eosc

FAIR ENABLING PRACTICES

Data Quality in Open Science and Cross-Disciplinarity Perspective

Chris Schubert
University of Technology Vienna
Head of Media Management & Library-IT

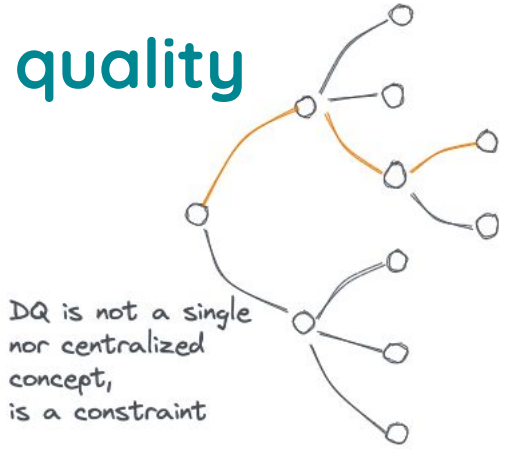
EOSC Task Force FAIR Metrics and Data Quality

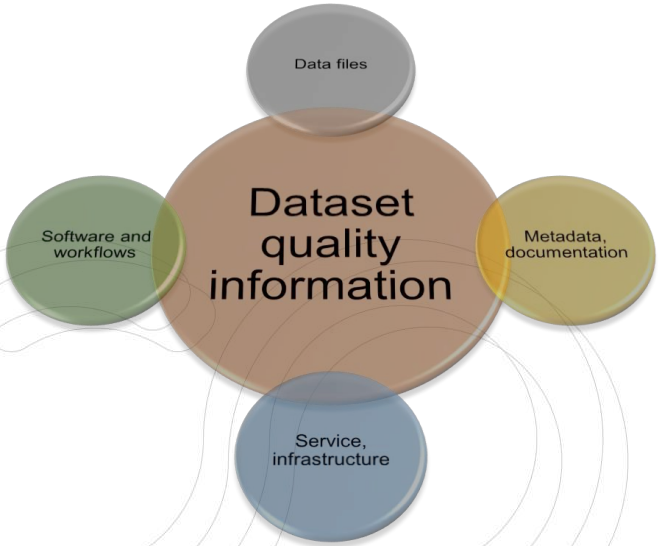
Dataset quality, not just data quality

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?

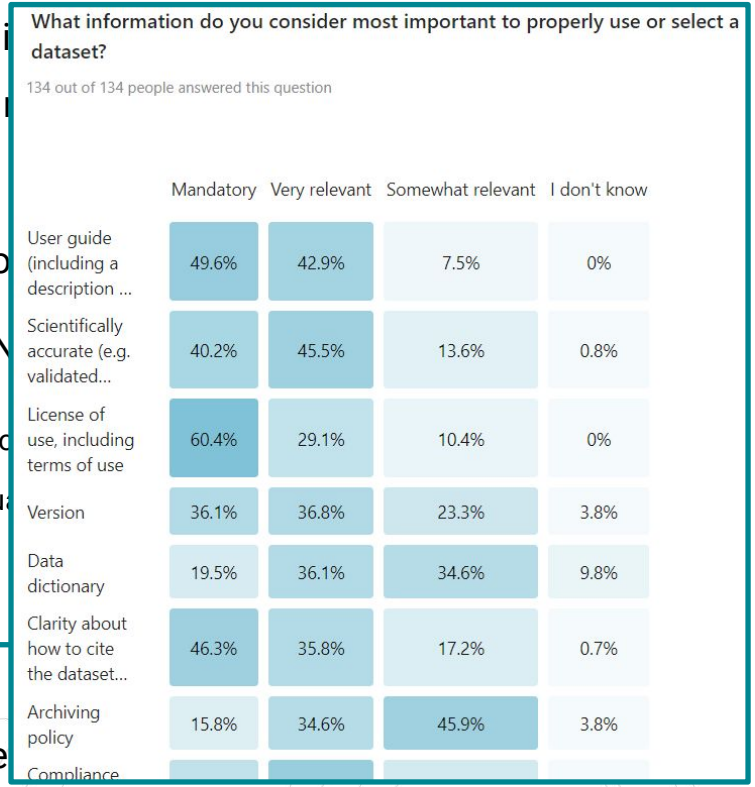
- Decision-making
  - Data use: Informing the reliability and usability of the dataset,
  - Data trust: Establishing the trust between data providers and consumers, policy-makers,
  - Influential data: Increase the value of the data for diverse users.
- Compliance reporting support
  - Consistently curated,
  - 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.

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, common semantic & crosswalks
- Gathering lessons learned, from **various initiatives** (e.g. INSPIRE, CoreTrustSeal, Energy Sector)
- Joint effort: International FAIR-DQI Community Guidelines Working Group, Information Quality Cluster of U.S. Earth Science Information Partners (ESIP IQC), Barcelona Supercomputing Center Evaluation and Quality Control (BSC EQC)
- Drafted a **survey** released in April: >700 views
- **RDA session** organized in June 2022
- Drafting a **recommendation document** – 1st version in December 2022

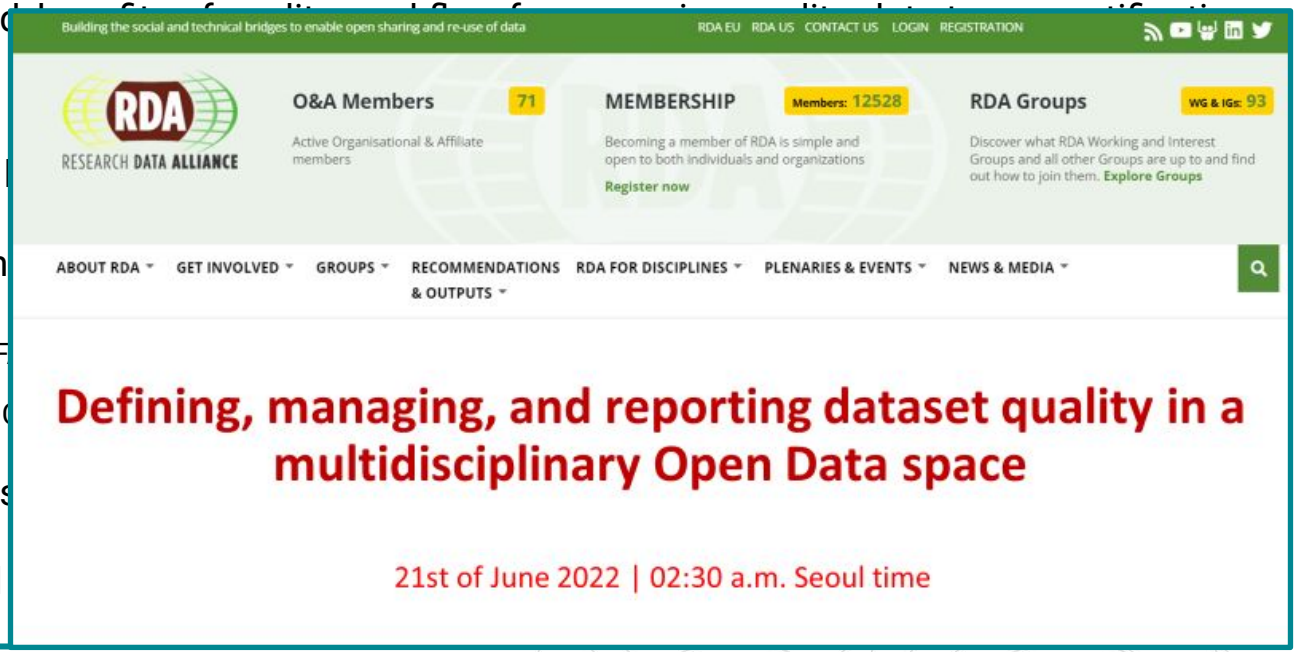
Data Quality Group: What has been done so far

- Pinning down **common ground understanding** about quali
  lifecycle, actors involved, benefits of quality, workflow for
  etc.
- **Desk research** of ISOs, literature, common semantic & cro
- Gathering lessons learned, from **various initiatives** (e.g. IN
- Joint effort: International FAIR-DQI Community Guidelines Working Gro
  Information Partners (ESIP IQC), Barcelona Supercomputing Center Evalu
- Drafted a **survey** released in April: >700 views
- **RDA session** organized in June 2022
- Drafting a **recommendation document** – 1st version in De



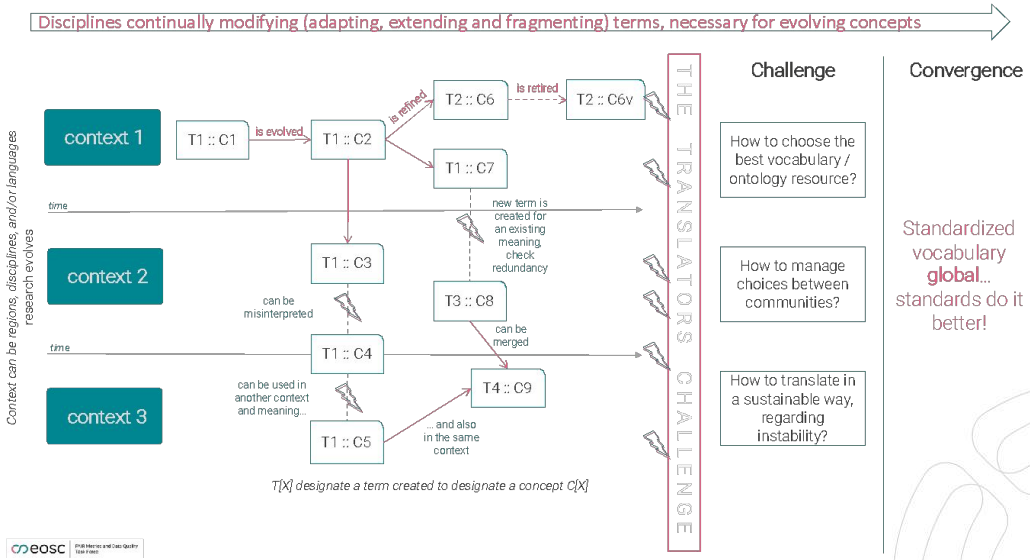
Data Quality Group: What has been done so far

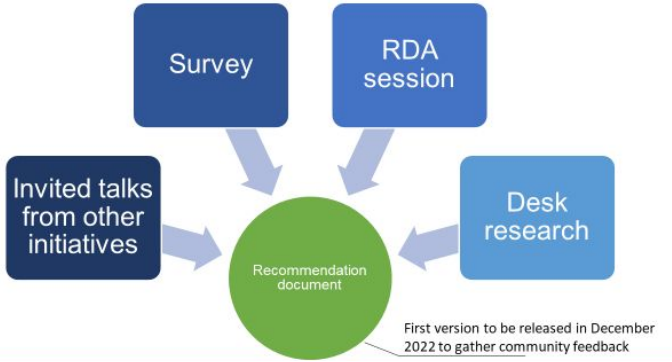
- Pinning down **common ground understanding** about quality approaches, what quality means, dataset lifecycle, actors involved
- **Desk research** of ISOs,
- Gathering lessons learn
- Joint effort: International F
- Drafted a **survey** releas
- **RDA session** organized
- Drafting a **recommendation document** – 1st version in December 2022



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

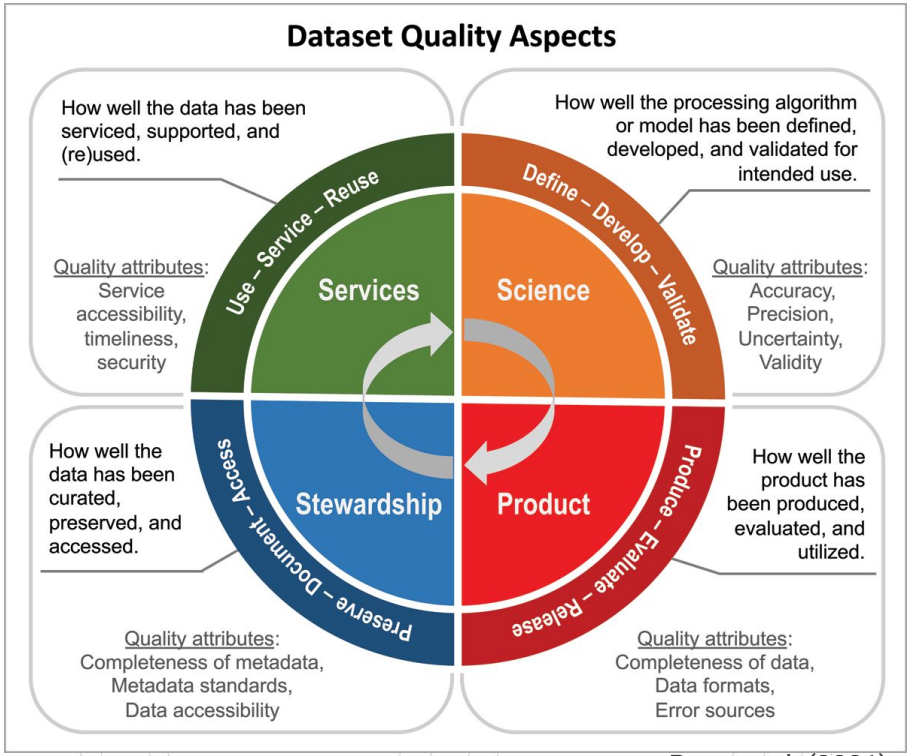




modified after R. David 2022, standardized /controlled / vocabulary evolution

Scientific Domains

- Use Cases & Area of Interest, TF FAIR Metrics & DQ gathered:
  - Climate Service
  - Climate Models & Research
  - Digital Collections for Humanities
  - Genomic Data
  - Spatial Data (INSPIRE –EU Directive)
  - Text Mining & Subject Indexing
  - Linguistic Research and Language Models
  - SI units, recognition of measurement (Metrology)
  - Biodiversity Data
  - Social Media Data Research
  - low carbon energy research



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 Earth Observation) Data Sharing & Data Management Principles, SG of Data WG;
ISO TC211, Austrian Standards Member;
chris.schubert@tuwien.ac.at