EOSC Task Force FAIR Metrics and Data Quality

EOSC Symposium

Carlo Lacagnina & Mark Wilkinson

Task Force co-Chairs

EOSC Task Force FAIR Metrics and Data Quality

Data Quality group

Current status

Data Quality Group: What has been done so far

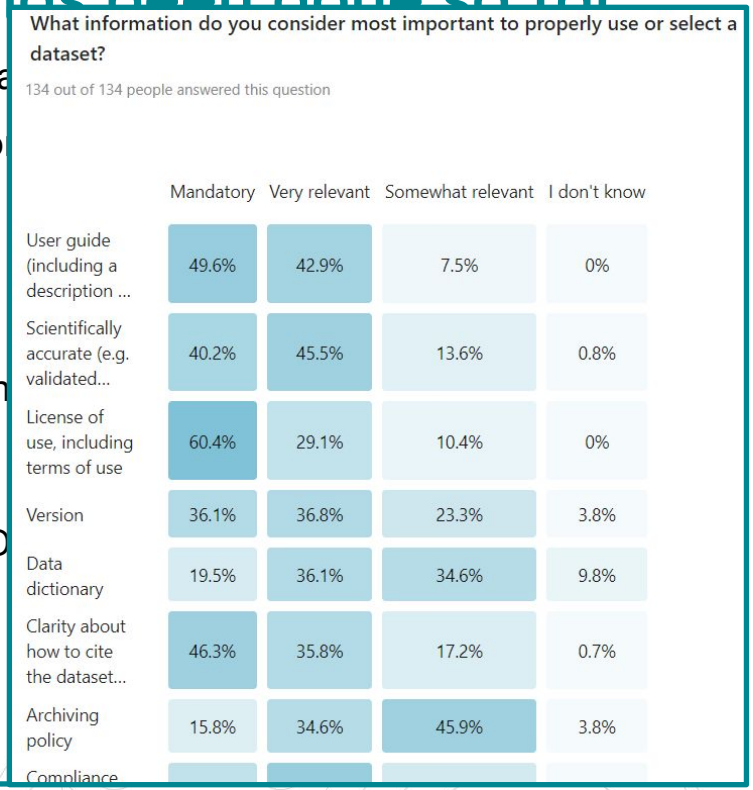
- Pinning down **common ground understanding** about quality approaches, what quality means, dataset lifecycle, actors involved, benefits of quality, workflow for managing quality, data types, certification, etc.
- **Desk research** of ISOs, literature, vocabulary
- Gathering inputs, lessons learned, agreed practices from **various initiatives** (e.g. RDA, INSPIRE, bioimaging, CoreTrustSeal, energy sector)
- Drafting a **recommendation document** – 1st version in December 2022
- **RDA session** organized in June 2022
- Drafted a **survey** released in April: >700 views

Data Quality Group: What has been done so far

- Pinning down **common**
  lifecycle, actors involve
  etc.
- **Desk research** of ISOs,
- Gathering inputs, lesson
  bioimaging, CoreTrustS
- Drafting a **recommenda**
- **RDA session** organized
- Drafted a **survey** released in April: >700 views

Data Quality Group: What has been done so far

- Pinning down **common ground understanding** about qua
  lifecycle, actors involved, benefits of quality, workflow fo
  etc.
- **Desk research** of ISOs, literature, vocabulary
- Gathering inputs, lessons learned, agreed practices from
  bioimaging, CoreTrustSeal, energy sector)
- Drafting a **recommendation document** – 1st version in D
- **RDA session** organized in June 2022
- Drafted a **survey** released in April: >700 views



What information do you consider most important to properly use or select a dataset?

134 out of 134 people answered this question

	Mandatory	Very relevant	Somewhat relevant	I don't know
User guide (including a description ...	49.6%	42.9%	7.5%	0%
Scientifically accurate (e.g. validated...	40.2%	45.5%	13.6%	0.8%
License of use, including terms of use	60.4%	29.1%	10.4%	0%
Version	36.1%	36.8%	23.3%	3.8%
Data dictionary	19.5%	36.1%	34.6%	9.8%
Clarity about how to cite the dataset...	46.3%	35.8%	17.2%	0.7%
Archiving policy	15.8%	34.6%	45.9%	3.8%
Compliance				

Survey: some insights

Biggest concern/barrier to provide quality assessed data:

Which practices should a discipline have to gauge its maturity in quality management?

- Metadata standards, agreed definitions, standard quality management framework, metrics to quantify quality, quality assessments are operational routine and funded

What level of data quality management do you expect from EOSC?

- Basic curation: e.g., data content sanity checks, control availability of basic metadata or documentation, basic metadata compliance checks. Allow (re)users to rate or leave comments on data quality

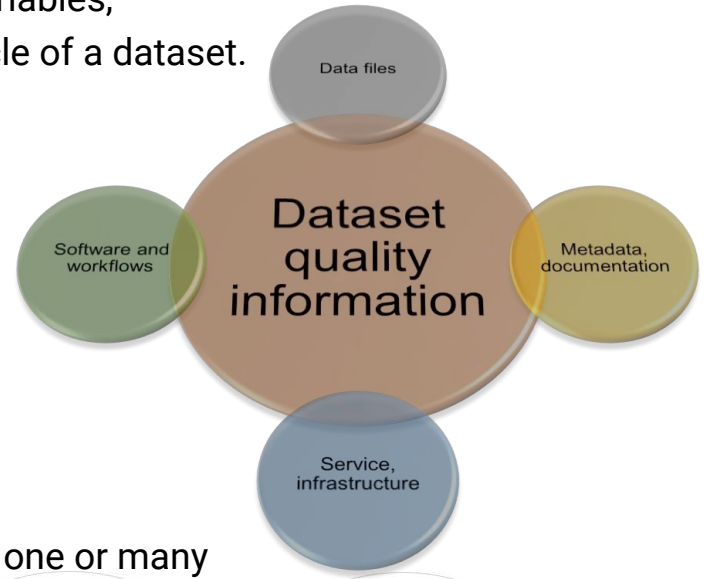
- **Some conclusions**
  - well advertisement that **quality does not refer to data content quality only**, a.k.a. scientific quality.
  - Striking **preference for no ranking**. If a ranking has to be applied, then priority should be placed on **showing the FAIRness level** of the datasets. **No data content assessment** is expected from EOSC, but check of documentation availability for data understanding.
  - The future quality assessments should be shown first to the **data provider**, to give a chance to **improve the data**, and then to the users. The methodology has to be the same for similar datasets.
  - Create a catalogue of community tests/methods to apply in quality analyses.
  - EOSC users expect tools and services being designed according to a user-centric model.

Dataset quality, not just data quality

Dataset quality information describes issues with instruments, variables, measurement, collection, access, use through the entire lifecycle of a dataset. It's about:

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

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).

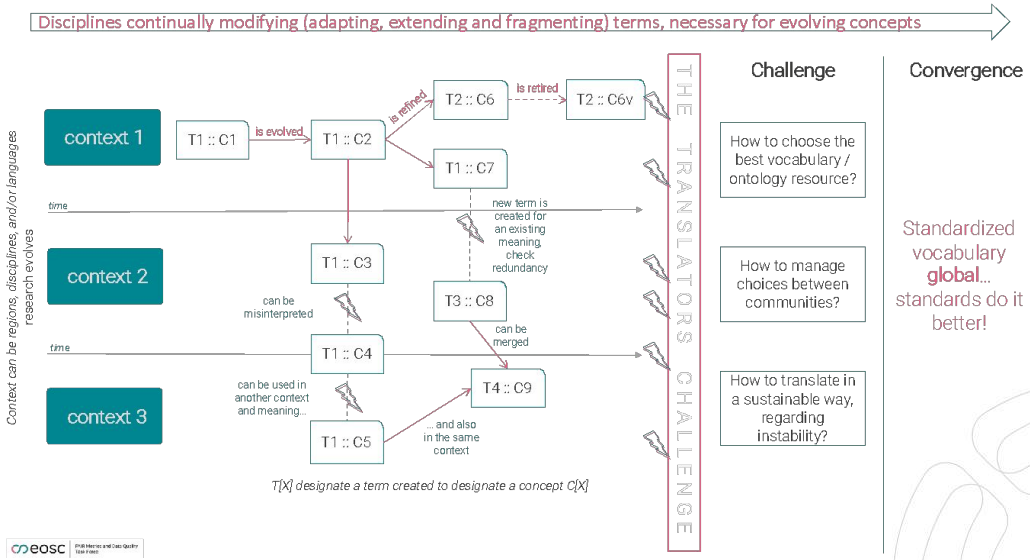


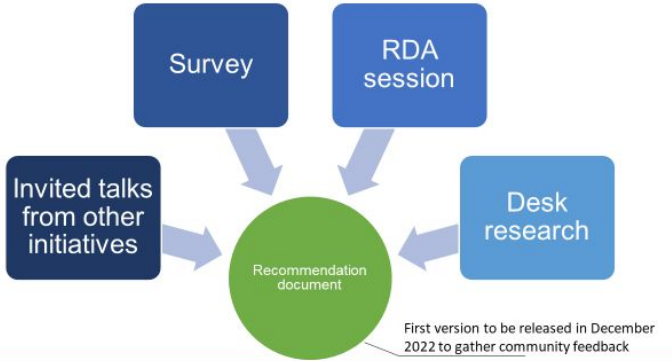
Why do we need quality information?

- Decision-making
  - Data use: Informing the reliability and usability of the dataset,
  - Data trust: Establishing the trust between data providers and consumers, policy-makers,
  - Influential data: Increase the value of the data for diverse users.
- Compliance reporting support
  - Consistently curated,
  - Readily available and understood by humans and machines,
  - Augmented understandability and clarity of data.
- Support data and information, sharing and reuse
  - Maximize the sharing of dataset quality information,
  - Interoperable dataset quality information also for utilizing Cloud and Machine Learning technologies,
  - Promote global access and harmonization of quality information.

Multidisciplinary understanding about data quality

- Each discipline is **unique** but may face similar **needs** and **challenges**
- **Common interest** in learning/sharing knowledge & best practices across disciplines

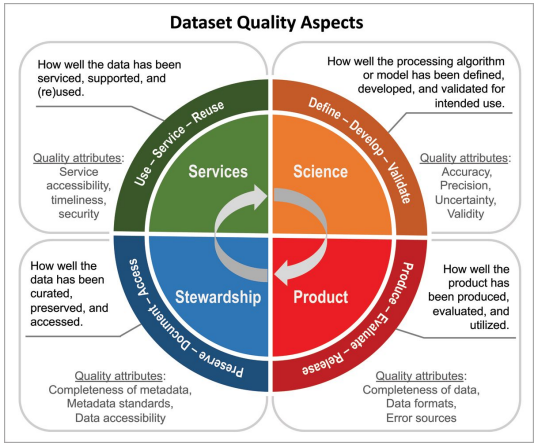




modified after R. David 2022, standardized /controlled / vocabulary evolution

Recommendation document

- Recommendations are a set of principles and guidelines for both EOSC and the next TF:
  - Datasets have to come with enough **contextualization** information to understand and correctly interpret them
  - EOSC is not in charge of **data content** assessments
  - Set clear **criteria** to prevent researchers concerns about how professionally their data will be managed, concerns are barriers to data sharing
  - Develop a **pre-operational quality function** tailored to the EOSC stakeholders' requirements
  - EOSC should support and push each community to agree on **community standards**, which form the basis for any quality assessment and FAIR sharing of research datasets
  - We have already identified **minimum requirements**; the next TF will need to identify the exact standards forming the baselines for these requirements assessment



Peng et al. (2021)

Thank you!

presented by Chris Schubert
University of Technology Vienna, Library
Head of Media Management & Library-IT
TF member;
Chair of GEO (Group on Eeath Observation) Data Sharing & Data Management Principles, SG of Data WG;
ISO TC211, Austrian Standards Member;
chris.schubert@tuwien.ac.at