidelines Document Maintenance and Update

Four key outputs so far



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

AUTHORS
Ge Peng, Carlo Lacagnina, Robert R. Downs, Ivana Ivánová, David F. Moroni, Hampapuram Ramapriyan, Yaxing Wei, Gilles Larnicol

Page: 1 of 27 Automatic Zoom

Version: v03r04 20200828

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

Ge Peng[1,2*], Carlo Lacagnina[3*], Robert R. Downs[2,4], Ivana Ivanova[5], David F. Moroni[2,6], Hampapuram Ramapriyan[2,7], Yaxing Wei[2,8], and Gilles Larnicol[3]

[1] Cooperative Institute for Satellite Earth System Studies (CISESS) between NOAA and North Carolina State University, at National Centers for Environmental Information (NCEI), Asheville, NC, USA
[2] Information Quality Cluster (IQC) of Earth Science Information Partners (ESIP), USA
[3] Barcelona Supercomputing Center (BSC), Barcelona, Spain
[4] Columbia University, Center for International Earth Science Information Network (CIESIN) and NASA Socioeconomic Data and Applications Center (SEDAC), Palisades, NY, USA
[5] Curtin University, Perth, Australia
[6] NASA Jet Propulsion Laboratory (JPL), Pasadena, CA, USA
[7] Science Systems and Applications, Inc. (SSAI) and NASA Goddard Space Flight Center, Greenbelt, MD, USA
[8] NASA Oak Ridge National Laboratory (ORNL), Oak Ridge, TN, USA

* Corresponding Authors: gpeng@ncsu.edu; carlo.lacagnina@bsc.es

1. **Case Statement**

Knowledge about the quality of data and associated information is important to support decision-making for taking solid, informed actions. Assessment of the data quality is key for ensuring that the available information is creditable and for establishing relations of trust between the data provider and various downstream users. Quality assessments inform the

FAIR DQI Workshop Summary Report.pdf Download previous versions ▾ Version: 2

Created: August 28, 2020 | Last edited: August 30, 2020

Download preprint

plaudit Nancy Ritchey has endorsed

Abstract

This document provides background for and summari

virtually

represe

Prepri

10.3121

Licens

CC-By A

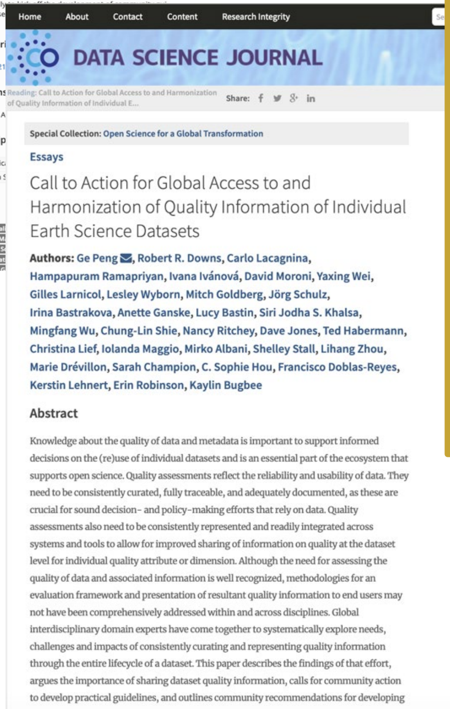
Discip

Physic

Earth S

Tags

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



Home About Contact Content Research Integrity

DATA SCIENCE JOURNAL

Reading: Call to Action for Global Access to and Harmonization of Quality Information of Individual E... Share:

Special Collection: Open Science for a Global Transformation

Essays

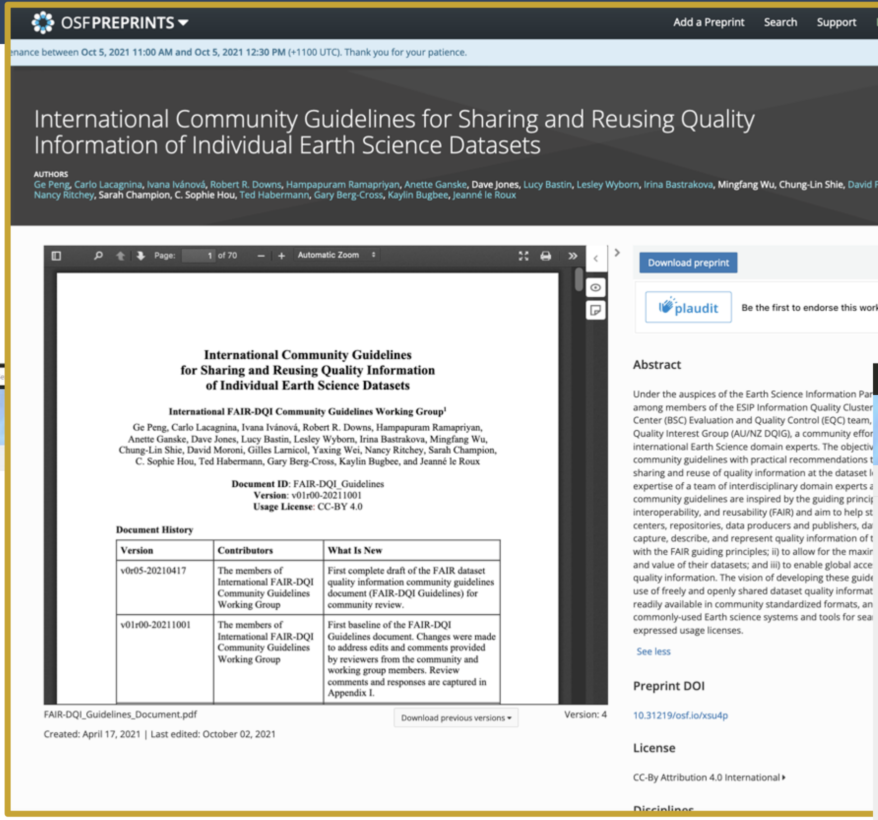
Call to Action for Global Access to and Harmonization of Quality Information of Individual Earth Science Datasets

Authors: Ge Peng, Robert R. Downs, Carlo Lacagnina, Hampapuram Ramapriyan, Ivana Ivánová, David Moroni, Yaxing Wei, Gilles Larnicol, Lesley Wyborn, Mitch Goldberg, Jörg Schulz, Irina Bastrakova, Anette Ganske, Lucy Bastin, Siri Jodha S. Khalsa, Mingfang Wu, Chung-Lin Shie, Nancy Ritchey, Dave Jones, Ted Habermann, Christina Lief, Iolanda Maggio, Mirko Albani, Shelley Stall, Lihang Zhou, Marie Drévillon, Sarah Champion, C. Sophie Hou, Francisco Doblas-Reyes, Kerstin Lehnert, Erin Robinson, Kaylin Bugbee

Abstract

Knowledge about the quality of data and metadata is important to support informed decisions on the (re)use of individual datasets and is an essential part of the ecosystem that supports open science. Quality assessments reflect the reliability and usability of data. They need to be consistently curated, fully traceable, and adequately documented, as these are crucial for sound decision- and policy-making efforts that rely on data. Quality assessments also need to be consistently represented and readily integrated across systems and tools to allow for improved sharing of information on quality at the dataset level for individual quality attribute or dimension. Although the need for assessing the quality of data and associated information is well recognized, methodologies for an evaluation framework and presentation of resultant quality information to end users may not have been comprehensively addressed within and across disciplines. Global interdisciplinary domain experts have come together to systematically explore needs, challenges and impacts of consistently curating and representing quality information through the entire lifecycle of a dataset. This paper describes the findings of that effort, argues the importance of sharing dataset quality information, calls for community action to develop practical guidelines, and outlines community recommendations for developing

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



OSFPREPRINTS ▾ Add a Preprint Search Support Donate Sign Up Sign In

nance between Oct 5, 2021 11:00 AM and Oct 5, 2021 12:30 PM (+1100 UTC). Thank you for your patience.

International Community Guidelines for Sharing and Reusing Quality Information of Individual Earth Science Datasets

AUTHORS
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 F. Moroni, Gilles Larnicol, Yaxing We
Nancy Ritchey, Sarah Champion, C. Sophie Hou, Ted Habermann, Gary Berg-Cross, Kaylin Bugbee, Jeanné le Roux

Page: 1 of 70 Automatic Zoom

International Community Guidelines for Sharing and Reusing Quality Information of Individual Earth Science Datasets

International FAIR-DQI Community Guidelines Working Group[1]

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

Document ID: FAIR-DQI_Guidelines
Version: v01r00-20211001
Usage License: CC-BY 4.0

Document History

Version	Contributors	What Is New
v0r05-20210417	The members of International FAIR-DQI Community Guidelines Working Group	First complete draft of the FAIR dataset quality information community guidelines document (FAIR-DQI Guidelines) for community review.
v01r00-20211001	The members of International FAIR-DQI Community Guidelines Working Group	First baseline of the FAIR-DQI Guidelines document. Changes were made to address edits and comments provided by reviewers from the community and working group members. Review comments and responses are captured in Appendix I.

FAIR-DQI_Guidelines_Document.pdf Download previous versions ▾ Version: 4

Created: April 17, 2021 | Last edited: October 02, 2021

Download preprint Downloads: 290

plaudit Be the first to endorse this work

Abstract

Under the auspices of the Earth Science Information Par
among members of the ESIP Information Quality Cluster
Center (BSC) Evaluation and Quality Control (EQC) team,
Quality Interest Group (AU/NZ DQIG), a community effor
international Earth Science domain experts. The objectiv
community guidelines with practical recommendations t
sharing and reuse of quality information at the dataset l
expertise of a team of interdisciplinary domain experts a
community guidelines are inspired by the guiding princip
interoperability, and reusability (FAIR) and aim to help st
centers, repositories, data producers and publishers, da
capture, describe, and represent quality information of t
with the FAIR guiding principles; ii) to allow for the maxir
and value of their datasets; and iii) to enable global acce
quality information. The vision of developing these guide
use of freely and openly shared dataset quality informat
readily available in community standardized formats, an
commonly-used Earth science systems and tools for sear
expressed usage licenses.

See less

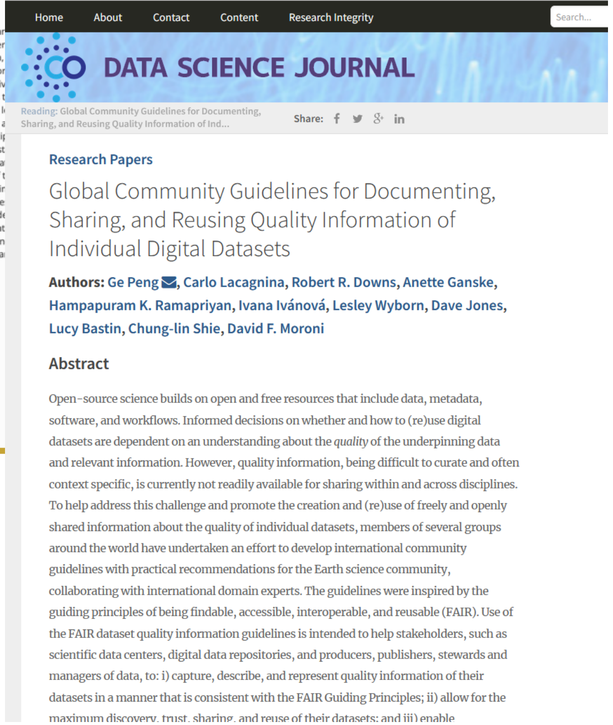
Preprint DOI

10.31219/osf.io/xsu4p

License

CC-By Attribution 4.0 International ▸

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



Home About Contact Content Research Integrity Search...

DATA SCIENCE JOURNAL

Reading: Global Community Guidelines for Documenting, Sharing, and Reusing Quality Information of Ind... Share:

Research Papers

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

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

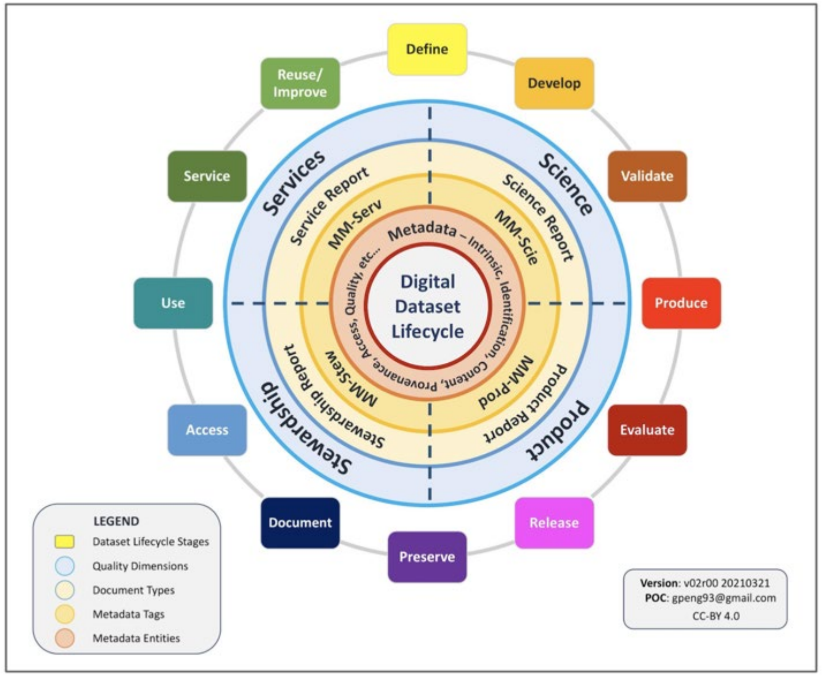
Abstract

Open-source science builds on open and free resources that include data, metadata, software, and workflows. Informed decisions on whether and how to (re)use digital datasets are dependent on an understanding about the *quality* of the underpinning data and relevant information. However, quality information, being difficult to curate and often context specific, is currently not readily available for sharing within and across disciplines. To help address this challenge and promote the creation and (re)use of freely and openly shared information about the quality of individual datasets, members of several groups around the world have undertaken an effort to develop international community guidelines with practical recommendations for the Earth science community, collaborating with international domain experts. The guidelines were inspired by the guiding principles of being findable, accessible, interoperable, and reusable (FAIR). Use of the FAIR dataset quality information guidelines is intended to help stakeholders, such as scientific data centers, digital data repositories, and producers, publishers, stewards and managers of data, to: i) capture, describe, and represent quality information of their datasets in a manner that is consistent with the FAIR Guiding Principles; ii) allow for the maximum discovery, trust, sharing, and reuse of their datasets; and iii) enable

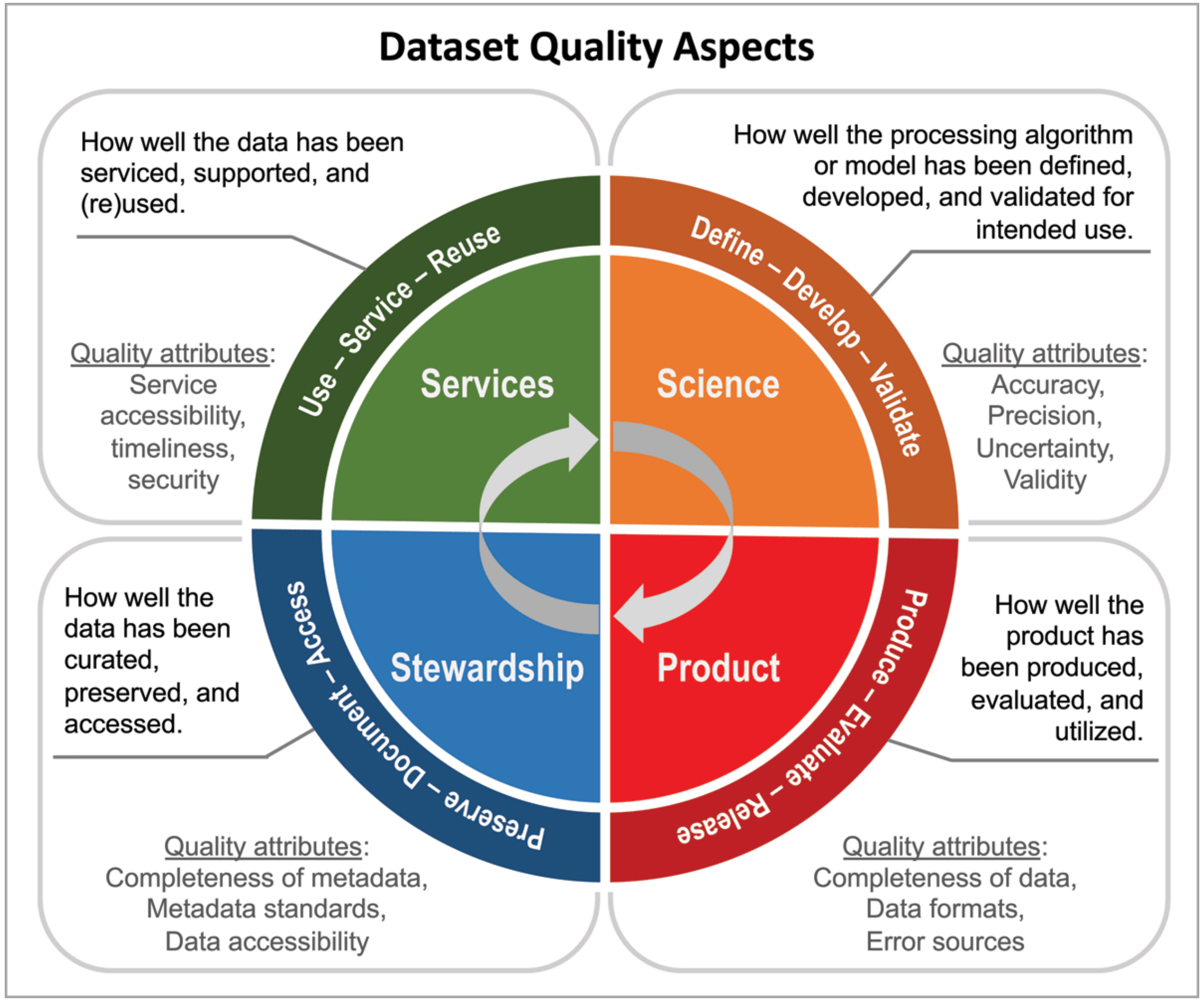
March 2022 - http://doi.org/10.5334/dsj-2022-008

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



Dataset Quality Aspects

How well the data has been serviced, supported, and (re)used.

Quality attributes: Service accessibility, timeliness, security

How well the processing algorithm or model has been defined, developed, and validated for intended use.

Quality attributes: Accuracy, Precision, Uncertainty, Validity

How well the data has been curated, preserved, and accessed.

Quality attributes: Completeness of metadata, Metadata standards, Data accessibility

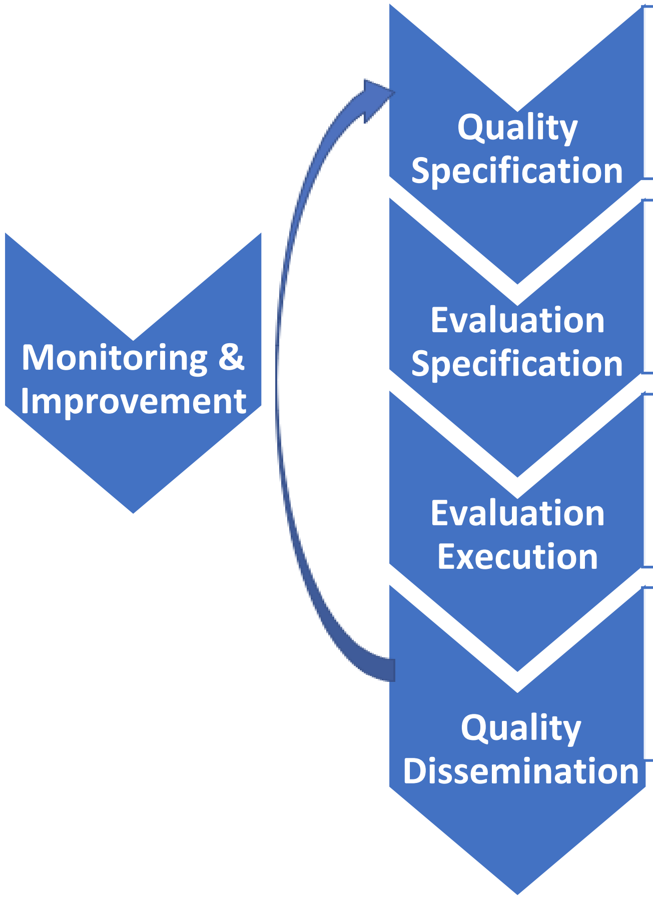
How well the product has been produced, evaluated, and utilized.

Quality attributes: Completeness of data, Data formats, Error sources

Use – Service – Reuse
Services
Define – Develop – Validate
Science
Access – Document – Preserve
Stewardship
Produce – Evaluate – Release
Product

Based on: https://doi.org/10.1045/july2017-ramapriyan

Basic workflow for curating and reporting DQI



Monitoring & Improvement

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

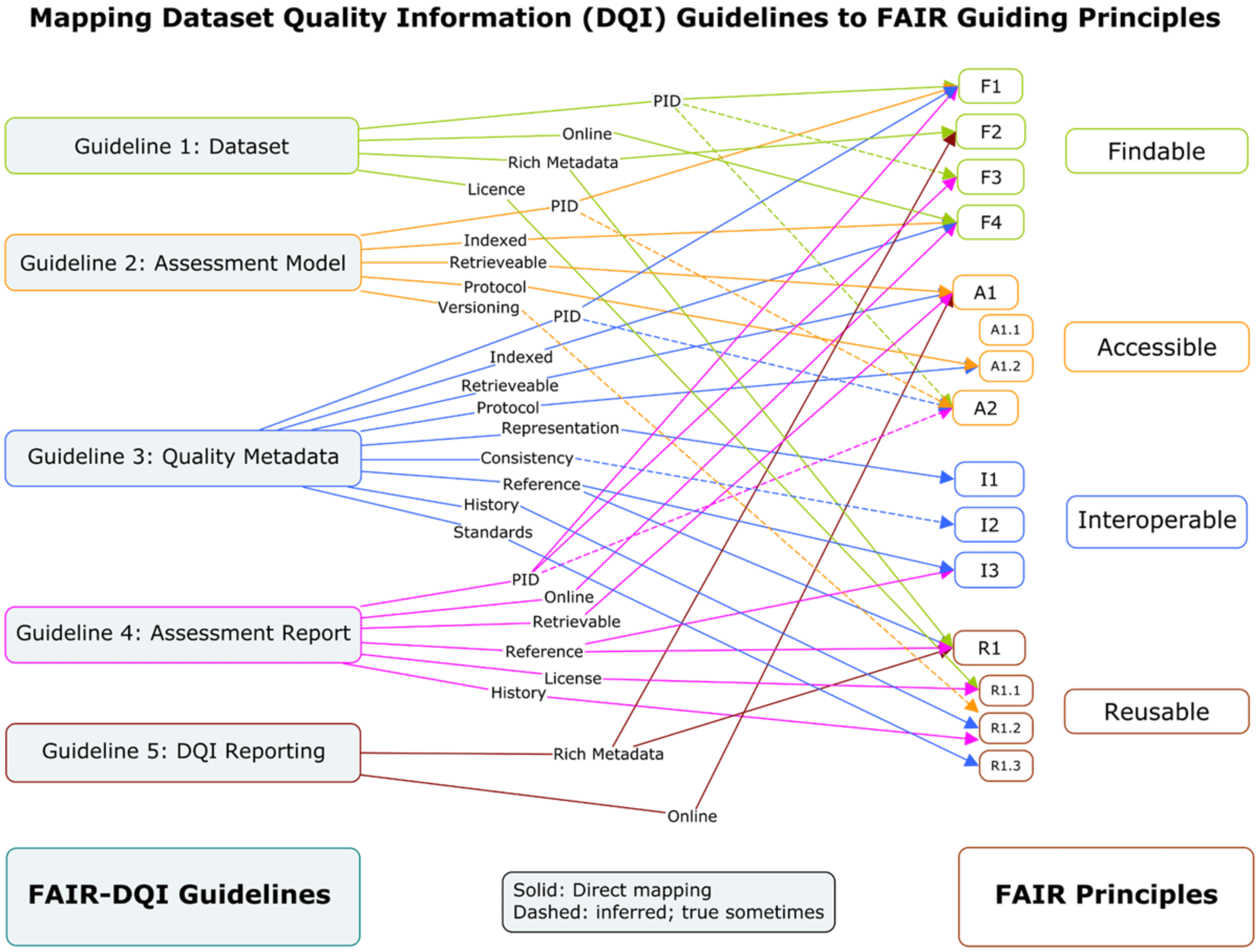
- **Make** the assessment results readily available and usable to stakeholders and collect feedback for improvement

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

FAIR DQI guidelines at a glance

- 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

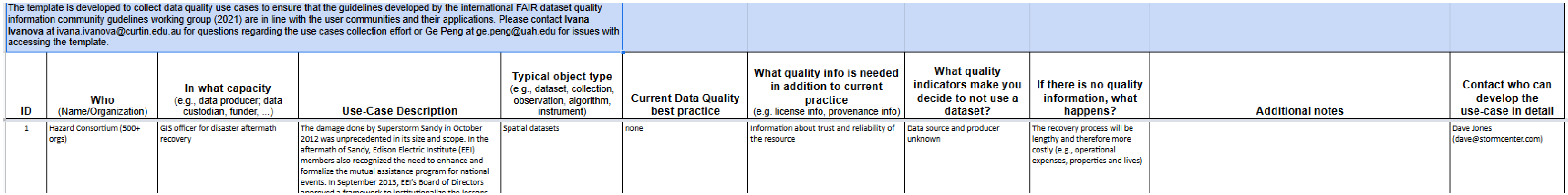
FAIR DQI guidelines are really FAIR



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

FAIR DQI use cases

- ❖ 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;



The template is developed to collect data quality use cases to ensure that the guidelines developed by the international FAIR dataset quality information community gudelines working group (2021) are in line with the user communities and their applications. Please contact **Ivana Ivanova** at ivana.ivanova@curtin.edu.au for questions regarding the use cases collection effort or Ge Peng at ge.peng@uah.edu for issues with accessing the template.

ID	Who (Name/Organization)	In what capacity (e.g., data producer; data custodian, funder, ...)	Use-Case Description	Typical object type (e.g., dataset, collection, observation, algorithm, instrument)	Current Data Quality best practice	What quality info is needed in addition to current practice (e.g. license info, provenance info)	What quality indicators make you decide to not use a dataset?	If there is no quality information, what happens?	Additional notes	Contact who can develop the use-case in detail
1	Hazard Consortium (500+ orgs)	GIS officer for disaster aftermath recovery	The damage done by Superstorm Sandy in October 2012 was unprecedented in its size and scope. In the aftermath of Sandy, Edison Electric Institute (EEI) members also recognized the need to enhance and formalize the mutual assistance program for national events. In September 2013, EEI's Board of Directors approved a framework to institutionalize the lessons	Spatial datasets	none	Information about trust and reliability of the resource	Data source and producer unknown	The recovery process will be lengthy and therefore more costly (e.g., operational expenses, properties and lives)		Dave Jones (dave@stormcenter.com)

- ❖ 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