EOSC Data Quality Framework

”Integrity, Provenance, and Trust”

Chris Schubert

TU WIEN TECHNISCHE UNIVERSITÄT WIEN

eosc

Open Science Day | 3rd of June 2025

EOSC Data Quality Framework

Chris Schubert
TU Wien, University of Technology Vienna,
Head of Media Management & Library-IT
EULiST local Coordinator at TU Wien (European University Alliance)
EOSC Task Force Co-Chair FAIR Metrics & Data Quality (2022 – 2024)
EOSC Task Force Member FAIR Metrics & Digital Objects (2024 – 2026)
EOSC OA3: FAIR Assessment and Alignment (2024 - …)
Member of GEO (Group on Earth Observation) Data Sharing & Data Management Principles
ISO TC211 – Terminology AG, Austrian Standards Member
CODATA Austria National Member, Chair

chris.schubert@tuwien.ac.at

TU WIEN Bibliothek Your knowledge hub

eosc FAIR Metrics and Data Quality Task Force

CODATA Committee on Data International Science Council

FAIR Metrics and Digital Objects Task Force

MDO

GEO Group on Earth Observations

ISO

RDA Research Data Alliance

DATA QUALITY - AN UNDERESTIMATED ISSUE[1]?

There is a need to address the challenges relating to the quality of research data:
- Data Quality as a burden or accelerator in trustworthy data productivity
- Is there a lack of high-quality data? In which domains? for AI training representatives, NLP, LLM, etc.?
- Explicit statements on Data Quality are extremely challenging – especially across interdisciplinary reuse.
- Impact on trusted data sources and providers (governance structures, costs, inclusion, etc.)

But what exactly is data quality in a scientific context?

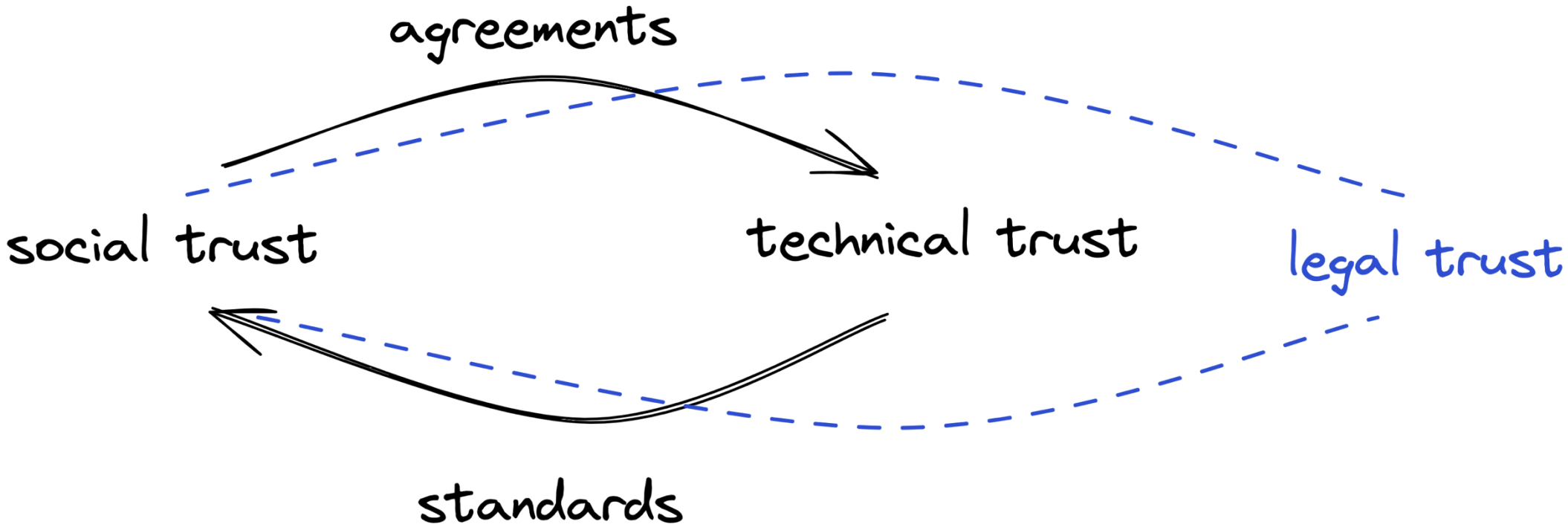
FAIR Data = Data Quality ? (…FAIR Data Package)

What motivates data providers to ensure FAIRness and data quality in their research data?
- assistance vs scoring

Do we need more guidelines and compliance rules for funding criteria, good scientific practice or other incentives?

[1] The Data Quality Challenge – February 2020- Recommendations for Sustainable Research in the Digital Turn, Rat für Informations Infrastrukturen 2021, https://rfii.de/download/herausforderung-datenqualitaet-november-2019/

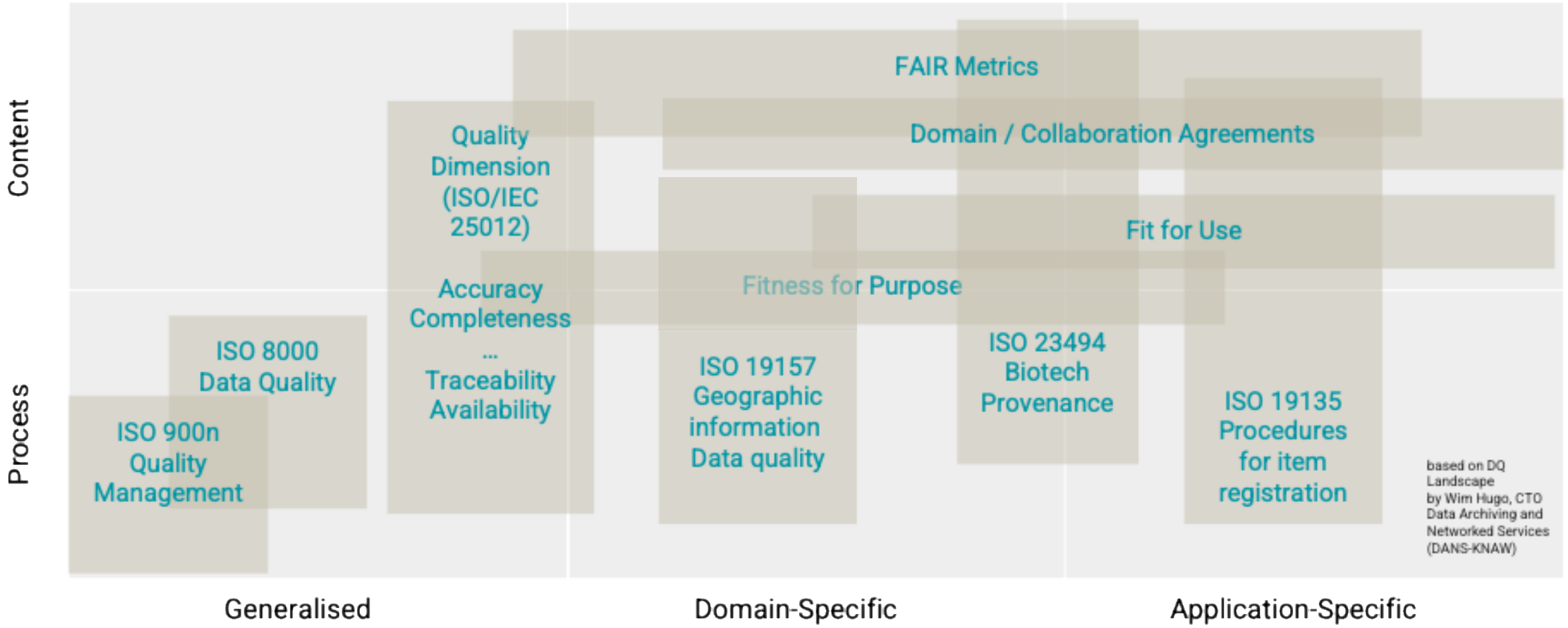
Trusted Data Environment



based on Mark A. Parsons, Quality Management In Open Science, SciDataCon Session 2023

The Landscape of Data Quality Implementation

„De jure“ und „de facto“ - Standards



AI Awareness - Data Literacy - Data Quality

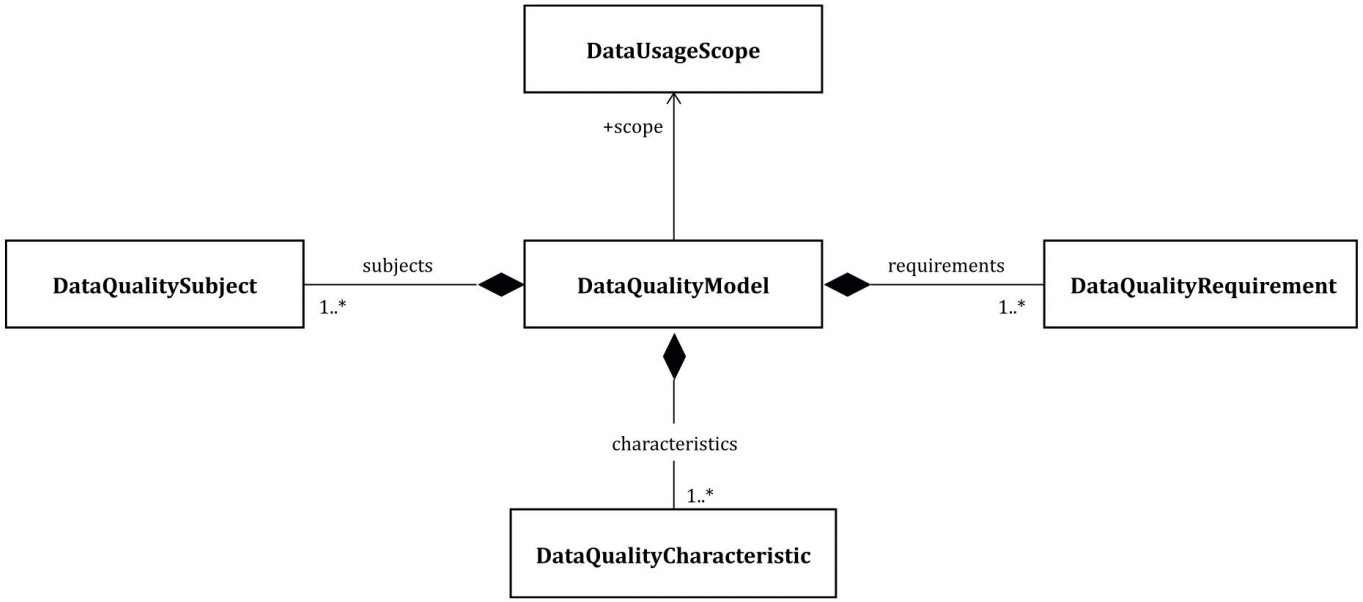


Figure 2 — Data quality model

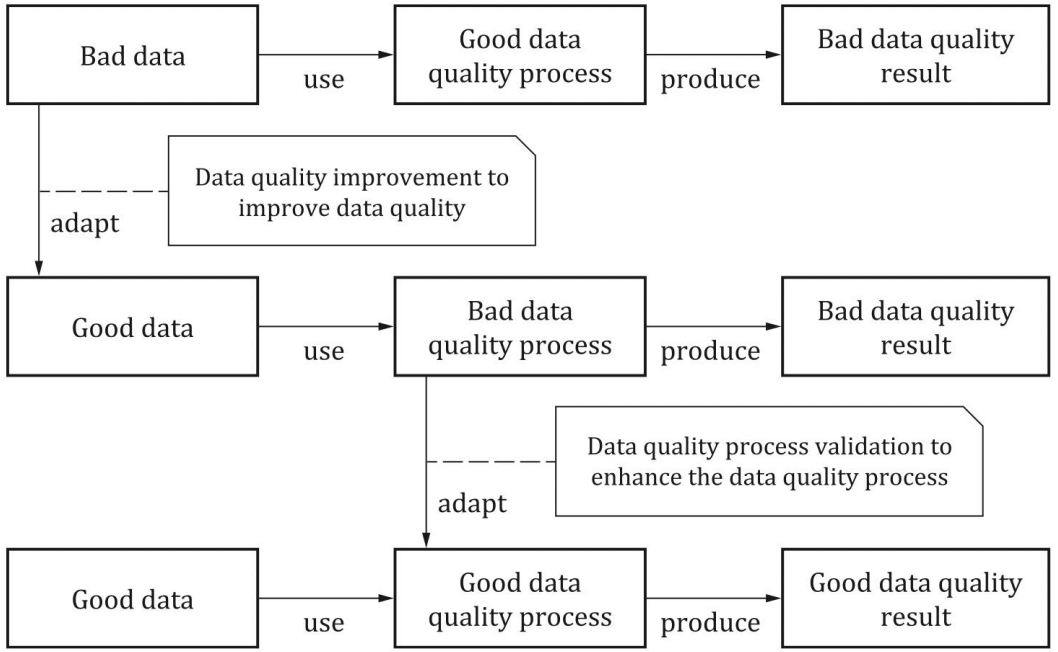
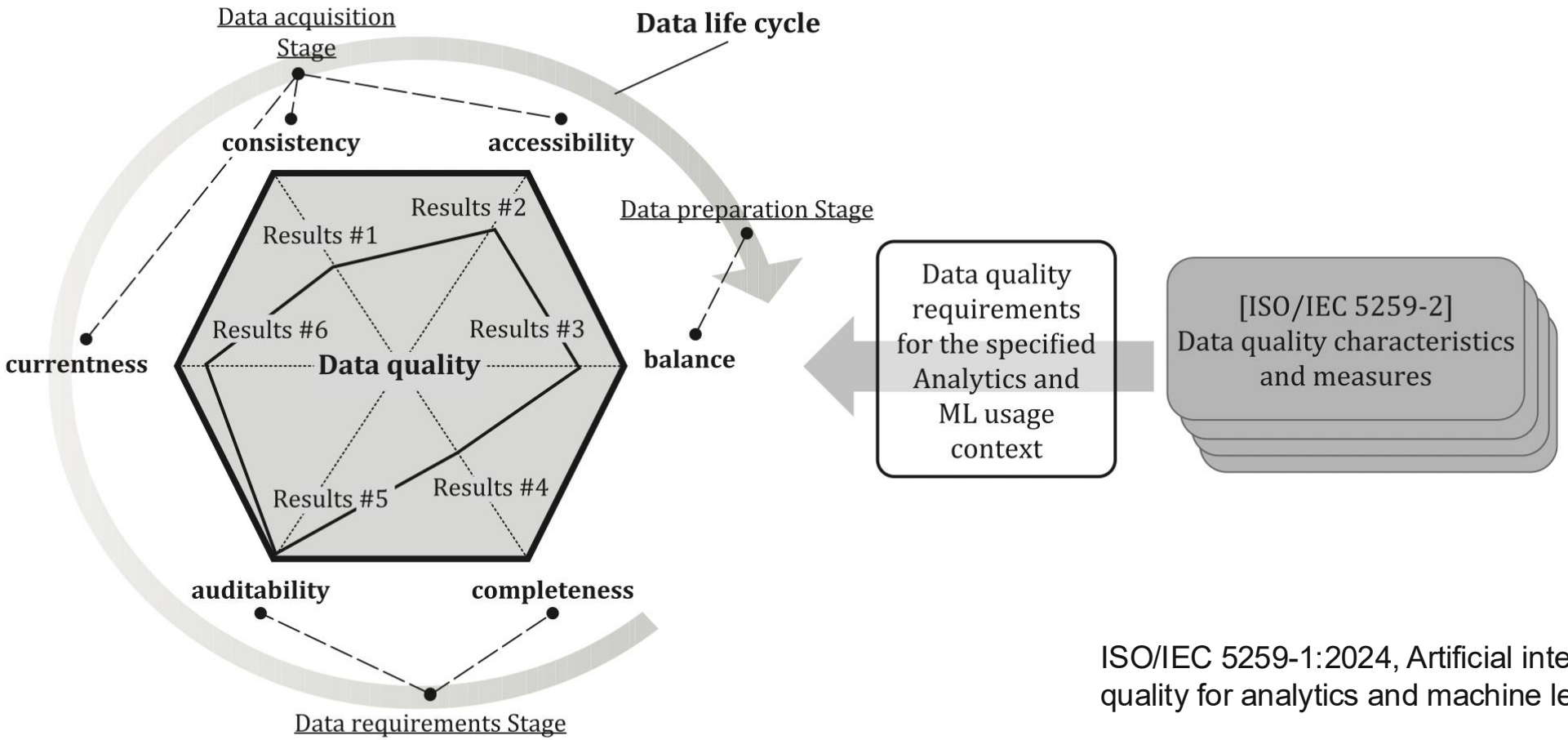


Figure 3 — Relationship between data quality and data quality processes

ISO/IEC 5259-4:2024, Artificial intelligence — Data quality for analytics and machine learning (ML)

AI Awareness - Data Literacy - Data Quality



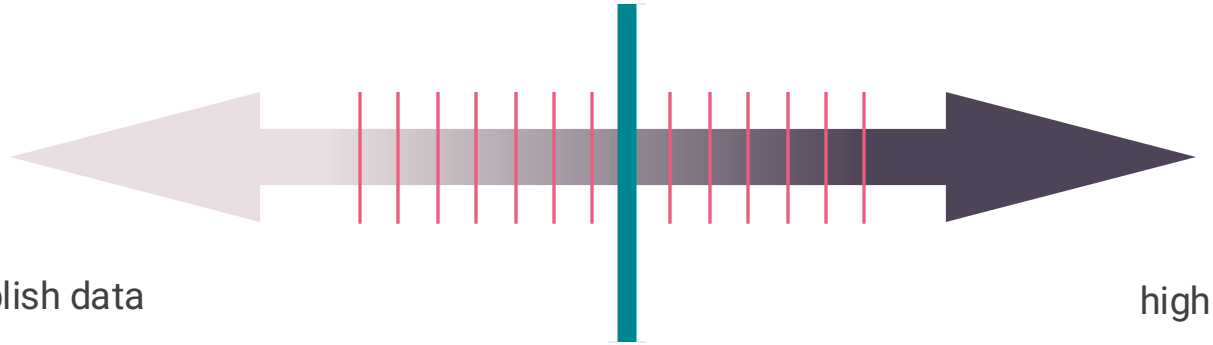
<A data quality model for analytics and ML>

ISO/IEC 5259-1:2024, Artificial intelligence — Data quality for analytics and machine learning (ML)

Data Quality in EOSC

data quality dilemma

for research data infrastructures



No a priori DQ-Requirements for inklusion

- no obstacles to publish data
- misleading outcomes
- bilateral trust, or massive investigation

Stringent a priori DQ-Requirements

- high barrier to publish data
- less (open) data available
- Governance, monitoring, Certification

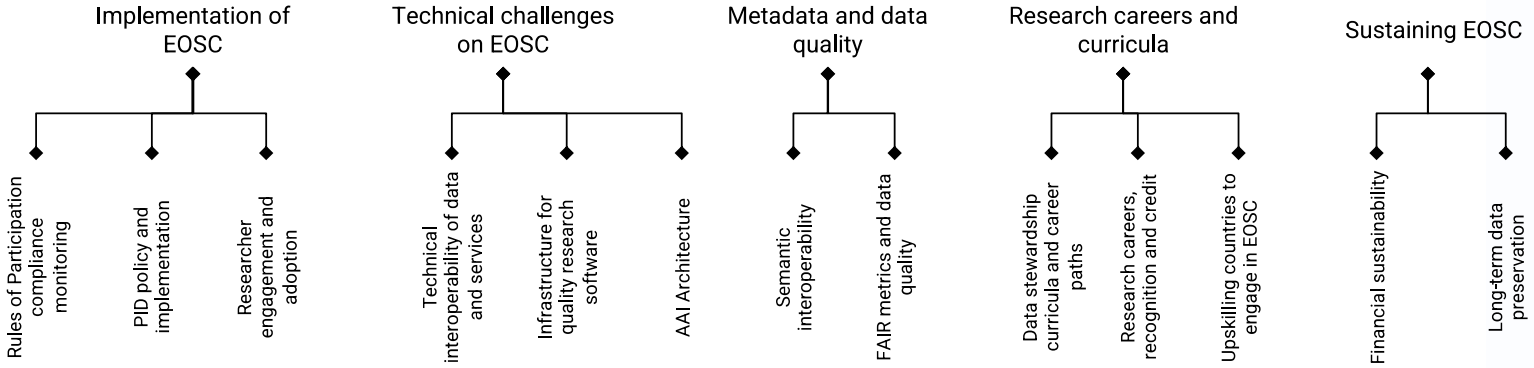
Taking a generic approach to data quality assessment can lead to arbitrariness

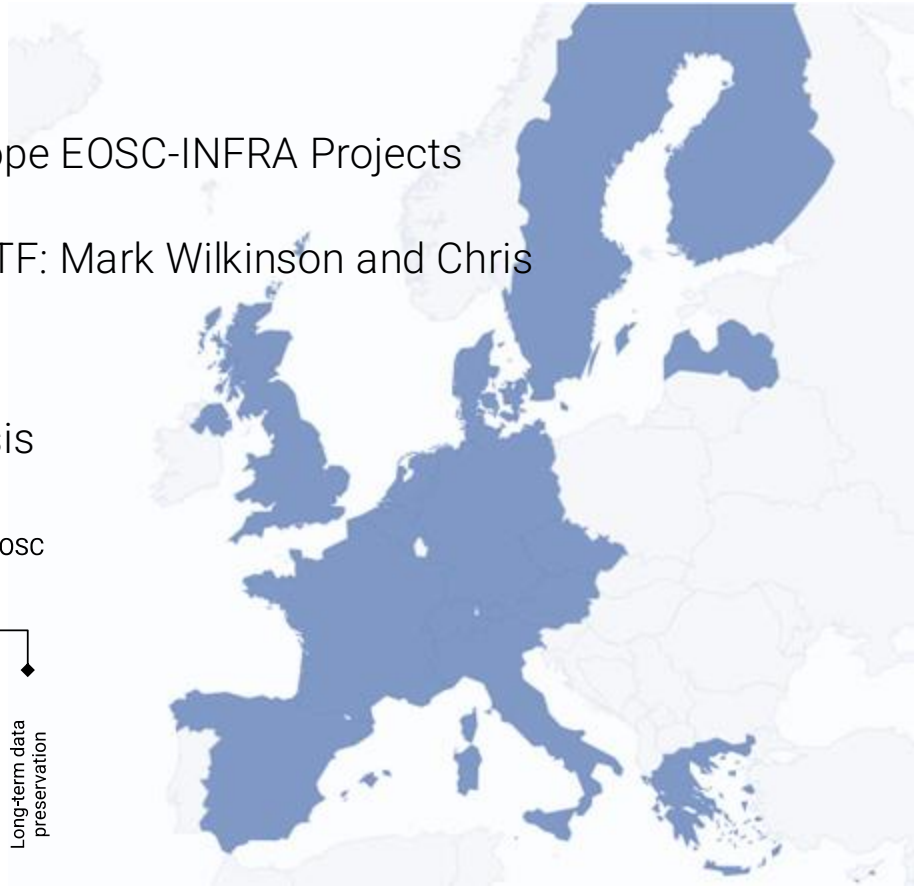
Domain/community standards + context + provenance Interaction is required

e.g. Metadata annotation as a formalised interoperable description (through AI?)

Data Quality in EOSC, Task Force FAIR Metrics and Data Quality

- 26 experts from 17 EU Staates from different domains
- The EOSC Task Forces contribute to the strategies of the EOSC Association
- Support the EOSC Association and are expected to interact with the Horizon Europe EOSC-INFRA Projects
- Started in 2021, ended in May/June 2024. Two co-chairs coordinated this EOSC TF: Mark Wilkinson and Chris Schubert
- At least biweekly work meetings over a period of two years, all on a voluntary basis





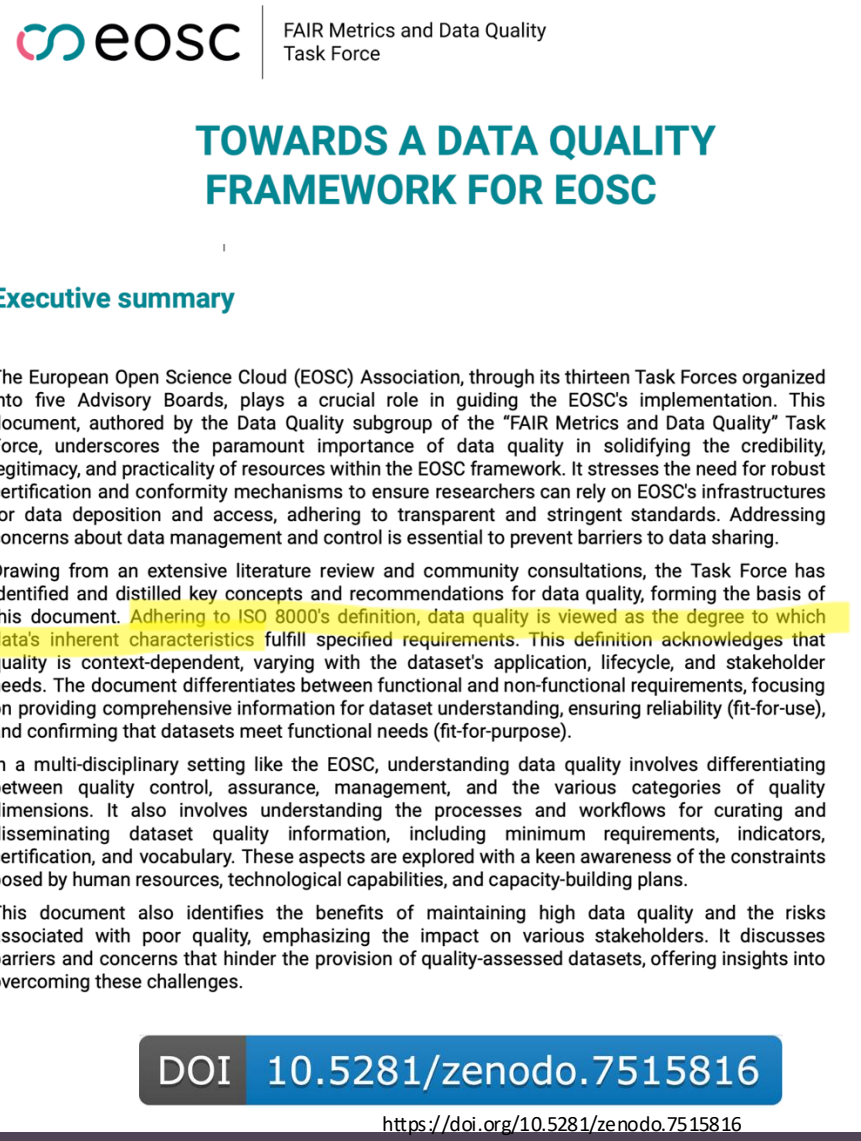
Data Quality in EOSC

our approach…

- State-of-the-art identification for a common understanding of data quality
- clear data quality definition - its use is very vague
- Focus on use cases to identify common elements, such as dimensions and indicators, for a framework document
- Motivation for quality management:

  Source Declaration | Unambiguity | Uncertainty | BIAS | Provenance

- Extraction of explicite information for common understandanding to reproducibility, training of algorithms and representatives for AI, but also validation protocols, certification, accreditation (DQ seal)?
- Raising awareness of data quality management as a unique selling point for the EOSC



eosc | FAIR Metrics and Data Quality Task Force

TOWARDS A DATA QUALITY FRAMEWORK FOR EOSC

Executive summary

The European Open Science Cloud (EOSC) Association, through its thirteen Task Forces organized into five Advisory Boards, plays a crucial role in guiding the EOSC's implementation. This document, authored by the Data Quality subgroup of the "FAIR Metrics and Data Quality" Task Force, underscores the paramount importance of data quality in solidifying the credibility, legitimacy, and practicality of resources within the EOSC framework. It stresses the need for robust certification and conformity mechanisms to ensure researchers can rely on EOSC's infrastructures for data deposition and access, adhering to transparent and stringent standards. Addressing concerns about data management and control is essential to prevent barriers to data sharing.

Drawing from an extensive literature review and community consultations, the Task Force has identified and distilled key concepts and recommendations for data quality, forming the basis of this document. Adhering to ISO 8000's definition, data quality is viewed as the degree to which data's inherent characteristics fulfill specified requirements. This definition acknowledges that quality is context-dependent, varying with the dataset's application, lifecycle, and stakeholder needs. The document differentiates between functional and non-functional requirements, focusing on providing comprehensive information for dataset understanding, ensuring reliability (fit-for-use), and confirming that datasets meet functional needs (fit-for-purpose).

In a multi-disciplinary setting like the EOSC, understanding data quality involves differentiating between quality control, assurance, management, and the various categories of quality dimensions. It also involves understanding the processes and workflows for curating and disseminating dataset quality information, including minimum requirements, indicators, certification, and vocabulary. These aspects are explored with a keen awareness of the constraints posed by human resources, technological capabilities, and capacity-building plans.

This document also identifies the benefits of maintaining high data quality and the risks associated with poor quality, emphasizing the impact on various stakeholders. It discusses barriers and concerns that hinder the provision of quality-assessed datasets, offering insights into overcoming these challenges.

DOI 10.5281/zenodo.7515816

https://doi.org/10.5281/zenodo.7515816

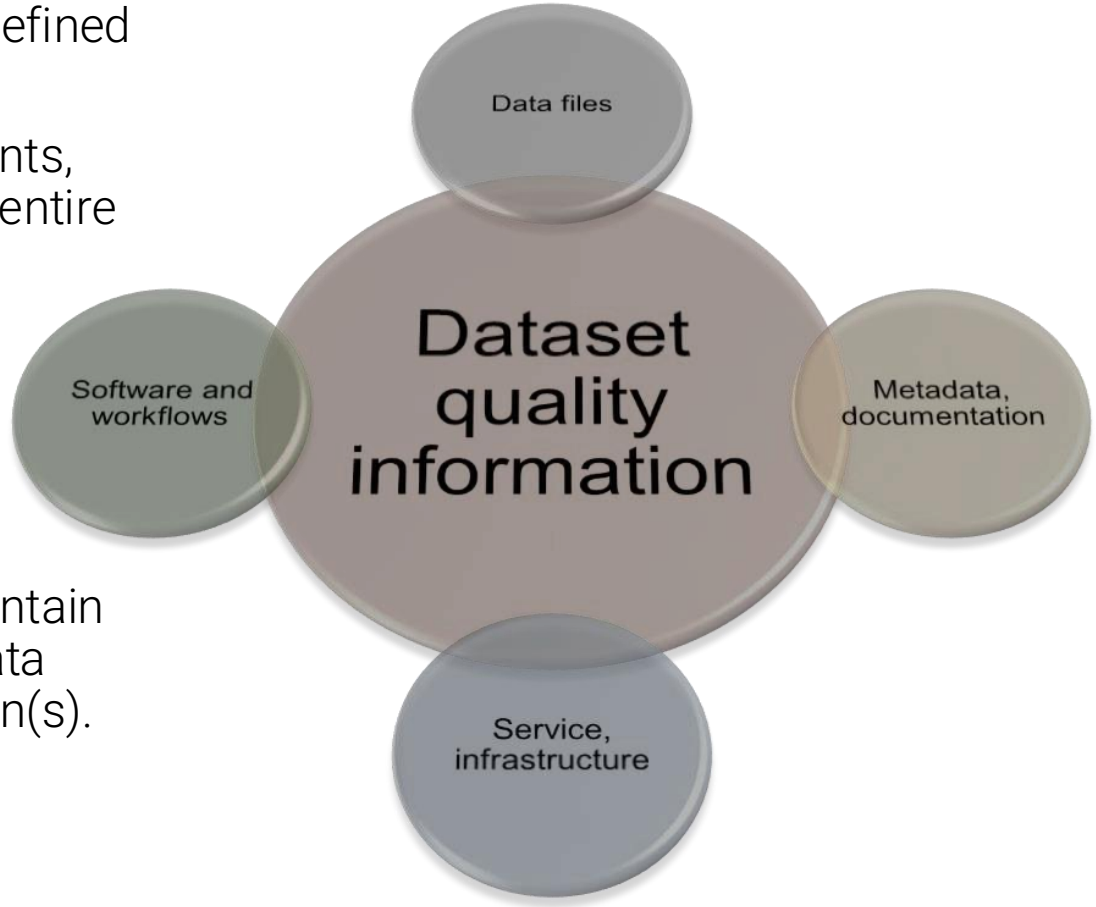
Dataset quality, not just data quality

"Data Quality" is commonly interpreted as "the degree to which a set of inherent characteristics of data fulfills requirements," as defined by ISO 8000.

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

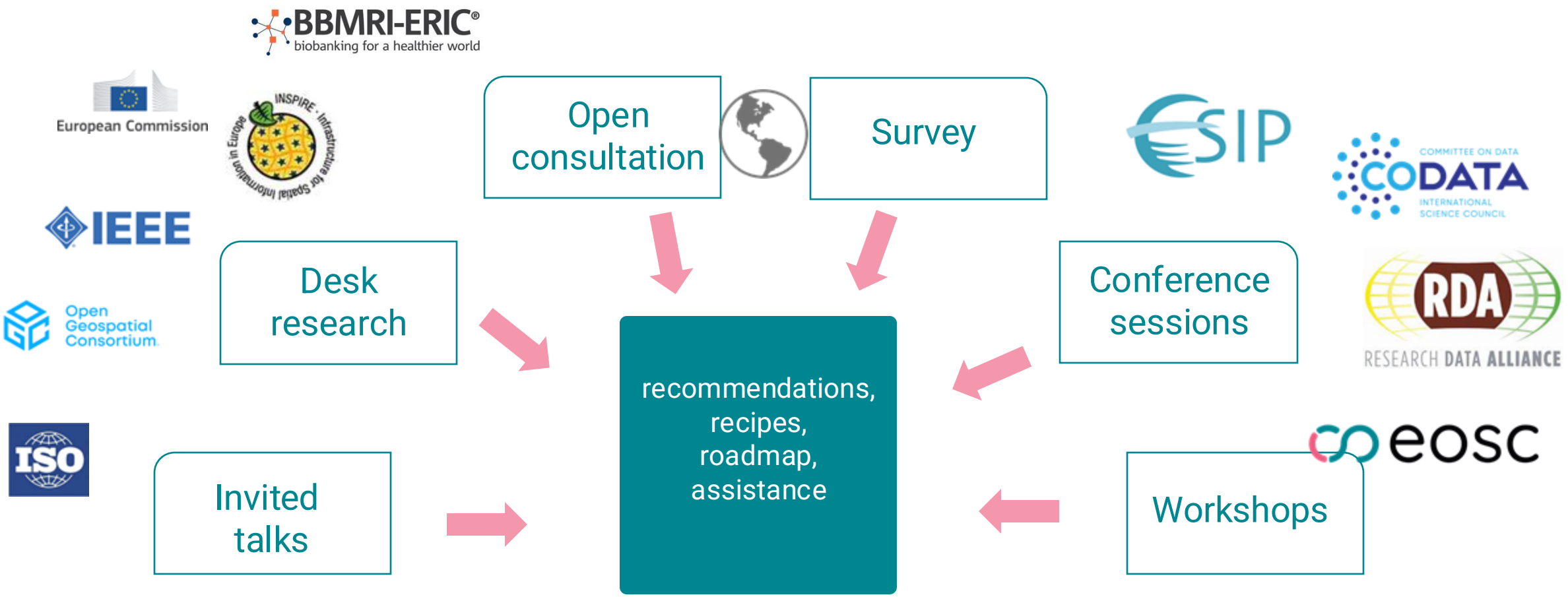
- Quality of data (input and output),
- Quality of metadata and documentation,
- Quality of software and workflows,
- Quality of procedures and processes,
- Quality of infrastructure, tools, and systems.

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).



> trustworthiness and transparency
> dataset quality by design

Community involvement for EOSC Data Quality Framework



Recommendations for EOSC Data Quality Framework

Quality information must be objective, transparent, consistent, and communicated effectively.

Serve formalised DQ information, published as registered item

Federated systems with essential core components like the operational use of semantic artefacts (controlled vocabulary, data-model-featureTypes, ontology, …)

Data Quality as topic for the entire data life cycle – not just as last mile

Data quality assessment needs common rules, EOSC should consider taking the opportunity to encourage communities to reach a consensus in using their standards.

Improving data quality as close to the source (i.e., producer or provider) as possible is highly recommended

			Tailored recommendations for different stakeholder categories					
			Practicioners		***Custodians***		***Decision & Policy makers***	
			... corresponds to those directly working on research, including		*... corresponds to stakeholders that will support FAIR in practice,*		*... encompasses stakeholders that will require access to*	
Recommenation Nr.	**General recommendation**	**Description**	***Data producer***	***Data consumer***	***Service providers & infrastructures***	***Support staff (e.g. data stewards)***	***Community-specific bodies***	***EOSC Association and other policy-makers***
1	Standardization and Governance	A community-specific standards and a governance structure for data quality should be developed and adopted, to ensure the exchange of high-quality and FAIR research data. Utilize platforms like FAIRsharing to monitor and encourage community adoption of these standards.	To facilitate the reuse of high-quality data, comply with quality standards and provide structured documentation and metadata to facilitate quality assessments.	To improve your data discovery, understanding, and reuse, rely on standardized data formats and documentation practices within your research community.	To maintain and provide information of capacities, like validation tools or information, to adherence a catalogue of supported and robust standards, compliance and improve data service quality. A clear information should be given on directives, available technical or	To facilitate the reuse of high-quality data, advocate for the adoption of community-specific standards and a governance structure. Promote platforms like FAIRsharing to monitor and encourage community adoption of these standards.	To foster and advocate the development and maintenance of community-specific standards, initiate and lead efforts to oversee these standards. Advocate for sufficient funding for this topic. To foster the development and maintenance of community-	To support work of practioners & custodians initiate the development of community-specific standards and governance structures, including legal frameworks and community-driven models. Advocate and promote the use of community-specific
		Data quality should be actively managed as an integral aspect of the full research data			To provide a catalogue of existing and supportive	Establish clear data quality criteria and encourage the	Ensure that quality is given high priority in the area of	~~Establish a quality management function to oversee the~~

ISO/IEC 30435:2023, Human Resource management – Data Quality

EOSC Strategic Roadmap

recent outcome of EOSC Winter School 2025, January in Seville

Strategic Pillars of MAR 2026-2027	Key outcomes and recommendations
1 **Sustaining and enhancing the EOSC Federation**	Discussions on Strategic Pillar 1 emphasised the need for **robust governance** and proposed setting up an authoritative body within the governance structure to be responsible for PIDs. Participants reiterated the need for **harmonised definitions and metadata standards** to achieve seamless semantic and technical interoperability and started identifying existing solutions within scientific domains. A **shared technical architecture model, common vocabulary, and open communication** among stakeholders as represented by the EOSC Federation Handbook were identified as crucial steps in fostering engagement, adoption, and long-term impact of the EOSC Federation. During the closing session the need to **promote existing working solutions** to avoid raising expectations was highlighted.
2 **Contributing to the web of FAIR data and the uptake of AI**	The exchanges on critical topics related to Strategic Pillar 2 centred on the intersection of FAIR principles and AI, emphasising the importance of **robust metadata, data quality and open source software for AI applications.** Participants distinguished between FAIR-for-AI and AI-for-FAIR data, underscoring the **need for standards to harmonise FAIR and AI practices**, along with efforts to integrate publications, research data, and metadata for improved resource discovery and usability within the EOSC Federation.
3 **Ensuring research security and sovereignty**	While assessing research security and sovereignty, Winter School participants acknowledged the **lack of high-quality data for training AI models.** They stressed the importance of defining responsibility and liability for the disclosure of sensitive data, suggesting that a general requirement is reaching an agreement on a **common model for metrics, tests and associated benchmarks** across FAIR tools that defines an authority for selecting metrics and community benchmarks.
4 **Linking with other Common European Data Spaces and beyond**	The attendees concluded that use cases are needed to explore the integration of Data Spaces within the EOSC Federation and vice-versa, and to define their interoperability. They argued that the usability of the EOSC EU Node should be assessed through **VREs for real use cases,** fostering existing communities to promote Open Science practices. The need for collaborations to establish trusted VREs between projects was highlighted. They also suggested to **engage with various Common European Data Spaces** and extend the involvement to include **EU Missions.**

EOSC Opportunity Areas & Task Forces

- OA1: PIDs
- OA2: Metadata, Ontologies & Interoperability
- OA3: FAIR Assessment & Alignment
- OA4: User & Resource Environments
- OA5: Skills, Training, Rewards, Recognition & Upskilling
- OA6: Open Scholarly Communication
- INFRAEOSC Projects: Sustainable pathways to impact
- Joint Communications & Outreach

FAIR Data Productivity

Data Quality in EOSC, ein Fazit und Überlegungen

- The EOSC Data Quality Framework document is a good and comprehensive knowledge base
- However, there are no specific agreed implementation rules (how far can the EOSC-A go and on what basis)
- The topic of data quality (like FAIR metrics) is difficult to define without governance structures
- Data Quality towards “fit for purpose” or “fit for use”? How do we achieve a balance?
- Data Quality of training data/representatives has a significant impact on the efficiency, accuracy and complexity of tasks in the field of machine learning (ML) – catalogue of QA training data
- Scientific domains still operate too much as ‘closed clubs’;
- there is at least potential for implementation in the direction of NFDI, just as an example

Thank You !

Chris Schubert
TU Wien, University of Technology Vienna,

chris.schubert@tuwien.ac.at

> Poor-quality data which is incomplete, incorrect, or unrepresentative, can result in misleading outcomes. Ensuring high-quality datasets to train AI systems involves addressing challenges such as trust, access, bias, availability, and interoperability across the data lifecycle.
>
> Science in the age of AI, May 2024, © The Royal Society