TOWARDS A DATA QUALITY FRAMEWORK FOR EOSC Authorship Community

Carlo Lacagnina, Romain David, Anastasija Nikiforova, Mari-Elisa Kuusniemi, Cinzia Cappiello, Oliver Biehlmaier, Louise Wright, Chris Schubert, Andrea Bertino, Hannes Thiemann, et al.

TOWARDS A DATA QUALITY FRAMEWORK FOR EOSC

Authorship Community:

Carlo Lacagnina | Barcelona Supercomputing Center, Barcelona, Spain | 0000-0001-9434-9809
Romain David | European Research Infrastructure on Highly Pathogenic Agents | 0000-0003-4073-7456
Anastasija Nikiforova | University of Tartu, Tartu, Estonia | 0000-0002-0532-3488
Mari Elisa Kuusniemi | University of Helsinki, Finland |0000-0002-7675-287X
Cinzia Cappiello | Politecnico di Milano, Milano, Italy | 0000-0001-6062-5174
Oliver Biehlmaier | Biozentrum, University of Basel, Basel, Switzerland | 0000-0003-0825-8500
Louise Wright | National Physical Laboratory, Teddington, UK | [No ORCID]
Chris Schubert| Vienna University of Technology, Library, Vienna, Austria| 0000-0002-4971-2493
Andrea Bertino | SWITCH, Zürich, Switzerland | 0000-0002-5080-036X
Hannes Thiemann | Deutsches Klimarechenzentrum Gmbh (DKRZ), Hamburg, Germany | 0000-0002-2329-8511
Richard Dennis | Novo Nordisk Foundation Center for Stem Cell Medicine- reNEW, University of Copenhagen, Copenhagen, Denmark | 0000-0002-4472-7194

All authors have reviewed the manuscript and approved the submission.

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

Table of contents

EXECUTIVE SUMMARY 3
1. INTRODUCTION 6
2. BASIC CONCEPTS 7
2.1 WHAT WE MEAN BY DATA 7
2.2 WHAT WE MEAN BY QUALITY 10
2.3 QUALITY DIMENSIONS 12
2.4 QUALITY CONTROL VS QUALITY ASSURANCE 14
2.5 FITNESS-FOR-USE VS. FITNESS-FOR-PURPOSE 15
2.6 IMPACT OF DATA QUALITY 17
2.7 TYPES OF QUALITY ASSESSMENT 20
2.8 QUALITY ASSESSMENT WORKFLOW 22
2.9 DATA QUALITY MANAGEMENT APPROACHES 24
2.10 CERTIFICATION 27
2.11 QUALITY INDICATORS 32
2.12 VOCABULARY CHALLENGES 33
3. LANDSCAPE, CHALLENGES, AND REQUIREMENTS 36
3.1 LANDSCAPE OF THE LEADING ORGANISATIONS IN DATA QUALITY 36
3.2 OVERVIEW OF THE RELEVANT ISO STANDARDS 41
3.3 TASK FORCE SURVEY 50
3.4 PRELIMINARY DATA QUALITY REQUIREMENTS FOR EOSC 53
3.5 GAPS AND WAYS OF GAUGING COMMUNITY MATURITY IN DATA QUALITY MANAGEMENT 54
4. PRINCIPLES AND RECOMMENDATIONS 57
4.1 PRINCIPLES 57
4.2 RECOMMENDATIONS 58
5. REFERENCE 61
ANNEX I: TERMS AND DEFINITIONS 67
ANNEX II: RECOMMENDATIONS FOR THE FUTURE EOSC TASK FORCE 71
ANNEX III: SURVEY DETAILS 72

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

Executive summary

The European Open Science Cloud (EOSC) Association leverages thirteen Task Forces (TFs), grouped into five Advisory Boards, to help steer the implementation of EOSC. This document is released by the Data Quality subgroup of the "FAIR Metrics and Data Quality" TF. Data quality is critical in ensuring the credibility, legitimacy, and actionability of resources within EOSC. Indeed, certification and conformity mechanisms must be established to assure researchers that the infrastructures they deposit and access data conform to clear rules and criteria. If researchers feel a loss of control and visibility or have concerns about how professionally their data will be managed, additional **barriers to data sharing** will emerge. Informed by the results of a systematic literature review and community consultation utilising surveys, presentations, and case studies, this TF identified key concepts and formulated recommendations. Let us start with a quick view of the critical concepts.

Following the definition given by ISO 8000, with "**data quality**" we mean the degree to which a set of inherent characteristics of data fulfils requirements. Aligning actual data characteristics with the desired requirements implies that quality depends on context (both dataset application and lifecycle) and stakeholders, central elements in setting the requirements. **Requirements** must have a clear target aspect (e.g., privacy) and a level to reach (e.g., GDPR standard) and can be distinguished into functional (targeting the question "are you producing the right thing for its application?") and non-functional (targeting the question "are you producing it right?"). The goal of data quality management is to ensure that: (i) valuable information to understand the dataset is available, (ii) the dataset is reliable and ready to be used according to non-functional requirements (**fit-for-use**), and (iii) the dataset meets functional requirements (**fit-for-purpose)** when the purpose is known; if the purpose is unknown, data quality management ensures that the information necessary for users to make self-assessments of fitness-for-purpose is available. Note that in a FAIR ecosystem, where datasets could be reused far from the original purpose, information about the intended new purpose and associated functional requirements is limited.

This document explains basic concepts to build a solid basis for a mutual understanding of data quality in a multidisciplinary environment. These range from the difference between quality control, assurance, and management to categories of quality dimensions, as well as typical approaches and workflows to curate and disseminate dataset quality information, minimum requirements, indicators, certification, and vocabulary. These concepts are explored considering the importance of evaluating resources carefully when deciding the sophistication of the quality assessments. Human resources, technology capabilities, and capacity-building plans constrain the design of sustainable solutions.

We identified several **benefits** (or risks) of having good (or poor) quality and which are the stakeholders impacted. Despite these benefits, barriers and concerns prevent the provision of quality-assessed datasets, we identified these issues in detail.

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

Distilling the knowledge accumulated in this Task Force, we extracted cross-domain commonalities, lessons learned, and challenges. The resulting main **recommendations** are:

1. Data quality assessment needs standards to check data against; unfortunately, not all communities have agreed on standards, so EOSC should assist and push each community to agree on community standards to guarantee the FAIR exchange of research data. Although we extracted a few examples highlighting this gap, the current situation requires a more detailed and systematic evaluation in each community. Establishing a quality management function can help in this direction because the process can identify which standard already in use by some initiatives can be enforced as a general requirement for that community. We recommend that EOSC considers taking the opportunity to *encourage communities to reach a consensus in using their standards*.
2. Data in EOSC need to be served with enough information for the user to understand how to read and correctly interpret the dataset, what restrictions are in place to use it, and what processes participate in its production. EOSC should *ensure that the dataset is structured and documented in a way that can be (re)used and understood*. Quality assessments in EOSC should not be concerned with checking the soundness of the data content. Aspects like uncertainty are also important to properly (re)use a dataset. Still, these aspects must be evaluated outside the EOSC ecosystem, which only checks that evidence about data content assessments is available. Following stakeholders' expectations, we recommend that EOSC is equipped with essential data quality management, i.e., it should perform tasks like controlling the availability of basic metadata and documentation and performing basic metadata compliance checks. The EOSC quality management should not change data but point to deficiencies that the data provider or producer can address.
3. Errors found by the curators or users need to be rectified by the data producer/provider. If not possible, errors need to be documented. *Improving data quality as close to the source (i.e., producer or provider) as possible* is highly recommended. Quality assessments conducted in EOSC should be shown first to the data provider to give a chance to improve the data and then to the users.
4. User engagement is necessary to understand the user requirements (needs, expectations, etc.); it may or may not be part of a quality management function. *Determining and evaluating stakeholder needs* is not a one-time requirement but a continuous and collaborative part of the service delivery process.
5. It is recommended to *develop a proof-of-concept quality function* performing basic quality assessments tailored to the EOSC needs (e.g., data reliability and usability). These assessments can also support rewarding research teams most committed to providing FAIR datasets. The proof-of-concept function cannot be a theoretical conceptualization of what is preferable in terms of quality. Still, it must be constrained by the reality of dealing with an enormous amount of data within a reasonable time and workforce.
6. Data quality is a concern for all stakeholders, detailed further in this document. The *quality assessments must be a multi-actor process* between the data provider, EOSC, and users, potentially extended to other actors in the long run. The resulting content of quality assessments should be captured in structured, human- and machine-readable, and standard-

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

based formats. Dataset information must be easily comparable across similar products, which calls for providing homogeneous quality information.

7. *A number of requirements valid for all datasets* in EOSC (and beyond) and specific aspects of a maturity matrix gauging the maturity of a community when dealing with quality have been defined. Further refinement will be necessary for the future, and specific standards to follow will need to be identified.

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

1. Introduction

The European Open Science Cloud (EOSC) is aligned with the vision of enabling a trusted, virtual, federated environment in Europe to store, share and reuse research outputs across borders and scientific disciplines. Priority areas to develop a functional infrastructure embracing this vision are being identified based on community feedback from an open consultation process, where advisory boards (grouping Task Forces) play a crucial role. One of these Task Forces (TFs) is named "FAIR Metrics and Data Quality". It assesses the applicability of the FAIR metrics across research communities and evaluates a range of tools to enable uptake. Recommendations are made to update metrics and adopt tools as appropriate. In addition, the group probes the state-of-art to establish a cross-disciplinary understanding of data quality. It conducts case studies and engages stakeholders to identify common features and dimensions to define a data quality approach for EOSC. A multidisciplinary (biology, metrology, climatology, data science and management, philosophy, computer sciences, etc.) international (seventeen European countries) group of twenty-six experts are participating for two years in a mixed-method approach consisting of virtual discussions, workshops organisation, and participation, case studies collection, and survey development and dissemination. The TF has been split into two subgroups; this document relates to the Data Quality subgroup.

Data quality is vital in achieving credibility, legitimacy, and actionability of the Open Science and EOSC. Indeed, the EOSC Declaration[1] advocates certification and conformity mechanisms to be established to assure researchers that the IT infrastructures where they deposit and access data conform to clear rules and criteria. If researchers feel a loss of control and visibility or have concerns about how professionally their data will be managed, additional barriers to data sharing will emerge. The centrality of data quality motivated this TF in the context of EOSC. Informed by the results of a continuous literature review (including ISO standards), we first identified key concepts, terminology, and major organisations dealing with data quality. We extracted cross-domain commonalities, lessons learned, and challenges during the first year and then produced the first version of a recommendation report (current version 1) that will be updated by embedding community feedback during the second half of the Task Force duration. The last version of the report (to be released in November 2023) aims to generate a cross-disciplinary understanding of data quality and identify a data quality strategy for EOSC, considering existing work in the area. Strategy elements for future developments and investments are also outlined.

[1] European Commission, Directorate-General for Research and Innovation, Turning FAIR into reality : final report and action plan from the European Commission expert group on FAIR data, Publications Office, 2018, https://data.europa.eu/doi/10.2777/1524

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

2. BASIC CONCEPTS

This section shows the common ground understanding reached within the TF about basic concepts in data quality. It is a necessary step to remove confusion and to generate an agreed terminology across the TF members and for the reader.

2.1 What we mean by data

A digital object is an object composed of a set of bit sequences[2]. This document focuses on a subset of digital objects: data. We define data according to the DIKW pyramid proposed by Rowley (2007). This well-known hierarchical model contextualises the concepts of data, information, knowledge, and wisdom concerning one another (figure 2.1.1). The underlying assumption is that data form the basis for making informed decisions.

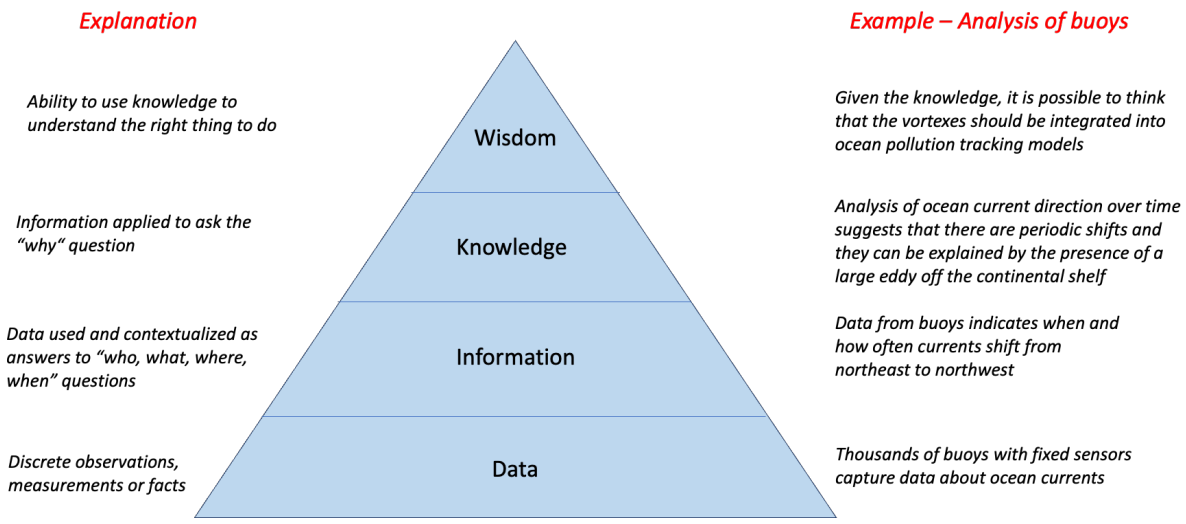

Figure 2.1.1. *The DIKW pyramid is explained with an example showing the transformation process from data to wisdom (a reworking of the image in https://www.i-scoop.eu/big-data-action-value-context/data-information-content-knowledge-input-insight-action-value/).*

Rowley (2007) defines data as "discrete, objective facts or observations, which are unorganised and unprocessed and therefore have no meaning or value because of lack of context and interpretation". ISO/IEC 2382-1:1993 offers another top-down perspective. It defines data as a reinterpretable representation of information in a formalised manner suitable for communication, interpretation, or processing.

As far as information is regarded, Wallace (2007) defines it as data endowed with meaning and purpose. On the other hand, the concept of knowledge is not so clearly defined in the literature: there is no agreement, and the definitions are often much more complex than those for data or information. Wallace (2007) reports:

- knowledge is the combination of data and information, to which is added expert opinion,

[2] https://public.ccsds.org/pubs/650x0m2.pdf

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

skills, and experience, to result in an asset, which can be used to aid decision making;
- knowledge is data and information that have been organised and processed to convey understanding, experience, accumulated learning, and expertise as they apply to a current problem or activity;
- knowledge builds on information extracted from data [...] While data is a property of things, knowledge is a property of people that predisposes them to act in a particular way;
- knowledge is information enriched with experience. Information is simply descriptive, while experience provides insights that are valuable for decision-makers.

Rowley (2007) does not discuss the wisdom concept, but some contributions refer to it as "knowing the right things to do" (Chrisholm and Greg 2007), which supports data-driven decision-making. Figure 2.1.2 provides a heuristic grasp of the DIKW concept, where data are sketched as a collection of points or values, information gives colour or meaning to the values by contextualising them, knowledge connects the meaningful values by finding patterns and relationships among the different elements to reach an understanding, the accumulated knowledge leads to extract new meaning to specific information (insight) that can be connected with experience to make an informed decision (wisdom).

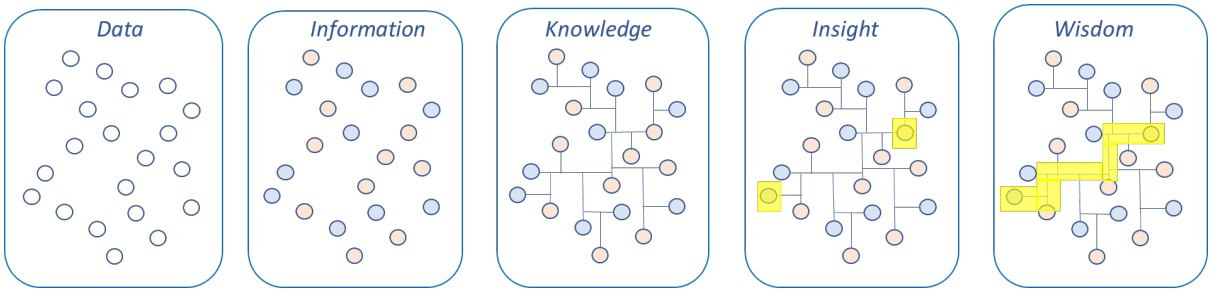

Figure 2.1.2. *The transformation from data to wisdom (adapted from https://www.netpresenter.com/knowledge-center/corporate-performance/how-to-transform-big-data-into-big-wisdom).*

The DIKW pyramid (figure 2.1.1) shows that data are the basis for extracting information. In analogy with the manufacturing environment, data represent the raw materials and information the product output (Wang et al. 1998). Therefore, it is possible to claim that data are collected "to be examined, considered, and used to help decision-making"[3] or "as a basis for reasoning, discussion, or calculation"[4]. As such, the quality of these data is fundamental for the decision-making process (see section 2.6).

To wrap up, in this context, we consider *data* as any observation, measurement or fact digitally represented as text, numbers, or symbols that form the basis to extract information. A dataset is a collection of data organised in the same format and associated with a unique body of work[5]. It is

[3] Cambridge dictionary
[4] Merriam-Webster dictionary
[5] https://www.usgs.gov/faqs/what-are-differences-between-data-dataset-and-database

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

associated with metadata that describes the dataset. Metadata is essential to support users in data usage and discovery. An organised collection of data stored as multiple datasets is a *database*. Those datasets are stored and accessed electronically from a computer system that allows the data to be easily accessed, manipulated, and updated[6].

Data classifications

In the literature, data have been classified using different perspectives. Here we report the most common approaches.

1. **Data classification from the data representation and storage perspective.**

In information technology, a data structure is a particular way of organising and storing data in a computer such that it can be accessed and modified efficiently. More precisely, a data structure is a collection of data values, the relationships among them, and the functions or operations that can be applied to the data[7]. Three types of data structures are distinguished:

- **structured data:** Data are in a standardised format; they are organised by following a precise structure that complies with a data model. Example: relational (tabular) data;
- **semi-structured data:** Data are not stored in a relational database; thus, they do not follow a tabular structure and a specific data model, but they are characterised by specific properties that facilitate their analysis. They can be transformed and processed in a way that can be stored in a tabular structure. Example: CSV, XML;
- **unstructured data:** Data do not follow a tabular structure and do not comply with a data model. Their properties do not make possible their storage in relational structures; they need other platforms and techniques for their storage and management. Example: free text, multimedia content.

2. **Data classification from a data source perspective.**

This classification originates from the needs in the "big data" context. Big data is a term used for defining data characterised by large volumes, variety and velocity that cannot be processed by conventional database systems (Dumbill 2013). UNECE (United Nations Economic Commission for Europe) classifies big data into three types of sources[8]: **human-sourced** (e.g., blog comments), **process mediated** (e.g., banking records), and **machine-generated** (e.g., sensor measurements).

3. **Data classification from a usage perspective.**

Data can also be classified by their frequency of access. Even if there is a high capacity today in the cloud, not all data require the same availability[9]:

- **hot storage** for frequently accessed data that must be stored on quick-to-access repositories;

[6] https://www.usgs.gov/faqs/what-are-differences-between-data-dataset-and-database
[7] https://www.bigdataframework.org/data-types-structured-vs-unstructured-data/
[8] UNECE. Classification of types of big data. https://unstats.un.org/unsd/trade/events/2015/abudhabi/gwg/GWG%202015%20-%20item%202%20(iv)%20-%20Big%20Data%20Classification.pdf.
[9] https://www.ctera.com/company/blog/differences-hot-warm-cold-file-storage/

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

- **warm storage** for data accessed less frequently, for instance, those used for reporting or meta-analyses. No need to be accessed as quickly as hot data so that they can be saved on slightly slower, capacity-optimised disks;
- **cold storage** for data that are rarely accessed and archived for compliance reasons. It can be stored on even slower, "cheap and deep" disks;
- **frozen data** are a particular type of cold data that will likely never be reused (e.g., sensitive data associated with genome sequencing).

4. **Data classification from a sensitiveness perspective.**

This classification concerns the access rights relevant to the discipline at stake. The Sensitive Data Expert Group[10] (2020) proposes one classification:

- **high risk data** require strong controls against unauthorised access, loss or modifications that could result in significant risk of harm to both data subjects and owners;
- **medium risk data** refer to confidential data, which requires strong controls against unauthorised access, loss or modifications that may put data subjects and owners at risk;
- **low risk data** need rules against unauthorised access, loss or alterations to ensure aspects related to data integrity rather than to risk prevention.

2.2 What we mean by quality

"*Data quality*" is commonly understood as "the degree to which a set of inherent characteristics of data fulfils requirements" (ISO 8000). This definition introduces essential concepts that illustrate what we mean when referring to quality. First, "*degree*" - when assessing quality, we seek to apply a measurement (typically in the form of a scale) of specific quality aspects, commonly referred to as dimensions. A *quality dimension* represents a single aspect or construct of quality (e.g., accuracy, completeness), which allows complex areas of data quality to be subdivided into groups, each with its specific way of being measured (Wang and Strong, 1996, see also section 2.3). Second, "*characteristics*" - a set of features of the dataset under evaluation that need to be compared against desiderata, i.e., "*requirements*." Aligning data characteristics with the requirements implies measuring how far the characteristics are against a scale, which means quantifying the dataset quality. Requirements are characterised by an apparent target aspect (e.g., privacy) and a level to reach (e.g., GDPR standard); requirements represent the translation of the "stakeholder's voice" into the product-making process.

These considerations led Fürber and Hepp (2016) to elaborate more on the definition of quality. According to their study, "data quality" is a comparison of the actual state of a particular set of data to the desired state, with the desired state typically referred to as "fit for use", "fit for purpose", "to specification", "standard", "meeting consumer expectations", "free of defect", "free of errors", or "meeting requirements". The desired state is defined by requirements, specifications and expectations set by the user, producer, law, business, etc., closeness to the desired state is measured utilising dimensions. Therefore, quality depends on the stakeholder - users and providers are usually interested in different dataset aspects - and on the context in which data are used (both

[10] https://zenodo.org/record/4690571

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

its application and lifecycle). Dataset lifecycle refers to a set of interrelated stages that a dataset encompasses from its generation to its delivery or deletion (figure 2.2.1). Depending on the stages of a dataset in its lifecycle, some quality dimensions tend to become more relevant than others (Peng et al. 2021). For instance, dimensions associated with scientific quality are highly regarded in the earlier stages of dataset definition and development. In contrast, aspects like the licence of use or preservation are better deemed when the dataset is served.

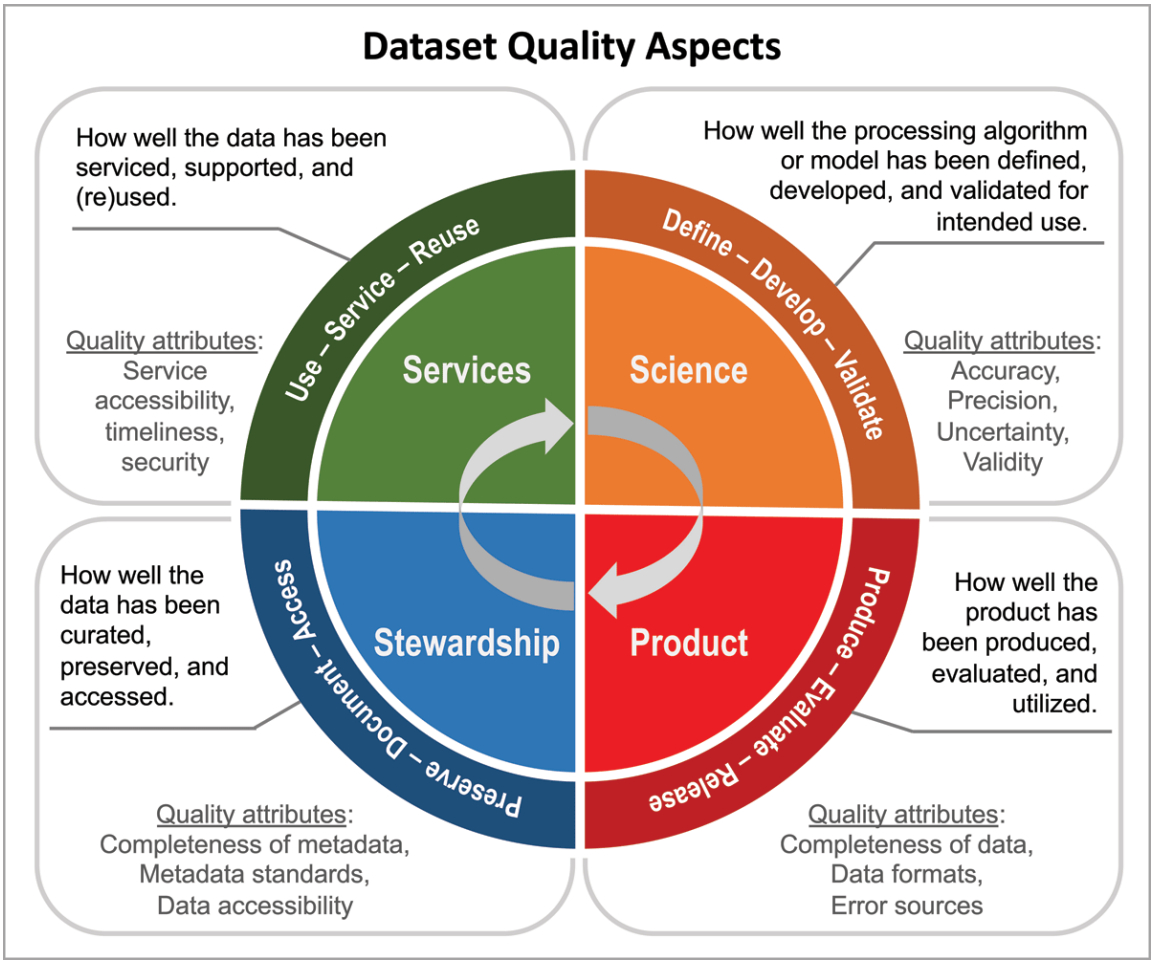

Figure 2.2.1. *Credits to Peng et al. (2021). A schematic diagram of four quality aspects (i.e., science, product, stewardship and service) throughout a dataset lifecycle, three key stages and a few attributes associated with each quality aspect (e.g., define, develop, and validate stages for the science quality aspect).*

In conclusion, quality depends on the application (albeit there are requirements valid regardless of application, see section 2.5), stakeholder, and dataset lifecycle. It implies that the same data may be suitable for one application or user but not for another (Tayi et al. 1998) because requirements differ. Therefore, achieving a level of quality at which the dataset fully satisfies all use cases - a.k.a. “absolute data quality” (Bouchana and Idrissi 2015, Nikiforova 2020) - is unfeasible. The following sections explore how to go beyond this difficulty to reach a minimum acceptable quality level. There are dataset characteristics that need to be met independently of the use case.

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

A final remark concerns *perceived* quality. The picture outlined above fits well what is known as "objective quality", where the curator of a dataset can work to improve its intrinsic characteristics. However, another aspect of quality deals with extrinsic attributes surrounding the product, determining its final value and impacting its perceived (as opposed to objective) quality. These extrinsic attributes are more related to the service provider's image and heritage, affective user judgments, cultural or social values, advertising, and marketing promotion techniques (Stylidis et al. 2020). These can become requirements once the user's feelings about the product are mapped into the design specifications to maximise the user's satisfaction (Mitsuo 1995). Unfortunately, latent user demands are rarely recorded because the users seldom, if ever, formulate their latent expectations sufficiently precisely and unambiguously (Lindemann et al. 2019). While this document does not dig into the topic of perceived quality because it is not in the remit of the dataset curator, it is important to mention its existence as an additional element impacting the overall dataset quality.

2.3 Quality dimensions

Data quality is often managed by looking at single data attributes as dimensions, each with its way of being measured (Wang and Strong 1996). There are many quality dimensions, e.g., completeness, and accuracy, prioritised differently depending on the author (e.g., Alexander and Tate 1999, Shanks and Corbitt 1999, Zhu and Gauch 2000, Knight and Burn 2005), and there are several ways to categorise the quality dimensions. The most typical academic way is distinguishing four categories: intrinsic, contextual, representational, and accessibility. Intrinsic quality implies that data have quality in their own right. Contextual quality highlights the request that quality is considered within the context of the task at hand; data must be relevant, timely, complete, and appropriate to add value. Representational and accessibility quality emphasises the importance of computer systems that store and provide access to data; that is, the system must present information in such a way that it is interpretable, easy to understand, easy to manipulate, and is represented concisely and consistently; also, the system must be accessible but secure (Lee et al. 2002).

ISO 25000 (covered in more detail in section 3.2) clusters quality dimensions into inherent and system dependent. The former is the degree to which data characteristics have the intrinsic potential to satisfy stated and implied needs when data are used under specified conditions. This category refers to (i) data values and restrictions (e.g., business rules governing the quality required for the given application), (ii) relationships of data values (e.g., consistency), (iii) metadata. The system-dependent category refers to the degree to which data quality is reached and preserved within an IT system when data are used under specified conditions. From this point of view, data quality depends on the technological domain in which data are used; it is achieved by the capabilities of computer systems' components such as hardware devices (e.g., to make data available or to obtain the required precision), computer system software (e.g., backup software to achieve recoverability), and other software (e.g., migration tools to achieve portability).

Similar categorization of quality dimensions is offered by ISO 19157 and manufacturing practices, like for instance the wine industry. The former classifies as direct the dimensions associated with inspecting the data values in a dataset, and classifies as indirect the dimensions focusing on

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

external knowledge of the dataset, such as lineage or documentation. The latter, the example in the manufactory, categorises quality dimensions in wine as intrinsic - flavour type and intensity and extrinsic - brand, packaging, etc., factors (Langstaff 2010). The parallelism between two very different products - datasets and wine - shows similarities in the approach to quality. It is common to distinguish between quality dimensions inherent to the product (e.g., data soundness - wine toxicity) and other dimensions closer to the system where the product is stored or shared (e.g., archiving - packaging).

All these examples of dimension categorization show that the most typical way of thinking about quality is to distinguish it into two prominent families: product-dependent and system-dependent. As shown above, Lee et al. (2002) further split these two families into intrinsic and contextual the former and representational and accessible the latter.

Dataset types

Defining the quality dimensions relevant to the specific context often represents one of the first steps in any quality assessment process (section 2.8). The data category under examination is crucial in selecting the appropriate quality dimensions. Data come from various sources, including social media, the web, or administrative registers, and tend to be characterised by quite different structures. Here we give a flavour of the most common dataset types to show the complex landscape EOSC will potentially need to consider:

- **structured** (or relational) data: the type of data for which various data quality models and dimensions have been proposed over the years (e.g., Wand and Wang 1996, Wang and Strong 1996, Redman 1996, Naumann 2002, Cai and Zhu 2015). For instance, Wang and Strong (1996) highlight the importance of considering intrinsic, contextual, representational, and accessibility quality. A few literature contributions state the definition of specific data quality dimensions for unstructured data (e.g., Batini and Scannapieco 2016);
- **texts** (e.g., free text, publications). A text comprises a set of sentences that are units of one or more words. Assessing the quality of text needs to consider syntax and semantics, and consequently, quality dimensions such as readability and coherence could be relevant (Batini and Scannapieco 2016);
- **data streams** (e.g., sensor data, social data) are defined as an infinite sequence of data items. Klein and Lehner (2009) propose a set of quality dimensions derived from the categories defined by Wang and Strong (1996) and show that it is necessary to include additional dimensions to capture specific aspects such as sensors' reliability, the presence of outliers, the temporal context, and the amount of captured raw data. Often, this type of data requires (almost) real-time quality assessments (on the fly);
- **multimedia** (images, audio, video) is classified as unstructured data. A data quality model must include the factors that may influence the perceived multimedia quality (e.g., lighting conditions, spatial resolution, sharpness contrast, colour balance, noise-to-signal ratio);
- **maps** or geodata, in general, are another type of unstructured data whose quality is related to measuring the degree of adherence of the existence of geographic data (features, attributes, functions, relationships) to the corresponding elements in the real world mapped;
- **linked data** are structured data that relate to other data through an RDF (Resource

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

Description Framework) representation. RDF allows the specification of the connected object's URI (Uniform Resource Identifier). An RDF triple (subject-predicate-object) provides mechanisms for describing groups of related resources and the relationships among them. These semantic data structured in the form of the “subject-predicate-object” expression consider the subject and the object as resources. The predicate defines the properties of the subject of the statement. Nodes are sets of subjects and objects. The subject is a URI or a blank node, the predicate is a URI, and the object is a URI or a literal or a blank node. Literals and blank nodes are RDF terms (Cappiello et al. 2016). From a data quality perspective, linked data differ from “traditional” structured data due to the need to evaluate the quality of the connection among sources and the possibility of navigating from one resource to another. Zaveri et al. (2016) survey highlighted the data quality dimensions needed to evaluate these peculiar aspects.

The list described above represents an incomplete starting point; the authors are working on expanding and improving it.

2.4 Quality control vs quality assurance

Although the term “quality” is often followed by “control” or “assurance,” there is a significant difference between the two concepts. The former is a *reactive* process focused on "detecting defects" and analysing the product or service. In contrast, quality assurance is a *proactive* process focused on "preventing defects" by examining the procedures for making the product. Quality control encompasses a set of steps applied to detect and identify errors. It verifies to what extent the requirements are met.

In contrast, quality assurance provides confidence that requirements will be met (ISO 19158:2012) because it deals with the processes involved in making the product. Although quality control and quality assurance have long been familiar concepts, these belong to the broader concept of quality management (figure 2.4.1), which also incorporates quality planning and quality improvement (WMO No. 1221). Quality management focuses not only on the quality of the product but also on the means (organisational structure, procedures, processes, and resources) to achieve it. The ISO 9000 family (see section 3.2) promotes adopting a quality management system (QMS), which emphasises the importance of continual improvement of processes based on objective measurements as well as understanding and meeting requirements. The customer plays a significant role in defining requirements as input to the quality management process. ISO 8000-2 offers a similar understanding of quality management, describing it as a set of coordinated activities organisations use to direct and control quality. These activities include formulating policies, objectives, planning control and assurance to achieve continuous improvement.

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

Figure 2.4.1. *Sketch showing the components of quality management.*

2.5 Fitness-for-use vs. fitness-for-purpose

A dataset is characterised by an inherent set of features that need to be fulfilled independently of the specific application and are generally defined by subject matter experts. For instance, a dataset containing geographically distributed monthly precipitation is expected to have non-negative precipitation; this dataset may fit monthly analyses but is not helpful for purposes requiring hourly precipitation values. In our example, the hourly requirement set by the user is not met. In contrast, the essential non-functional requirement of having a reliable dataset (non-negative precipitation in our case) is met. We may say that the former requirement defines whether the dataset is *fit-for-purpose*, whereas the latter requirement defines whether the dataset is *fit-for-use*. The distinction between fitness-for-purpose and fitness-for-use is well-known in manufacturing. Indeed, ITIL[11] identifies service value as made by "utility" (fit-for-purpose) and "warranty" (fit-for-use). The utility is targeted when the service or product is fit for purpose, which means that the service/product must fulfil customer needs. For instance, it does not matter that you can rent a specialised black-and-white printer for half the average market price if the user needs a colour wax printer.
On the other hand, fit for use, or warranty, means that service is available as the user needs it. A good example is a mobile; it needs to be ready to use wherever you want to make a call. If the connection keeps dropping every time, it is worthless.

The distinction between fitness-for-purpose and fitness-for-use can be better understood by introducing the concept of functional and non-functional requirements. This distinction stems from

[11] ITIL v3 (Information Technology Infrastructure Library) https://www.axelos.com/certifications/itil-service-management

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

software quality and is portable to data quality. According to Chen et al. (2013), software structural quality refers to how software meets non-functional requirements that support the delivery of the functional requirements. Verifying non-functional requirements proves software performance (e.g., robustness, modularity). For instance, a script, independently of its final purpose, should be secure or reactive. These requirements are typically defined by developers, software architects, and technical leaders. Structural quality is evaluated by analysing the software structure and source code. In contrast, functional requirements describe what the software should do as the user dictates. For instance, a notification email is required whenever the user registers for the first time on some website.

Another management model, V-cycle[12], offers a different perspective to distinguish between the two components of fitness by saying that fitness-for-purpose responds to “are you building the right thing?" (addressed by validating against the user needs). Fitness-for-use responds to "are you building it right?” (addressed by verifying against non-functional requirements). The guidance offered by both software quality and service management highlights relevant elements portable to data quality: requirements are the basis for defining quality (see section 2.2), and these can be distinguished in functional (pointing to fitness-for-purpose) and non-functional (pointing to fitness-for-use) requirements. In a FAIR ecosystem, where datasets may be reused far from the original purpose, information about the intended new purpose and associated functional requirements are often limited. In addition, these datasets may be used by audiences not involved in the initial data collection and having various scientific backgrounds (Baker et al. 2016). In this case, the focus in quality management on the fitness-for-purpose aspect is to provide value-added information, including metadata, documentation, and guidance to enable users to assess the respective suitability and fitness themselves[13] (Whitfield 2012). Therefore, the goal of data quality is to ensure that requirements are met and, when the functional requirements are not known because the user’s purpose is unknown, data quality must provide the typical information[14] (e.g., how to use the dataset, its limitations) the user needs to make informed decisions about the dataset to select. Knowing the typical information the users need for their self-assessment of dataset fitness-for-purpose may not be a task for practitioners in data quality because it requires user profiling methods targeted by user engagement, market research, psychology, and communication teams. These teams are also best placed to define ways to present and disseminate the information that quality practitioners make available. According to Whitfield (2012), maximising the utility of datasets needs not simply to increase the number of times they are used but rather to ensure that they are used wisely. The added-value information disseminated with data quality is crucial to offer such usage guidance plus guidance in the selection process among similar datasets. See section 2.6 to overview the advantages of establishing data quality management practices.

[12]https://www.cio.bund.de/Web/DE/Architekturen-und-Standards/V-Modell-XT/vmodell_xt_node.html
[13] http://qa4eo.org/docs/QA4EO_Principles_v4.0.pdf
[14] How and for what purpose the dataset has been collected and whether the data are in a raw form or is processed, if anonymization techniques were put in place, what are the attributes, whether there are specificities to consider, connections between attributes, etc.

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

In summary, data quality management ensures that (i) valuable information to understand the product is provided, (ii) the product is reliable and usable (fit-for-use), and (iii) the product is fit-for-purpose when the purpose is known; if the purpose is unknown, data quality management ensures that the information necessary for users to make self-assessments of fitness-for-purpose is available.

2.6 Impact of data quality

Service or product value originates from various factors where quality plays a key role. Some of these factors are:

- **reliability and usability.** The primary goal of data quality is to ensure that products are reliable and usable for decision-making (a.k.a. fitness-for-use, see section 2.5). Only reliable data can power accurate analyses, necessary ingredients for trusted decisions (in line with the "garbage in, garbage out" principle). Having reliable and usable data contributes to making the service or product provider an authoritative source of information. When organisations constantly deliver erroneous products, their reputation value decreases quickly;
- **compliance.** Data quality also ensures compliance with legal regulations (e.g., GDPR or intellectual property) and guarantees that the product has the expected characteristics, either explicit or implied. In other words, before opening the box, the dataset, in this case, the stakeholder already trusts its content;
- **trust.** Trust[15] is generated when data comply with standards because standards are a compendium of the knowledge accumulated over the years mediated by consensus from different experiences. Conformity assessments of standard compliance deliver the confidence needed for user acceptance and thus create trust between the user and provider. Conformity assessments become increasingly relevant when the designated community is multidisciplinary, where users may not have the personal experience to evaluate quality from the data alone[16];
- **cost prevention.** Having reliable data prevents the costs of remediation or correction. Quality management comes with a cost that can be considered a preventive cost or an investment incurred by organisations to reduce poor data quality (Batini et al. 2009). Making inadequate products does not save money; quality management decreases time and resources spent on recurring problems, as many are resolved permanently. In the long run, it is more expensive to find mistakes after they have been made than to prevent them in the first place. A lack of data quality leads to a waste of time and money, poor decisions, and frustration for the user who cannot work properly and the provider whose datasets have less uptake. In turn, the providers' reputation is damaged, and they do not know how to deliver products that meet the user's requirements and relevant regulations. It is important to note that the wasted

[15] Trust can be gained through a set of shared ethical, procedural and technical norms, which tacitly or explicitly define what is appropriate behaviour by all parties involved and how things should work (McLeod C. (2020). Trust. In: The Stanford Encyclopedia of Philosophy plato.stanford.edu/archives/fall2020/entries/trust/)

[16] CoreTrustSeal Trustworthy Data Repositories Requirements 2020–2022

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

time is not only related to basing decisions on inaccurate data, but also to the need for the user to post-process data to remove errors or structure data according to the available tools. Quality management should avoid being perceived as an additional burden to the data provider/producer with little investment return. The substantial number of regulations and documentation demands may reduce intrinsic quality motivation; thus, simplicity is necessary. Quality management risks being seen as an additional layer with lower priority, whereas it should be part of the production process, making it more efficient, enhancing teamwork and capturing corporate knowledge;

- **removal of barriers to data sharing.** Curating data using well-defined standards and measuring closeness to requirements facilitates interoperability and boosts continuous improvement. In this respect, raising quality promotes data sharing and assures researchers that the infrastructures they deposit and access data conform to clear rules and criteria. If researchers feel a loss of control and visibility or have concerns about how professionally their data will be managed, additional barriers to data sharing will emerge[17];
- **uptake and misuse.** When data quality information is disseminated to the users, it guarantees sufficient metadata to understand how to use a product. Better understanding helps to reduce misuse and to increase uptake. Information about how to use a product enables the users to self-assess whether the product applies to their specific needs. Moreover, by assuring a product meets user requirements, quality increases the chances for product uptake and calls for necessary identification of the user needs;
- **comparability and guidance.** In many cases, there is information asymmetry between producer and user because dataset characteristics may be hidden before the user can explore them, or some characteristics may be known to the data producer only (e.g., data collection process), or some users may not be able to judge the dataset characteristics themselves (being not an expert in the field the dataset was produced). Quality information can fill this gap. Without quality information, the user's best guess for a given product is that it is of average quality (Akerlof 1970), diminishing the ability to discriminate during the dataset selection process (Langstaff 2010). Lack of quality information is also detrimental for the data producers, who will not improve the dataset to increase its uptake. This reduces the average quality of the datasets available, generating a negative feedback loop. The result is that a dataset exchange market, in which there is asymmetric information concerning quality, shows characteristics like those described by Gresham's Law: the bad drives out the good (Phlips 1983). Therefore, disseminating quality information reduces the information asymmetry between supply and demand, diminishing the risk of going towards poorer quality and benefitting the users by guiding them in the data selection process.

The benefits of data quality affect a variety of stakeholders who define the requirements. Some of these requirements are complementary, overlapping, and contradictory, with stakeholders potentially taking multiple roles simultaneously. The main stakeholders (figure 2.6.1) benefiting

[17] European Commission, Directorate-General for Research and Innovation, Turning FAIR into reality : final report and action plan from the European Commission expert group on FAIR data, Publications Office, 2018, https://data.europa.eu/doi/10.2777/1524

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

from quality are:

- **users:** avoid wasting time and money on poor decisions based on unreliable and unusable datasets, have access to quality information to understand the product better and select which one meets their needs. Researchers and reviewers are often data users;
- **data providers/producers:** have feedback about which requirements their products need to meet, improve the quality of their products, and increase data uptake. Data producers and providers enter both in this category; researchers are often data producers;
- **data service** becomes a trusted, authoritative source of information that understands the users' needs, the service delivery is improved. Data services are wide-ranging from the typical dataset catalogue and related repository to scholarly institutions to information or digital journal publisher;
- **funders:** get a measure of how compliant the funded products or services are with (their) requirements. Funding authorities must measure the quality of the data produced against the initial project or service goals and management plans.

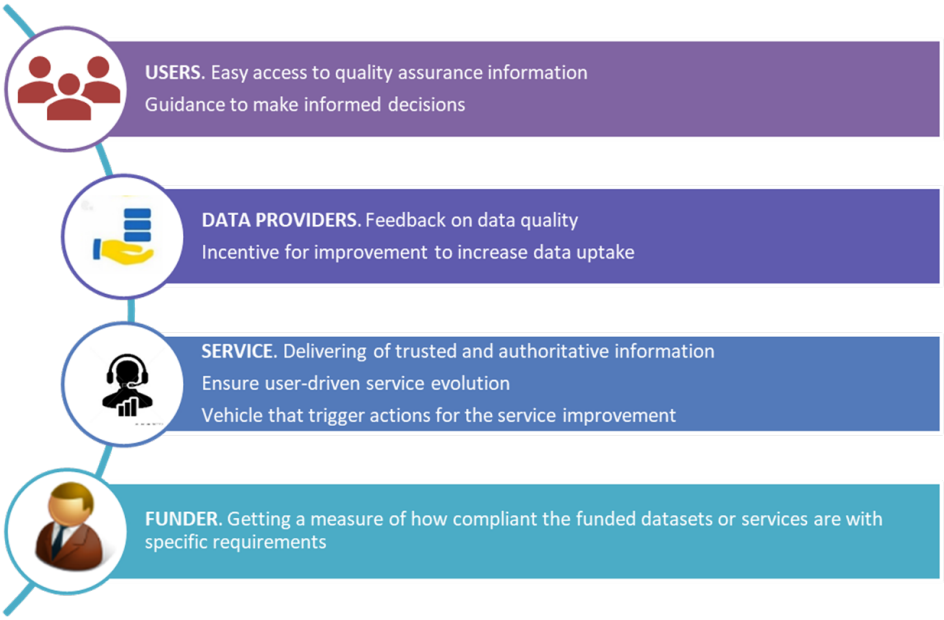

Figure 2.6.1. *A sketch of the actors benefitting from quality management in a data delivery service. Actors (e.g., publishers, reviewers, repositories) typically involved in research can enter into one or more of these categories. For instance, researchers can be either data producers or users.*

Despite these benefits, barriers and concerns prevent the provision of quality-assessed datasets. A survey launched by this TF (see section 3.3) revealed some of these barriers, which include (i) lack of organised community methods; (ii) concerns about how to make accessible and reusable (due to lack of context) the quality assessment results; (iii) resource limitation putting quality to the second order of importance because of no reward in providing standard-compliant datasets or no funding dedicated or quality is less critical compared to other academic achievements; (iv) lack of experts in quality; (v) time consuming; (vi) lack of tools for the evaluation; (vii) standards are more effective when tailored to the specific characteristics, needs, users of the organisation involved but this leads to the development of practices local to the specific organisation, complicating the realisation of a

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

common framework. In turn, this leaves small organisations or projects with no standards and limited resources to invest in creating their ad-hoc standards, which most likely will be used by them only and will not usually be mapped to others.

These exciting research paths may stimulate the community focused on quality in investigating ways to reduce the barriers that prevent thorough quality management.

2.7 Types of quality assessment

When dealing with data quality, it is essential to define which facet of a dataset the practitioner wants to assess. From the perspective of quality, dataset facets are (i) data content, (ii) metadata distributed either within the files or external as documentation, (iii) the ecosystem where it is served. Each dataset's facet is associated with quality dimensions that can be the same across facets, but may be measured differently depending on the selected facet[18] so that the resulting quality assessment can be either technical, scientific, or stewardship. This is illustrated in figure 2.7.1. There are diverse ways to categorise types of assessments; the one described here is the most typical (Peng et al. 2021); another approach is outlined later in the text.

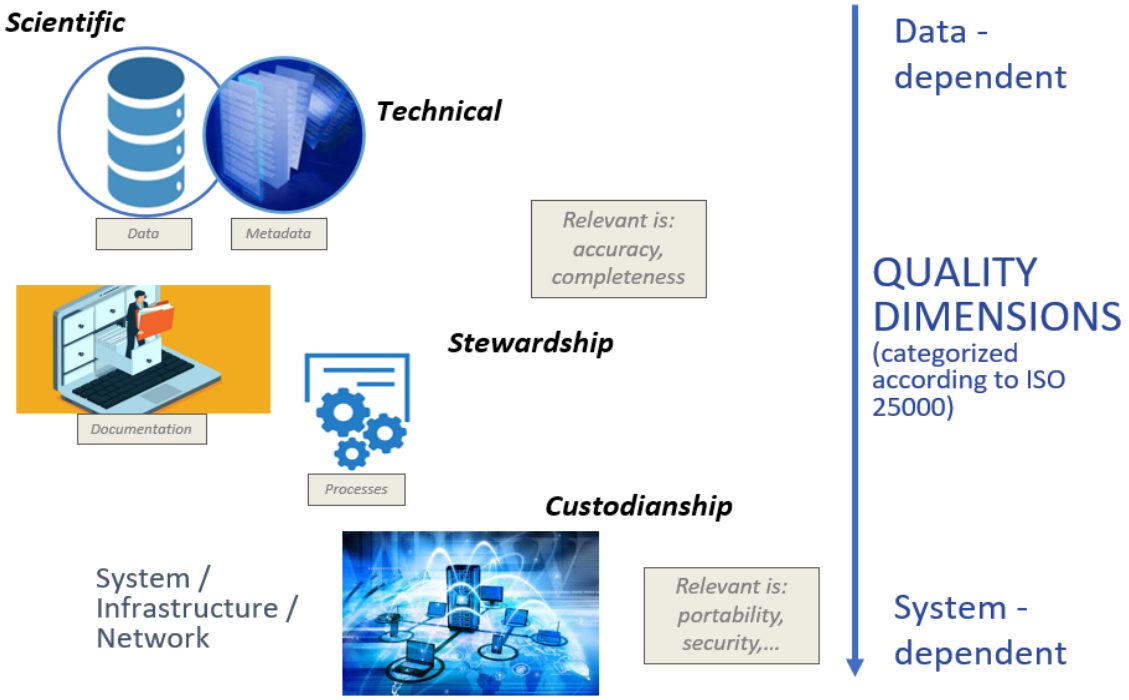

Figure 2.7.1. *A sketch showing the different facets characterising a dataset: the content (data values), how the content is structured and documented within the files (metadata), additional information about the dataset given external-to-files (documentation and processes), the ecosystem where the dataset is transmitted, stored, and safeguarded (system/infrastructure/network). In black are reported the types of quality assessments and typical dimensions that move from more data-dependent to more system-dependent. These two categorizations come from ISO 25000.*

[18] For instance, the dimension of completeness is measured differently when referring to gaps in the data points or to a lack of comprehensiveness of the metadata model.

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

The *technical assessment* regards the metadata and data file checks, while the scientific assessment consists of data content and cross-data content checks (Stockhause et al. 2012). For example, when the evaluator looks for metadata standard compliance, such as compliance against the Attribute Convention for Data Discovery (ACDD[19]), the evaluator checks whether the attributes describing the files comply with a set of community-recognized metadata characteristics (e.g., specific date-time format). In this case, there is no file content check or scientific evaluation: a metadata conformity check is a purely technical assessment (Evans et al. 2017). On the other hand, when the evaluator plots the dataset variable and checks for a faithful representation of a physical process (e.g., El Niño events) against skill metrics, here the scientific soundness of the data content is considered (Haiden et al. 2019). The technical assessment generally checks data files' consistency and completeness among the distributed data and metadata repositories, adherence to formal standards, and the use of identified formats. Checks will vary according to the specific needs of the service and may include compliance against community standards (conformity), temporal and spatial checks for unexpected gaps (completeness), and identification of corrupted values (integrity). Lawrence et al. (2011) postulate a generic technical and scientific assessment checklist.

As far as the *scientific assessment* is concerned, this refers to scientific analyses of the physical content described by the dataset to check for its soundness (Lacagnina et al. 2022). Given the nature of this assessment, it is typically conducted by experts in the domain of the data producer. Analyses may include uncertainty characterization, validation against reference datasets, and reproducibility of temporal/spatial patterns.

The technical and scientific assessments are often accompanied by the *stewardship assessment* to guarantee the accessibility and understandability of the dataset distributed. Here we refer to stewardship as anything of relevance in dataset quality that is not associated with the data and metadata file content, e.g., documents accompanying the dataset describing how to use it and the context of the dataset creation. Typical examples regard the description of the algorithms or models used to produce and process the data, provision of the DOI and licence of use, verified network address to access the data and information about the archiving procedures. The goal is to ensure that the dataset is well documented and contextualised, the processing chain is visible, and the data is readily obtainable and usable.

Stewardship assessment may be further distinguished from *custodianship assessment*, this latter to emphasise its components are farther from the data themselves and much more related to the ecosystem where the dataset is shared or archived. Whilst dimensions like security or portability are heavily system-dependent, assurance of adequate dataset contextualization, typically associated with completeness of the documentation, is closer to the inherent data. This further distinction is inspired by the quality dimensions categorization described in ISO 25000 (see section 3.2) and it is adopted from data governance practices, where responsibilities for data management are divided between data steward and data custodian[20][21]. The former is responsible for what is stored with a

[19] https://wiki.esipfed.org/Attribute_Convention_for_Data_Discovery_1-3
[20] https://www.cmu.edu/iso/governance/roles/data-custodian.html
[21] https://oit.ncsu.edu/it-security/data-framework/data-management-procedures-summary-and-guidance/

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

data field (content and context), and the latter is responsible for the dataset environment (storage, transmission, safeguard).

Instead of looking at the different types of assessments separately, it is possible to perform overarching analyses encompassing all a dataset's facets. This is typically achieved using *maturity models*. These are formal approaches to support compliance verification, usually defined in discrete stages to evaluate practices applied in organisations, services or products. Maturity is meant as a desired or anticipated evolution from a more ad-hoc approach to a more managed process (Peng 2018). Datasets associated with high maturity are produced following best practices of the community and in a more managed fashion, increasing user trust in the data record provided. It should be noted that a low maturity rating does not necessarily imply low scientific value for a dataset. It can happen especially for datasets managed by a single investigator, which may be flagged as having low maturity due to inadequate quality in metadata or accessibility but still have a high scientific value.

ISO 19157 offers a distinct perspective to group types of quality assessment. These can be "direct", as based on inspection of the items or data values in the dataset or "indirect", as based on external knowledge associated with the dataset, such as lineage, documentation, or metadata. Direct evaluation is further classified by the source against which the assessment is done: "internal" if only the dataset under inspection is used during the evaluation or "external" if there is a reference to external data (benchmark). As illustrative examples, checking for duplicate records in the dataset is a way to assess the quality dimension of completeness using a direct and internal type of assessment. Checking for the misclassification rate of labelled items (e.g., water vs land) is a way to address the quality dimension of accuracy in a topographic dataset using a direct and external type of assessment.

2.8 Quality assessment workflow

To help practitioners address the challenge of where to start when curating and reporting dataset quality information, Peng et al. (2021) developed a typical workflow made of five steps:

- **quality specification**. Curating dataset quality information should start with defining the quality dimensions that will be assessed, the standards considered, determining the level of granularity (variable, ensemble member, model, or algorithm), and identifying which data and quality dimension should be prioritised. This step will need some profiling, i.e., an initial analysis of the available data to understand the challenges and the most critical issues to set priorities and determine the appropriate strategy to deploy. Quality dimensions and so requirements are determined by the system environment and the stakeholders (e.g., users, funding authorities);
- **evaluation specification**. This step involves identifying or developing an approach to evaluate the identified quality dimensions. The framework for the evaluation, including methods, protocols, and workflows with the responsible roles, is defined. Consider that many quality issues are solved only by having a cross-actor view, i.e., involving providers, users, management, etc. A well-documented approach to quality helps to increase transparency, verifiability, reproducibility, and resilience of the quality evaluation process;

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

- **evaluation execution.** During this stage, the actual assessments are performed based on the tools, approaches and priorities defined in the previous phases. It is recommended to produce a glossary aiming to establish the metadata used to achieve commonly understood definitions;
- **quality dissemination.** The results of the assessments represent the core of the dataset quality information and need to be disseminated with the data to benefit the users. For reproducibility purposes, the operations performed to produce the quality information are also recommended to be published. In this step, the way the quality information is disseminated (e.g., metadata, web page, KPIs) is implemented and put into practice. How well a user interprets dataset quality information depends on how it is presented. A consultation process with the users helps to select the most suitable dissemination format that meets consumer needs and points to dataset issues not identified so far;
- **monitoring and improvement.** The feedback collected in the previous step and the experience gained during the assessments are rationalised to consider improvements of the protocols, tools, and approaches and to redefine priorities in the assessment process. Setting performance indicators and scheduling regular reviews help to improve the curation of quality information.

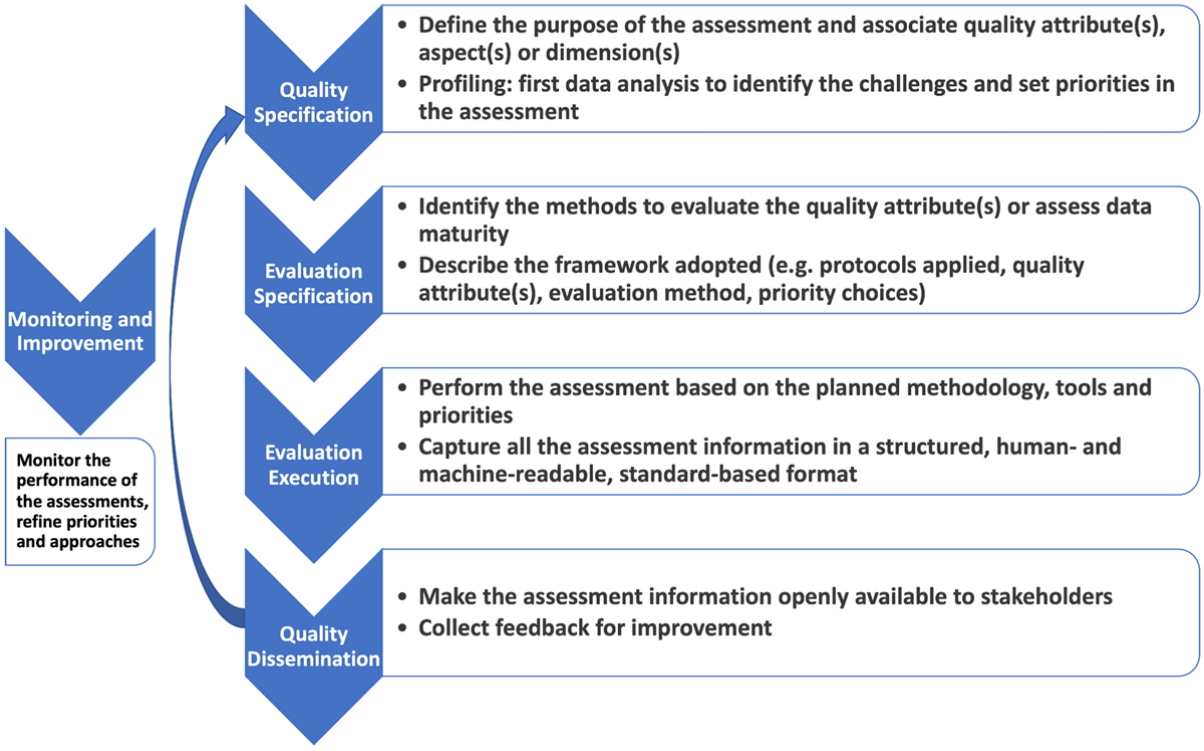

Figure *2.8.1.* *Credits to Peng et al. (2021). A schematic workflow and essential ingredients for curating and disseminating dataset quality information.*

There are a number of stakeholders to involve in the quality assessments conducted within a data service. The quality assessments are a multi-actor process between:

- the service provider (EOSC system here) reviews data for accepting/rejecting them in the

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

system based on well-defined requirements and guarantees the completeness of the dataset documentation;
- the data provider is in charge of the dataset rectification, performing the scientific quality assessments, and producing the dataset documentation requested by the service system;
- users provide feedback because they introduce the requirements and improve the dissemination of the quality information.

Minimum requirements

Quality specifications (the first workflow step in figure 2.8.1) rely on numerous dataset requirements. However, resource limitations (e.g., time, workforce) dictate a prioritisation of the requirements to be considered to specify the scope of quality. It is important to recognize which requirements are so fundamental that, when missing, the dataset is not usable/understandable and thus unservable. Fulfilling minimum requirements (MRs) guarantees a sufficient, but not necessarily optimal, quality of the dataset. While the definition and further assessment of all requirements are ideal, time and resources constrain the depth of the quality assessments. Therefore, establishing which requirements are more critical than others, i.e., prioritisation, is a way to define the scope of the evaluations with MRs characterising the first level of the assessment's depth.
The Bioimaging community offers an example where the latest efforts define tier levels for metadata standards[22], going from minimum information to advanced and complete. This is illustrated with a case study in section 3.5. Another example comes from the Earth science community with CMIP5 (Stockhause et al. 2012). Here quality is screened with increased detail (named level) depending on the requirements considered.

Several reasons may lead an organisation to assign priority to the requirements, ranging from user needs to time of assessment dedication. Automated assessments tend to be prioritised compared to human-based assessments because resources play a significant role when determining requirements priority. It is essential to evaluate human resources, technology capabilities, and capacity-building plans when deciding the sophistication of the quality assessments (Lacagnina et al. 2022). The audience's expertise can also determine the assessment depth: the lower the quality is evaluated, the more limited is the audience. This is the case of CMIP5 (Stockhause et al. 2012), where only a few selected users can access data associated with the lowest level of quality scrutiny.

2.9 Data quality management approaches

Usually, it is not hard to get everyone to agree that having data of excellent quality is important. However, when it comes to the essential questions about who is responsible for data quality (DQ), who must do something about it and who will fund the necessary activities, then the going gets tough. Quality management needs to be a mandatory piece of a larger governance strategy. Without it, an organisation will not successfully manage and govern data. *Data governance* is an authoritative function controlling the management and use of data within an organisation. It standardises how data is collected, stored, and analysed or disseminated for a specific use; it defines who can/must take what action, upon what data, in what situations, and using what methods. Data governance is

[22] https://quarep.org/special-issue-in-nature-methods-released/

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

beneficial to increase transparency around data and their availability, standardise data systems, policies, and procedures, resolve data issues, and ensure regulatory and organisational compliance. Data governance is about ensuring the organisation can maximise the value it gets from its data while controlling security and regulatory risks. Data governance is expected to bring together data providers and data users on the same platform, where policies, requirements, and rules are created and agreed on. Once you know how the data flows through the organisation and what the standards are, asking the DQ team to translate these standards into DQ rules and run them on the data in those systems is more streamlined. Indeed, DQ maintenance and improvement are the most significant driving forces behind data governance activities.

A standardisation of DQ approaches of the different research domains under the umbrella of the EOSC undoubtedly requires a governance structure that considers the diversity of disciplines on the one hand and the connecting aspects on the other. At the same time, governance will give this activity a unified voice to the outside world and help ensure that DQ is adequately considered in the constantly evolving research world over time. Different approaches to such a governance structure exist, as further elaborated in the sections on CTS (CoreTrustSeal) and INSPIRE crosswalk (see section 2.10). While CTS has a community-driven, i.e., bottom-up approach benefitting from the RDA activities, INSPIRE is a legal framework created and maintained by the European Commission. In the context of setting DQ processes in EOSC, it will be fundamental to analyse governance models of similar activities. This includes aspects such as efficiency, effectiveness, and inclusivity, as well as questions about how generic and specific competencies are mapped in governance. For example, will it make sense to have different advisory boards or just one central council? Which actors will be sought for participation? What voice should existing networks get?

Data quality approaches

Governance is expected to define the most appropriate approach to DQ. Over the years, several techniques for DQ assessment and improvement have been proposed. The DQ methodologies help select, customise, and apply such methods. These methodologies usually offer a systematic approach that, through the combination of relevant actors, processes, and technology, provides guidelines and procedures to detect relevant DQ issues, understand the causes and solve them. Most of the popular DQ methodologies, mainly designed to deal with relational databases, have been described and analysed by Batini et al. (2009). The methods they refer to are Total Data Quality Management (TDQM) (Wang 1998), Data Warehouse Quality Methodology (DWQ) (Jeusfeld et al. 1998), Total Information Quality Management (TIQM) (English 1999), A methodology for information quality assessment (AIMQ) (Lee et al. 2002), Canadian Institute for Health Information methodology (CIHI) (Long and Seko 2005), Data Quality Assessment (DQA) (Pipino et al. 2002), Information Quality Measurement (IQM) (Eppler and Munzenmaier 2002), Activity-based Measuring and Evaluating of product information Quality (AMEQ) (Su and Jin 2004), Cost-effect Of Low Data Quality (COLDQ) (Loshin 2001), Data Quality in Cooperative Information Systems (DaQuinCIS) (Scannapieco et al. 2004), Methodology for the Quality Assessment of Financial Data (QADF) (De Amicis and Batini 2004), Comprehensive methodology for Data Quality (CDQ) (Batini and Scannapieco 2016).

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

These methodologies have been compared in several aspects, such as methodological phases and steps, strategies and techniques, DQ dimensions, data type, and types of information systems considered by each methodology. The similarities and differences in these methodologies have led the authors to classify them into four categories:

- **complete methodologies** (e.g., TIQM, CDQ) support the assessment and improvement phases. These also provide guidelines to consider the DQ management initiative from an economic point of view. Particularly emphasised is the need to compare the cost of inferior quality with the costs and benefits of the DQ initiative. These cost/benefit analyses rule subsequent deployment of the DQ processes;
- **operational methodologies** (e.g., TDQM, DWQ) focus on the technical issues of both the assessment and improvement phases but lack an economic perspective. The operational methodologies specialise the DQ assessment on identifying issues for which their improvement approach works best;
- **audit methodologies** (e.g., AIMQ, CIHI) focus on the assessment phase and have minimal focus (if any) on the improvement phase. A more accurate assessment phase characterises the audit methodologies compared to the previous categories;
- **economic methodologies** (e.g., COLDQ) focus on defining and estimating the costs related to poor DQ and DQ investments. These methodologies elaborate detailed cost-benefit analyses and suggest the most appropriate improvement techniques using "data quality scorecards."

Batini et al. (2009) noted that many methodologies proposed in the literature are domain-specific (such as the DWQ designed for data warehouses or the CIHI focused on the healthcare sector), which limits their application. In fact, the selection of DQ dimensions and techniques is customised for the scenario and, therefore, is not applicable in general (Nikiforova, 2020). TDQM is a general-purpose methodology that offers a complete set of relevant dimensions and improvement techniques, which can be applied in various contexts. Complete methodologies, as the name suggests, provide a comprehensive framework for guiding and solving significant DQ problems in organisations that process critical data and give DQ a high strategic priority (e.g., banks and insurance companies). On the other hand, they significantly lack personalization for specific application domains seeking a high-level and context-independent approach. Completeness decreases as methodologies focus on specific contexts. The specialisation of operational methodologies reduces their applicability compared to complete methodologies while increasing the efficiency of the proposed techniques, particularly when focused on a specific DQ problem.

It shall be noted that the evolution of the approaches to DQ often reflects the development of ICT (Information Communications Technology) technologies. For example, the advent of the Web required the definition of new quality dimensions able to reflect the difficulties of accessing and controlling the data sources in this environment. The most recent approaches to DQ often focus on Web data and semi-structured and unstructured data that are increasingly used in the Big Data era. For example, there are DQ assessment methodologies proposed for Linked data (Debattista et al. 2016), crowdsourcing data (Meek et al. 2014), Big data (Taleb et al. 2015), or for supporting decisions in an industry context (Gunther et al. 2019). In addition to focusing on new data types, these approaches differ from the ones mentioned above because they emphasise the relevance of

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

information created during the DQ assessments and the related provenance, which witnesses all the operations performed during these assessments. The typical DQ workflow discussed in section 2.8 highlights the role of dissemination in quality information and is well aligned with the latest approaches to quality.

2.10 Certification

Certification is a third-party attestation that an object or activity meets specific requirements, the process of demonstrating that the object fulfils those requirements is called conformity assessment (ISO 17000). Third-party means that the conformity assessments or inspections (a.k.a. audits) are conducted by independent bodies grounded on community-recognized standards laid down by (usually external) authorities (Jahn et al. 2004). The third-party auditing organisations that check conformity and compliance against standards must have the technical competence and integrity (i.e., independence and impartiality) to conduct these assessment services. This is recognized with an accreditation attestation by a designating authority, an organisation usually established within the government or empowered by the government (ISO 17000). ISO 17000 provides the set of rules to be used by the accreditation bodies (note that ISO 17065 supersedes ISO guide 65). Because ISO only sets the standards and has no authority to control accreditation and certification activities, the ISO logo cannot be used in connection with certification or certificates, nor used on labels[23]. For an overview of the certification and accreditation bodies, see section 3.1.

Fundamental in the certification systems is the existence of *standards* boosting the production of state-of-the-art, high consensus deliverables widely accepted by the market (Regulation 1025/2012 on European standardisation[24]). The standards have no legal status, but they serve to establish a benchmark for audits (Jahn et al. 2004). In this respect, there are standardisation bodies (e.g., ISO, CEN, CENELEC, ETSI) networking with large numbers of technical experts from industry, associations, public administrations, academia, and societal organisations to develop widespread consensus. There can be more than one standard for the same process, product, or service. Jahn et al. (2004) explored the reasons for this differentiation. Important forces driving this differentiation are (i) the disparate target groups: based on the profile of the user, different standards may best suit the needs of the stakeholders; (ii) protectionism: standards are built up in various countries and regions to protect local producers; (iii) depth of coverage: some standards cover one step of the value chain, others several steps and others all the chain. It is easier to achieve a consensus on only one value chain level, as the interests of the parties involved tend to be more homogeneous. That is why only a few approaches include all stages. Once the root causes for the differentiation of standards have been identified, an intriguing question arises about why some standards are used and others not. To this end, this TF issued a survey (section 3.3) where this question was explicit.

[23] https://www.fao.org/3/y5136e/y5136e08.htm#TopOfPage
[24] https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32012R1025&from=EN

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

Figure 2.10.1. *Word cloud diagram showing the most frequent responses to the survey's question, "Why do you think some standards are used, and some are not?" Details about the survey are available in section 3.3.*

As shown in Fig. 2.10.1, we got several answers to the survey question concerning the importance of data standards. The most frequent ones rotate around these terms: "complexity of implementation," "no tools available," "unclear specifications," "not understandable," "too expensive," "no incentive of being compliant," and "demand-driven." Clustering the most frequent responses offers new insights: the complexity of implementation is alleviated by developing appropriate software tools. Indeed, software tools can produce standard expressions of the information they operate on; even complex standards can be made easy for users. Software tools make the specifications dictated in the standards transparent to the users, reducing the burden of their understandability. Reasoning in these terms, the remaining responses indicate that what pushes for adopting a standard is when the cost of its use is less than the benefit. In the medical industry, for instance, any company wishing to sell its own medicine in the European market must apply for marketing authorisation from the European Medicines Agency (EMA). Conformance to the stringent standards promoted by EMA requires additional work on top of the internal quality management processes performed in the company. Still, the need for market authorisation is an incentive for complying with the EMA standards. When market authorisation is not necessary, other incentives drive compliance with standards. It is the case of the food sector, where customers ask for specific quality-related labels, hence compliance with the associated standards, which enhances trust in and affiliation with the product (Jahn et al. 2005). Now, putting us in the case of standards that exhibit benefits outweighing costs, there may still be standards that are preferred to similar ones. Whether a standard becomes accepted will depend on many factors, including the public recognition of the authority setting the standard; the standard-setting process, especially stakeholder consultation; and the publicity around the standard.

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

Standards form the baseline for conformity assessments that lead to certifications. According to the EA (European co-operation for Accreditation)[25], certifications enhance confidence in product and service quality and reduce technical trade barriers. Adapting this concept to research and data, "trade" may be replaced with "interoperability" and "reusability." We may say that certification reduces barriers to the circulation of datasets because these are produced according to standards that make datasets easy to integrate. Conformance with these standards shall enhance the uptake of the supplied dataset. Providers and users should be fully informed about the commodities exchanged, datasets in our case. This exchange is often characterised by far-reaching information deficits and quality uncertainties from which the consumer is at risk, impeding a smooth exchange (Jahn et al. 2005) and market equilibrium efficiency[26]. This information asymmetry is reduced by issuing reliable quality indications using certification. In other words, certifications simplify the creation of standardised information exchange interfaces within the supply chain. Once awarded the certificate, the suppliers are entitled to make use of quality signals (e.g., labels, attestations) demonstrating the compliance of their products. In summary, the central task of certification is the reduction of information asymmetry between the data provider and the user. It ends with issuing a certification document, often synthesized in the form of a label recognized by the consumer as a quality surrogate.

The process for certification of a product is generally summed up in five steps, where we refer to CoreTrustSeal certification as an example (being in line with common trends followed by most certification bodies including, but not limited to, ISO), further elaborated in the next section:

- **application**: the organisation wishing to get certified makes the request. In the example of the CoreTrustSeal certification scheme, the organisation also submits the documentation requested to verify conformity to the requirements;
- **evaluation**: the auditor reviews the product to check whether it meets the qualification criteria. In the case of the CoreTrustSeal certification, this is done by external reviewers selected by the CoreTrustSeal board;
- **decision**: the certification body grants the certification or asks the applicant for improvements before final acceptance;
- **surveillance**: the certifier checks periodically that the product continues to meet qualification criteria. This is not performed in the case of the CoreTrustSeal. However, the CoreTrustSeal application documents are publicly available after certification so that users can verify compliance with the criteria at any time;
- **recertification**: the certification may expire after a certain period. It is the case of the CoreTrustSeal certification expiring after three years.

[25] https://european-accreditation.org/
[26] https://www.fao.org/3/y5136e/y5136e07.htm#fnB19

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

Case study - Standards in CoreTrustSeal certification and INSPIRE directive
CORE TRUST SEAL (CTS)

The term “Core Trust Seal” is used with two different meanings. On the one hand, it denotes an international, community-based, research domain-independent, non-governmental and non-profit certification organisation. On the other hand, a CoreTrustSeal certification is offered by this organisation. This certification is based on a catalogue of data repository requirements criteria, which reflects trustworthy data repositories' core characteristics. Both criteria and organisation have been established, under the umbrella of the Research Data Alliance (RDA), in a joint effort between the ICSU World Data System (ICSU-WDS) and the Data Seal of Approval (DSA).
The CoreTrustSeal (CTS) organisation has become a legal entity under Dutch law. CTS is governed by a Standards and Certification Board, which is, inter alia, responsible for developing and maintaining the CTS certification criteria and processes. The Assembly of Reviewers elects CTS members. The participation of the community is essential in the development process, as they are continuously called upon to make suggestions regarding the criteria as well as the entire process of certification. CoreTrustSeal Requirements are evolving and updated every three years.

At the time of writing (2022), the sixteen requirements are split into three topic areas (see figure 2.10.2): Organisational Infrastructure, Digital Object Management, and Technology (CoreTrustSeal Standards and Certification Board. 2019).

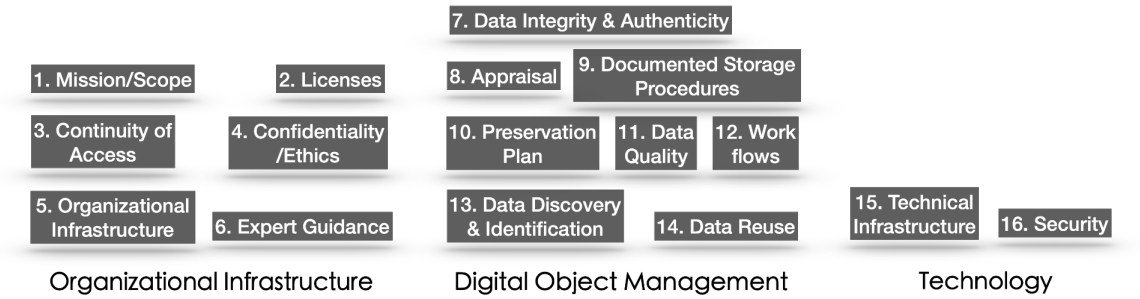

Figure 2.10.2. *The allocation of the 16 CTS requirements towards the topic areas: Organisational Infrastructures, Digital Object Management and Technology.*

The applications for certification, which consist of a self-assessment, are reviewed by multiple members of the so-called Assembly of Reviewers. In addition to the documents submitted, their review is based on publicly accessible records that provide evidence for the respective criteria. The review process is explicitly designed to enable an exchange between the reviewers and the repository. This setup forms an inclusive process that gradually leads to the performance improvement of repositories and, thus, to the preservation of the initial investments in collecting data.

INSPIRE

The INSPIRE conformity assessment procedure is an interesting example in the context of our TF because of the following characteristics: range of thematic data, operative obligation on data and data services of EU-member states, and the comprehensive application of different existing

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

standards. As an EU directive entered into force in 2007, INSPIRE[27] aimed to enhance the sharing of environmental spatial information among public sector organisations and facilitate public access to environmental information across Europe.

Based on a collection of established standards (W3C, ISO/TC211, OGC, etc.), the INSPIRE Technical Guidelines (TGs) describe the legal requirements for data and data services and recommendations on implementation. The INSPIRE directive is underpinned by technical Implementing Rules (IR) as a legal framework for a European Spatial Data Infrastructure (SDI). It enables spatial information to be interoperable, discoverable, viewable, and downloadable to assist in policy-making across boundaries. In addition to the IRs, detailed TGs are dedicated to each of the 34 INSPIRE Themes and consider the characteristics of the different domains.

Three central infrastructure components were developed to ensure streamless data access: the INSPIRE Geoportal, INSPIRE Validator and INSPIRE Registry (figure 2.10.3). The INSPIRE Geoportal harvests national catalogues and evaluates the metadata compliance and described data service (OGC conformal View-, Catalogue-, Download-Services).

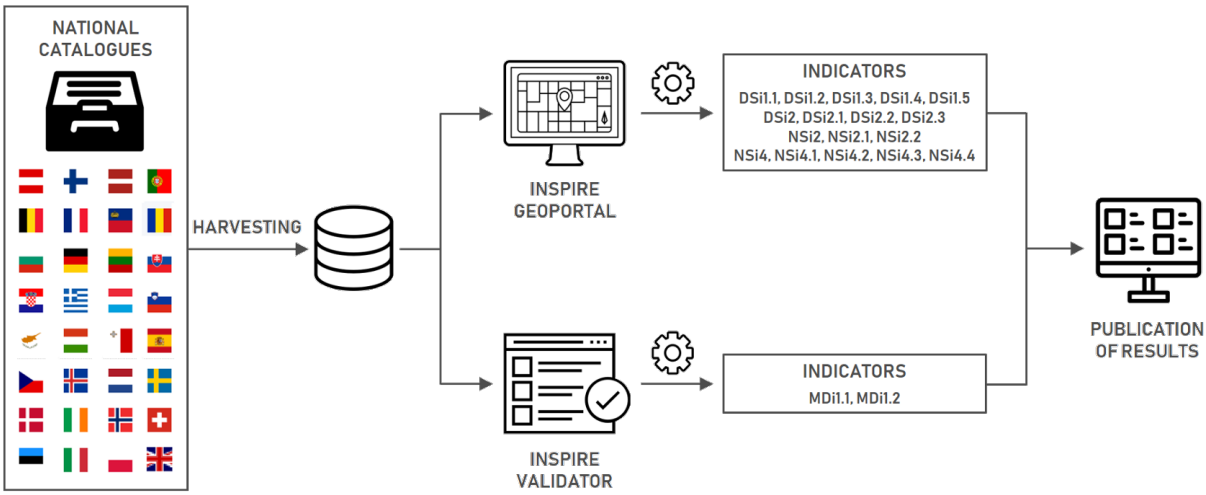

Figure 2.10.3. *The evaluation of metadata and data service compliance embedded in an on-demand monitoring and reporting process. Abbreviation of the indicators pointing to the scope of the content of measures: MD = Metadata, DS = Data Services (FeatureTypeCatalouge), NS=Network Services (OGC based, like View-, Download, etc.) (source: EC-JRC 2022).*

The conformity assessment of data content is proceeded via an ATS (Abstract Test Suite). The ATS is a check against the requirements defined as conformance classes based on the developed INSPIRE data models and application schemas[28]. With the development of the INSPIRE Testing Framework and the Executable Test Suites (ETS) in 2020, a web-based assessment tool is fully operational[29]. Metadata elements, Feature Types of the thematic data models and their constraints,

[27] http://inspire.jrc.ec.europa.eu/
[28] https://inspire.ec.europa.eu/portfolio/data-models
[29] https://inspire.ec.europa.eu/validator/home/index.html

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

and the implementation of the code list (INSPIRE Registry[30] & Registry federation[31]) are considered in the validation mechanism, evaluated for compliance, and issued as a report.

Conclusions

In contrast to the CoreTrustSeal, INSPIRE does not aim to assess the characteristics of a trustful repository in an explicit sense. Indeed, INSPIRE's objective is to set up a streamless spatial data infrastructure with interoperable data access and reliable data.
However, both examples are characterised by an open review process, the involvement of experts, and multidisciplinary requirements that lead the way for similar practices to be adopted in EOSC. This TF performed a preliminary mapping between CTS and INSPIRE requirements and extracted commonalities valid to the several dataset types in EOSC, as outlined in section 3.4.

2.11 Quality indicators

Since quality dimensions are rather abstract concepts (section 2.3), the assessment metrics rely on *quality indicators* (QIs) that allow the assessment of the quality of a data source w.r.t. the criteria. An assessment score is computed from these indicators using a scoring function (Zaveri et al. 2016, Parmelli et al. 2021). QIs are characterised by thresholds, ranges or other targets permitting comparisons and evaluations of the performance of the same quality aspect in different datasets. These characteristics make QIs significant assets to monitor, evaluate, and eventually improve the quality of the particular aspect they were designed for. Thus, QIs must serve as input for an action plan. Elements to consider when creating a QI:

- define monitoring frequency;
- benchmarks or thresholds for acceptable performance;
- measure quality aspects as rates or fractions;
- graphic displays are an effective mechanism for presenting the indicator's values;
- variations in quality indicators exceeding defined thresholds should trigger an investigation to identify the root cause of the variation.

QIs may indicate that a specific quality requirement of the product has been successfully met. This achievement can sometimes be indicated with labels signalling the dataset's compliance with defined requirements. In the same way, as it happens for certifications (see section 2.10), labels simplify the creation of standardised information exchange about the product quality status that can be easily compared across similar datasets. To do this, labels must be visually striking, convey information quickly and intuitively, and be easy to recognize and understand. Most labels rely on graphics to draw attention and communicate information quickly and memorably. There are three general approaches to these graphics - one based on the range of quality of the products available, another based on predefined quality categories and the last based on progress towards a target level (IEA 2000).

An example of QI is NUSAP (van der Sluijs et al. 2005). NUSAP (Numeral, Unit, Spread, Assessment, Pedigree) is an analytical tool based on a specific notational system for multidimensional

[30] https://inspire.ec.europa.eu/registry
[31] https://inspire.ec.europa.eu/register-federation/

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

uncertainty assessment and classification. It helps users to prioritise uncertainties according to quantitative and qualitative metrics for the limitations of underlying models or observations. The pedigree aspect can also be extended to cover more abstract areas of uncertainty (e.g., assumptions and socio-economic influences). The European Food Safety Authority (Bouwknegta and Havelaara 2015) and the Netherlands Environmental Assessment Agency (van der Sluijs et al. 2015) are notable users of the NUSAP system. By design, NUSAP requires a manual selection and assessment of the relevant documentation. It only provides the user with a standardised approach to aid the analysis and evaluation.

A dataset having successfully passed the minimum target set by the QI may be labelled with a mark alerting the stakeholders about the dataset quality. In our survey (details in section 3.3), we explored community interest in having any indicator or ranking system to apply to the datasets in EOSC. The results showed a striking preference for no ranking rather than ranked quality should be documented. If a hierarchy must be used, then the priority should be placed on ranking the FAIRness level of the datasets. Assessment of the data content, a.k.a. scientific assessment (see section 2.7), should be avoided. 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 assessments should be accompanied by the methodology applied and by the profile of the authors. The methodology must be the same for similar datasets.

2.12 Vocabulary challenges

A wide range of independent terminology resources across research domains and communities has emerged, with over eight hundred across all disciplines (FAIRsharing 2022). The lack of a common strategy to define vocabulary concepts makes them not directly comparable with the consequence of uncertain re-use. Consequently, the data annotated with these concepts are not interoperable in a sustainable way and requires substantial manual efforts to integrate larger terminologies. In other words, our ability to exploit these data as a common resource is hampered by a lack of interoperability in how researchers describe data variables in a common expression (David et al. 2021). The proliferation of semantic resources poorly aligned is a source of confusion for users (Magagna et al. 2021).

In any process of sharing information, the sender and the receiver need to use common codes and signifiers. For instance, if a researcher were to filter a database using the word “female,” records that refer to females but have “f.” as a sex value would not be returned. Also, even if they retrieved all records to avoid that gap, they would currently have to determine what “f.” stands for manually. Other examples of mechanisms of discrepancies are reported in figure 2.12.1 (David et al. 2022). According to Cox et al. (2021), ontologies are often not machine readable (some are available in PDF format) or do not adhere to the FAIR principles (for instance, some contain two or more terms with identical or similar definitions). The lack of common vocabulary concepts renders data without apparent reference less discoverable and challenging to use.

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

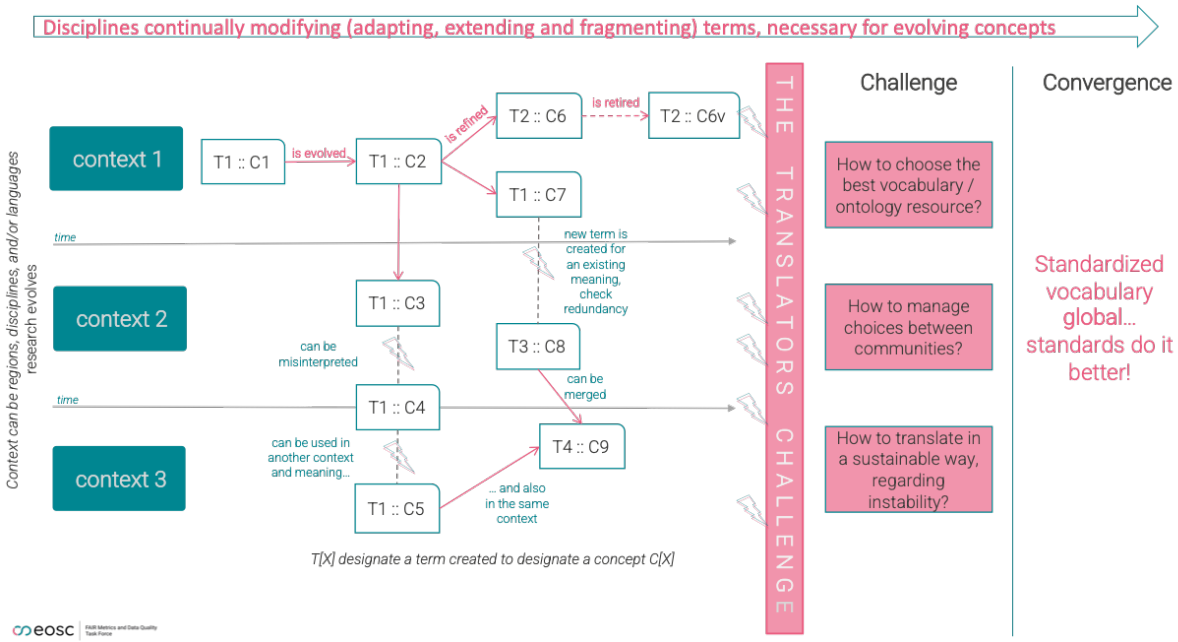

Figure 2.12.1. *Discrepancies between regions and groups (culture, content, workflows, language, semantics, translation, funds…). Typical issues relate to polysemy (one term, multiple meanings), confusion (multiple terms, one meaning; or 'false friends' between two languages), plus existing and evolving nuances (inexact matches both between languages and over time). Furthermore, terms are often adopted from another language with different contexts and disciplinary realms (that might decrease interoperability) and impede the translation of all versions simultaneously. Specifically, the critical point is that translation occurs at the concept level, not as a simple one-to-one translation of (consecutive) words. Modified according to David et al. (2022).*

Without common rules and provenance models to enable the mapping of community vocabularies and the capture of information in myriad ways, there is a risk of being incomplete and inaccurate in transmitting the information. Certain meta-information, like the source of the definition of a term, the reason for adaptation for a definition, the contributor and a timestamp, are at least basic information to keep provenance information. If a particular value does not refer unambiguously to a specific concept, it creates a barrier to assessing whether a record containing such a value is dependable. In this context, the construction and use of vocabularies represent a significant matter of data quality (Chapman et al. 2020 and Cox et al. 2021).

Finally, governance is gaining momentum in the last few years. Similarly to how W3C works with web languages (e.g., HTML, XML, CSS validators), semantics for research data should be evaluated and certified by recognized authorities to achieve comparability and findability of semantic resources. For example, the ISO Technical Committee (TC 211- Geographic information) developed "ISO 19135 - Procedures for item registration", a maintenance model for controlled vocabularies, where roles of responsibilities were determined, as shown in figure 2.12.2.

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

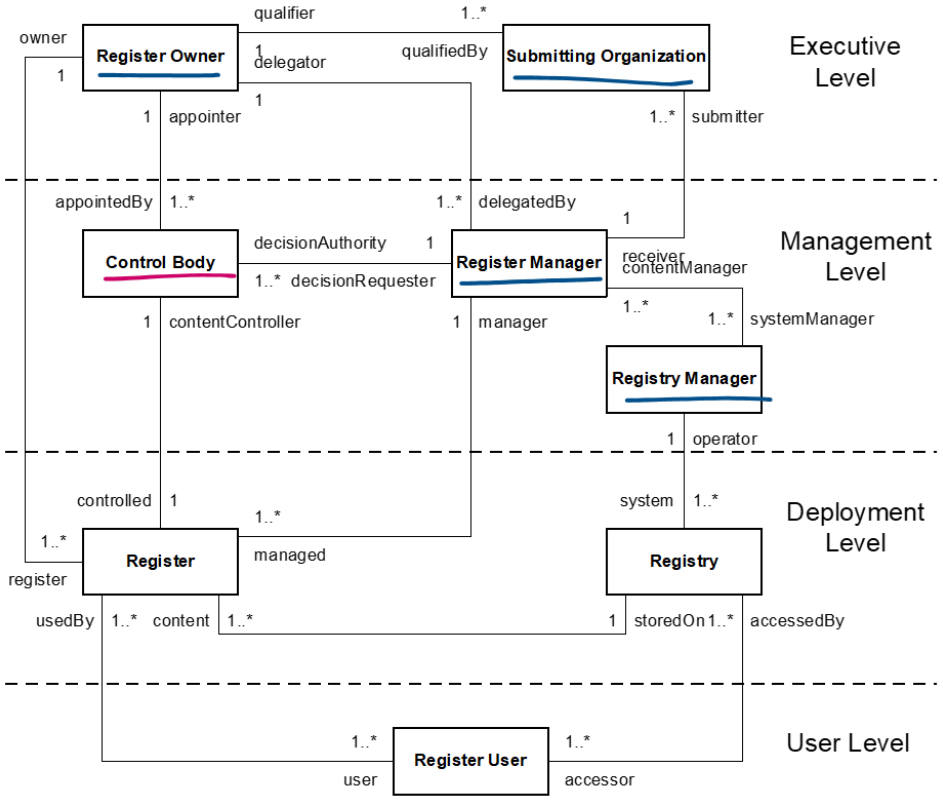

Figure 2.12.2. *Levels of responsibility of stakeholders in the ISO 19135 maintenance model. Regarding Data Quality for vocabularies, the Management Level and Executive Level mirror the needed interaction within a governance process. The Registry Owner is an organisation that establishes a register, which is a set of files containing identifiers assigned to items. The Submitting Organisation gets abstract permission to deal (publish) with the register and delegates tasks to the Management Level for the Registry as Software Framework and the Register itself. The Control Body is usually the instance to make decisions about new concepts, change proposals, corrections, semantic relations, etc. ISO 19135-1:2015 Geographic Information — Procedures for item registration — Part 1: Fundamentals, 2015*

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

3. LANDSCAPE, CHALLENGES, AND REQUIREMENTS

3.1 Landscape of the leading organisations in data quality

Most domains identified their specific criteria starting from general principles and set the standards, guidelines, or recommendations for data quality by bringing together domain experts, initiatives, and projects. Identifying the actors involved in the quality assessments and organisations seen as a reference point in various disciplines facilitates the collection of community requirements and the identification of standards to follow. The resulting overview is far from exhaustive, but it gives a flavour of the complexity of the fragmented landscape. The relevant initiatives and organisations can be clustered under five broad headings:

- **providers of standards** with national or international acceptance: organisations that develop and maintain standards for data gathering and measurement in particular technical areas. Examples include the International Standards Organisation (ISO) and the European Committee for Standardisation (CEN);
- **providers of accreditation/certification**: certification bodies issue certificates recognizing that a product complies with specific standards; in turn, accreditation bodies accredit the certification bodies as being able to issue trustworthy certificates (see also section 2.10). Examples of accreditation bodies include the International Laboratory Accreditation Cooperation (ILAC) and the International Accreditation Forum (IAF);
- **providers of reference data**: organisations that deliver data deemed as a reference in particular fields. An example is the International Union of Pure and Applied Chemistry (IUPAC), which holds databases of reference data for chemical analyses;
- **providers of good practice and guidelines**: organisations that deliver guidance, tools, templates, and recommendations but no formal standards. An example is *OpenAIRE*, which (amongst other services) provides guidance documents and training to help researchers make their data FAIR and shareable. Another example is the *Digital Preservation Coalition (DPC),* providing resources to educate various public and private entities on the best practices for long-term digital preservation, or the *International Committee for Documentation (CIDOC),* providing the museum community with advice on good practices and developments in museum documentation;
- **discussion groups**: initiatives that function as the discussion for promoting the adoption of good practices or standards. Members discuss and define which criteria and aspects of data quality are essential in their field. Just a few examples to show the massive number of bodies entering this group: *International Geothermal Association (IGA)*, *European Energy Research Alliance (EEEA)*, *INetQI*, *International Energy Agency (IEA)*, *International Council of Museums (ICOM)*, *Alliance of Digital Humanities Organizations (ADHO)*, *International Federation of Library Associations and Institutions (IFLA)*, *International Organization of Legal Metrology (OIML)*.

This section focuses on the first three categories, where organisations are at the forefront of defining data quality standards. The organisations belonging to the last two categories are not discussed due to their diversity and number, making their full description not feasible within the scope of this document.

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

Providers of Standards

Growing international exchange stimulated the development of internationally recognized standards, which led to the *International Organisation for Standardisation (ISO),* established in 1947. ISO is a network of national standards institutes, one member per country, with a Central Secretariat in Switzerland that coordinates the system. ISO is a non-governmental organisation that creates a bridge between the public and private sectors. On the one hand, many of its member institutes are part of the governmental structure of their countries or are mandated by their government. On the other hand, other members have their roots uniquely in the private sector, having been set up by national partnerships of industry associations. The ISO standards that are directly related to data quality are reviewed in section 3.2. Other prominent international standardisation bodies often collaborating with ISO are the *International Electrotechnical Commission (IEC)* and the *International Telecommunication Union (ITU)*. The former covers a vast range of electrotechnologies (e.g., power generation, transmission and distribution to home appliances and office equipment, semiconductors, fibre optics, batteries, solar energy, and nanotechnology), the latter focuses on information and communication technologies (e.g., radio spectrum, satellite orbits, broadband Internet, wireless technologies, aeronautical and maritime navigation, and TV broadcasting).

Similar bodies exist at the continental and national levels. In the case of Europe, there are three standardisation organisations: (i) *European Telecommunications Standards Institute (ETSI)* is the recognized regional standards body dealing with telecommunications, broadcasting and other electronic communications networks and services; (ii) *European Committee for Electrotechnical Standardization (CENELEC)* prepares standards that cover the electrotechnical area, (iii) and *European Committee for Standardization (CEN)* produces standards in areas such as air and space, chemicals, construction, consumer products, defence and security, energy, the environment, food and feed, health and safety, healthcare, ICT, machinery, materials, pressure equipment, services, smart living, transport and packaging.
Professional associations also define standards, typically for technical measurements and other procedures relevant to their members. An example is the *American Society of Mechanical Engineers (ASME)*. Technical committees develop the standards created by these bodies made up of experts in their field who reach a consensus on what process to follow to achieve target requirements and document that process. Similarly, many worldwide bodies, such as the *World Meteorological Organisation (WMO)* and the *World Health Organisation (WHO),* issue data collection and reporting standards that ensure that the resulting datasets have the aspects of data quality that are most important for their applications. Equivalent approaches are also present at the continental level, an example being the *European Food Safety Agency's* (*EFSA*) standardised system for classifying and describing food, FoodEx2.

Many communities have developed their recommendations for data and metadata, which have often evolved to become de facto standards through their breadth of adoption. Examples include:

- the *International Virtual Observatory Alliance (IVOA)* develops standards to ensure the interoperability of astronomical data collected by its members, enabling the international utilisation of astronomical archives as an integrated and interoperating virtual observatory;
- the *Spectrum Community* develops and maintains a standard for collection management

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

activity in museums;
- the *Biodiversity Information Standards (TDWG)* develops and maintains open standards for biodiversity data, including the Darwin Core standard;
- the *Health Level Seven International (HL7)* is dedicated to standards and solutions that empower global health data interoperability;
- the *Text Encoding Initiative (TEI)* develops and maintains a standard for the representation of texts in digital form;
- the *Open Geospatial Consortium (OGC)* addresses interoperability challenges for the geospatial community, such as publishing map content on the Web, exchanging critical location data during disaster response & recovery, and enabling the fusion of information from diverse Internet of Things (IoT) devices;
- the *Internet Engineering Task Force (IETF)* is an open standards organisation that develops and promotes voluntary Internet standards, particularly the technical standards that comprise the Internet protocol suite (TCP/IP).

In some cases, these standards become ISO recognised; for instance, ISO has recognised the DICOM standard for medical imaging and related data as the ISO 12052 standard. Others, such as the Resource Description Framework (RDF), find use beyond their initial purpose. The RDF was initially developed as a data model for metadata, but it has come to be used as a general method for describing and exchanging graph data.

Providers of Accreditation/Certification

There are many certification and accreditation bodies; here, we limit ourselves to mention those most known in scientific research.
Among the accreditation bodies, the international organisation for laboratory accreditation bodies operating in calibration, testing, and proficiency testing according to ISO/IEC 17011 is called *ILAC (International Laboratory Accreditation Cooperation)*. It maintains a mutual recognition agreement to promote the acceptance of technical tests and calibration data for exported goods. A similar body, the *International Accreditation Forum (IAF)*, brings together bodies interested in conformity assessment in management systems, products, processes, services, personnel, validation, and verification. These bodies have their equivalents at the continental and national level; for instance, the *European co-operation for Accreditation (EA)* has members across Europe operating a peer evaluation scheme to ensure compliance of the EA national accreditation members to European regulations international standards. The EA is a member of ILAC and IAF, as sketched in figure 3.1.1.

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

Figure 3.1.1. *The accreditation landscape: the first three lines regard accreditation bodies at various levels of geographical recognition. This cascade ends with the certification bodies.*

The services ILAC and its members provide are relevant to data quality because their work underpins the traceability of measurements back to national standards. Most countries have a *National Measurement Institute (NMI)* or equivalent Designated Institute (DI) that is responsible for holding national standards for (some of) the seven International System of Units (SI) base units and for disseminating them within their country. This dissemination is achieved through a traceability pyramid, which involves accredited laboratories as a vital step between national standards and end users. The pyramid means that the measurements made in research and industry can be traced back to national standards through a chain of trusted accredited measurements. The NMIs and DIs are accredited for specific measurement processes by their national bodies but also participate in international peer review processes and intercomparisons with their equivalents in other countries. The definition of the base SI units disseminated by NMIs and DIs, and a mutual recognition agreement of measurement capabilities, are agreed internationally during the *General Conference on Weights and Measures (CGPM)*, based on technical information provided by the *International Committee for Weights and Measures (CIPM)*. CGPM consists of delegates of the governments of its member states, and CGPM elects CIPM members. CIPM runs several consultative committees comprised of technical experts from DIs and NMIs to obtain the technical information it provides to CGPM.

As far as the certification bodies are concerned, the leading certification body for research repositories is the *CoreTrustSeal foundation* (details in section 2.10). This organisation was established in 2018 as a single independent, international, community-based, non-governmental, and non-profit foundation promoting sustainable and trustworthy data infrastructures. The foundation is truly global and interdisciplinary, including members from social and political sciences,

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

geomagnetism and seismology, astronomy, Earth sciences, crystallography, medical sciences, economics, psychology, and archaeology[32]. Nestor (Dobratz and Schoger 2007) offers an alternative scheme. This competence network for digital preservation provides a certification process for comprehensive self-evaluation based on the DIN 31644 "Criteria for trusted digital repositories". This scheme was initially set up by a German Federal Ministry for Education and Research funded project and has been operated by the original project partners since 2009. Whilst its members are primarily based in Germany, it has links to similar initiatives in other countries, such as the *Digital Preservation Coalition (DPC)*, a British coalition for long-term digital preservation.
Since certification is characterised by a third-party attestation by an external audit body (details in section 2.10), the auditors must have the requisite skills and ability to perform the audit effectively. Qualification as an auditor may align with internationally accepted standards, e.g., ISO 19011:2018, International Standards on Auditing (ISA) issued by the International Auditing and Assurance Standards Board (IAASB). The International Organization of Supreme Audit Institutions (INTOSAI) is the most crucial intergovernmental organisation in the audit field. One of the regional working groups of INTOSAI is the European Organization of Supreme Audit Institutions (EUROSAI), comprising the audit institutions of the member states plus the European Court of Auditors.

Providers of Reference Data

Other bodies provide reference datasets or conduct monitoring of data quality, typically within a specific domain of science and sometimes gathered from a particular country or continent. Many of the examples given are either funded directly by individual governments or the EU or are well-established international bodies whose work is indirectly supported by governments through member bodies:

- The *WIGOS Data Quality Monitoring System*: monitors the availability, quality, and timeliness of data from the WMO International Global Observing system. WIGOS ensures data and products are dependable and correspond to an agreed set of needs;
- The *International Union of Pure and Applied Chemistry (IUPAC)* reviews determinations for atomic weights of chemical elements and has several databases of reference data relevant to chemical analysis calculations;
- The *International Science Council Committee on Data (CODATA)* has a Task Group on Fundamental Physical Constants that periodically provides the scientific and technological communities with a self-consistent set of internationally recommended values of the fundamental constants and conversion factors of physics and chemistry based on all relevant data available at a given point in time;
- The EMBL's *European Bioinformatics Institute (EBI)* holds freely available molecular data resources compiled in collaboration with international partners.

[32] https://www.coretrustseal.org/about/assembly-of-reviewers/

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

3.2 Overview of the relevant ISO standards

Standards can be of different types: basic, terminology, testing, product, process, service, interface (ISO Guide 2:2004), and provide three essential benefits (International Energy Agency):

- a consistent and transparent framework describing technologies and good practices in the fields concerned, including, inter alia, terminology, classifications, test methods, performances (along with the modalities of the presentation of test results and performance levels) and good management practices;
- state-of-the-art knowledge formalised by recognized experts in the field, based on international consensus from a balance of interests reflecting the technological, economic, and public interest conditions in most of the countries of the world;
- tend to reduce uncertainty for all the players, supporting international trade of goods and services, developing new markets, and helping to improve consumer/user understanding and confidence, thus influencing consumer/user behaviour and choices.

These standards are tailored to the different domains, leading to many criteria. Just a few examples to show the wealth of standards available across disciplines: e.g., MIAPPE for plant phenotype, Web Map Tile Service (WMTS) for georeferenced maps, Journal Article Tag Suite (JATS) used to describe scientific literature published online, Fast Healthcare Interoperability Resources (FHIR) for electronic health records, Europeana Data Model (EDM) for the cultural heritage sector, Metadata Object Description Schema (MODS) for bibliography, Metadata Encoding and Transmission Standard (METS) for digital libraries, CIDOC Conceptual Reference Model (CRM) for cultural heritage, Climate and Forecast (CF) in climatology, Flexible Image Transport System (FITS) in astronomy, Brain Imaging Data Structure (BIDS) for neuroimaging experiments, and Public Participation in Scientific Research (PPSR) Core in citizen science.

A comprehensive overview of the wide variety of standards available in all disciplines is a work in progress by FAIRsharing (2022), which maps the landscape of data and metadata standards, interlinking them (e.g., terminologies to models and minimal information guidelines) and repositories and policies that implement and recommend them, as relevant. Here we limit ourselves to exploring the standards directly related to quality coming from the most prominent standardisation body, ISO. This body is further described in section 3.1.

ISO 900n: Quality management

The ISO 9000 family is the most widely recognized of the many standards issued by ISO. Given its scope and popularity, it serves as a knowledge base for many other standards, including but not limited to those dealing with data quality, adapted to specific sectors and industries, including medical, automotive, software engineering, petroleum, aerospace, petrochemical and natural gas, railway, etc. areas.
It can be used by any organisation (regardless of its size or its field of activity) and therefore is stated to be a universal standard. It is a quality management standard and not a product quality standard. Consequently, it does not address the quality of products but defines requirements related to the organisation's management structure. ISO 9001 defines and relies on seven domain-agnostic principles for establishing and maintaining quality (see figure 3.2.1). In other words, this process-

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

based standard specifies the terms and definitions that apply to all quality management standards developed by ISO/TC 176 (i.e., Technical Committee responsible for Quality management and assurance).

Figure 3.2.1. *ISO 9001:2015. Principles apply to the Quality Management Standard (source: https://www.smartsheet.com/ultimate-guide-iso-9000).*

The requirements defined in ISO 9001 for products and services refer to[33]:

- customer communication and respective feedback received as a result of this communication leads to defining general requirements according to the customers' needs and expectations (it addresses questions like "Do we understand what the customer wants?", "Are we able to do the job on time and in full?", "Has the customer changed their requirements?");
- determining the requirements for products and services, where it is expected that the

[33] ISO 9001 - Clause 8.2: What are the Requirements for Products and Services? (iso-9001-checklist.co.uk)

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

organisation implements a process to define the needs for the products or services that it intends to offer to customers. It may include the requirements from interested parties and statutory or regulatory requirements relating to the product;

- reviewing the requirements for products and services, which refers to a review of customer's needs ensuring (a) product requirements are defined, (b) product requirements are agreed upon, (c) any amendments to the specifications are agreed upon, (d) any amendments to the specifications are communicated, (e) the organisation can achieve the stated requirements.

ISO 9001 is built on the Deming Cycle (figure 3.2.2), where the set of "plan", "do", "check", and "act" activities are taken, thereby contributing to continuous improvement, one of the principles defined in ISO 9001. More precisely: (1) *plan* - establish the strategy and implementation activities as necessary to deliver results according to the data requirements; (2) *do* - perform the processes; (3) *check* - monitor and measure data quality and process performance against the strategy and data requirements, while reporting the results; (4) *act* - based on the previous step results, take actions to improve the process performance.

Figure 3.2.2. *The Deming Cycle adapted in ISO 9001:2015 (source: https://the9000store.com/iso-9001-2015-requirements/).*

Over one million companies and organisations in over 170 countries are certified to ISO 9001. It is the only standard in the ISO 9000 family that can be certified as it is the only one having specific requirements; the other standards in this family provide complementary / clarification material supporting the 9001 implementations. The certification is granted for three years and must be recertified afterwards to maintain the certification status. Thus, the requirements imposed by ISO 9001 should be regularly maintained and brought to the organisation under question.

ISO 250nn: Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE)

Another standard relevant in this document refers to the ISO 25000 family "Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE)", where ISO 25012 "Data quality model" is of the highest interest for our TF. It defines the fifteen quality dimensions considered; see figure 3.2.3. In this ISO, dimensions are named characteristics (section

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

2.3 describes the concept of dimensions in data quality).

Characteristic	Description
Inherent	
Accuracy	The degree to which data has attributes that correctly represent the true value of the intended attribute of a concept or event in a specific context of use.
Completeness	The degree to which subject data associated with an entity has values for all expected attributes and related entity instances in a specific context or use.
Consistency	The degree to which data has attributes that are free from contradiction and are coherent with other data in a specific context of use.
Credibility	The degree to which data has attributes that are regarded as true and believable by users in a specific context of use.
Currentness	The degree to which data has attributes that is of the right age in a specific context of use.
Inherent and system dependent	
Accessibility	The degree to which data can be accessed in a specific context of use, particularly by people who need supporting technology or special configuration because of some disability.
Compliance	The degree to which data has attributes that adhere to standards, conventions or regulations in force and similar rules relating to data quality in a specific context of use.
Confidentiality	The degree to which data has attributes that ensure that it is only accessible and interpretable by authorized users in a specific context of use.
Efficiency	The degree to which data has attributes that can be processed and provide the expected levels of performance by using the appropriate amounts and types of resources in a specific context of use.
Precision	The degree to which data has attributes that are exact or that provide discrimination in a specific context of use.
Traceability	The degree to which data has attributes that provide an audit trail of access to the data and of any changes made to the data in a specific context of use.
Understandability	The degree to which data has attributes that enable it to be read and interpreted by users, and are expressed in appropriate languages, symbols and units in a specific context of use.
System dependent	
Availability	The degree to which data has attributes that enables it to be retrieved by authorized users and/or applications in a specific context.
Portability	The degree to which data has attributes that enable it to be installed, replaced or moved from one system to another preserving the existing quality in a specific context of use.
Recoverability	The degree to which data has attributes that enable it to maintain and preserve a specified level of operations and quality, even in the event of failure, in a specific context of use.

Figure 3.2.3. *The data quality characteristics or dimensions as defined in ISO/IEC 25012.*

ISO 25012 classifies the 15 characteristics or dimensions considered in two categories: (1) Inherent Data Quality, where the data quality refers to (1a) data domain values and possible restrictions (e.g., business rules governing the quality required for the characteristic in a given application), (1b) relationships of data values (e.g., consistency), (1c) metadata, and (2) System-Dependent Data Quality, referring to the degree to which quality is reached and preserved within a computer system when data are used under specified conditions, thereby depending on the technological domain in which data are used, although overlap of both is also possible (see figure 3.2.4).

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

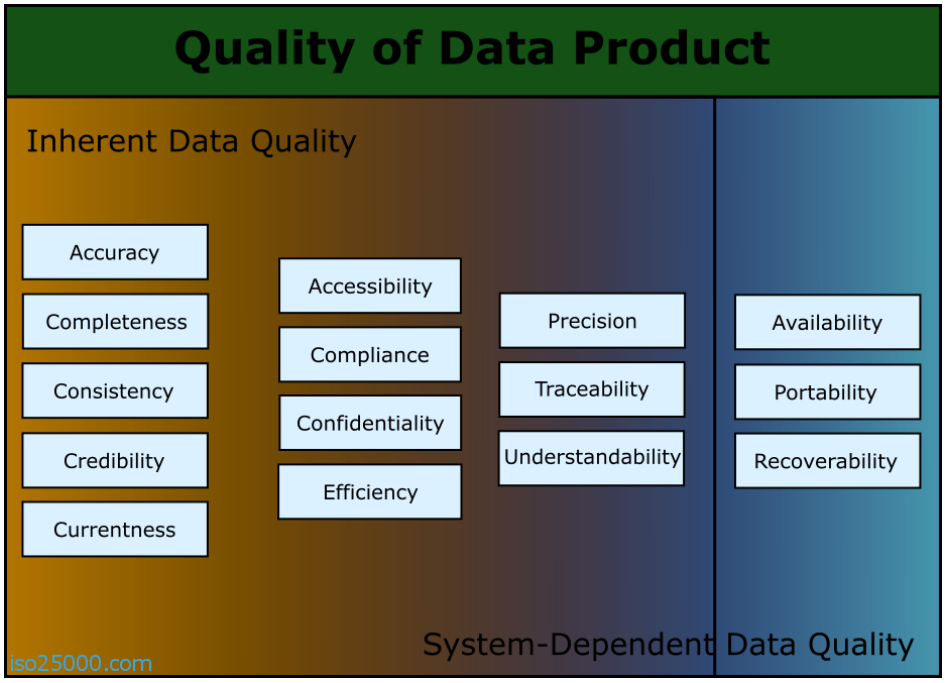

Figure 3.2.4. *The picture shows the fifteen quality characteristics or dimensions relevant to the data quality model defined in the ISO/IEC 25012 standard. Note the classification of the quality dimensions in two main categories: inherent data quality and system-dependent data quality (source: https://iso25000.com/index.php/en/iso-25000-standards/iso-25012/136-iso-iec-2012)*.

ISO 191nn: Geographic information/Geomatics and Geographic information

Compared to ISO900n, the 19100 family provides a more detailed understanding of data quality measurement and assessment processes tailored for digital geographic information, where ISO 19157:2013 ("Geographic information - Data quality") is of most interest to us.
Compared to ISO 250nn, ISO 19157:2013 provides a shorter list of six data quality characteristics called "elements": completeness, thematic accuracy, logical consistency, temporal quality, positional accuracy, and usability. Each element comprises several sub-elements, for example, completeness (commission and omission), logical consistency (conceptual, domain, format, topological), etc. A list of predefined measures is defined for each data unit (topographic dataset or street network), as shown in table 3.2.5.

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

Data quality unit	Data quality element	Data quality measure	Method
Topographic dataset	Completeness (commission)	Measure 1: Excess item Measure 2: Number of excess items Measure 3: Number of duplicate feature instances	Direct external Direct external Direct internal
Topographic dataset	Completeness (omission)	Measure 1: Missing item Measure 2: Number of missing items	Direct external Direct external
Topographic dataset	Thematic accuracy (correct classification)	Measure 1: Number of incorrectly classified features Measure 2: Misclassification rate	Direct external Direct external
Street network	Logical inconsistency (topological inconsistency)	Measure 1: Number of missing connections due to undershoots Measure 2: Number of missing connections due to overshoots Measure 3: Number of invalid self-intersect errors Measure 4: Number of invalid self-overlap errors	Direct internal Direct internal Direct internal Direct internal

Table 3.2.5. *Example of data quality measures defined in ISO 19157:2013.*

ISO 19157 distinguishes methods for the data quality evaluation procedures: direct (based on inspection of the items in the dataset) or indirect (based on external knowledge, such as lineage metadata). Direct evaluation is further classified by the source against which the evaluation is done: “internal” if only the data in the dataset is evaluated or “external” if there is a reference to external data (e.g., ground-based measurements).

A data quality evaluation process conforming to ISO 19157:2013 comprises steps very closely resembling the typical steps reported in section 2.8: (1) specify the data quality units to be evaluated;

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

(2) specify the data quality measures to be used to describe the quality of each data quality element of a data quality unit; (3) specify the data quality evaluation procedures; (4) perform the data quality evaluation; (5) report the results of the data quality evaluation.

A data quality measure conforming to ISO 19157:2013 is structurally well-defined and modelled as specified in the standard. Such a measure is described by at least an identifier, a name, an element name, a definition, and a value type. Optional descriptors are an alias, description, value structure, example, basic measure and one or more source references and parameters. Full inspection is most appropriate for small populations or tests that can be accomplished by automated means—for larger populations, checking a representative part of the data and reporting the quality result as a percentage rate is more appropriate and practical.

ISO 8000: Data quality

ISO 8000 defines characteristics of information and data that determine their quality and provides methods to manage, measure, and improve them. ISO 8000 stresses the importance of documenting the tailoring of standardised methods concerning the expectation and requirements pertinent to the use case at hand. This ISO family consists of “parts” applicable to all types of data (e.g., both structured and less structured, such as images, audio, video, and electronic documents) and parts applicable to specific types of data (e.g., ISO/DIS 8000-117 - Data quality - Part 117:

Application of ISO 8000-115 to identifiers in distributed ledgers including blockchains).
There is a limit to data quality improvement when only the nonconformity of data is corrected since the nonconformity can recur. However, when the root causes of the data nonconformity and the related data are traced and corrected through data quality processes, the recurrence of the same type of data nonconformity can be prevented. Therefore, a framework for process-centric data quality management is required to improve data quality more effectively and efficiently (see section 2.4). Furthermore, data quality can be enhanced through assessing processes and improving under-performing processes identified by the assessment.

The following fundamental principles apply to managing the quality of data:

- **process approach**: the processes that use, create and update data are defined and followed. These processes become repeatable and dependable by also defining and operating procedures for managing data quality;
- **continuous improvement**: data are improved through effective measurement and correction of data nonconformities that arise from data processing. Such modifications do not prevent the same nonconformities from occurring repeatedly. Sustained improvement arises from analysing, tracing, and removing the root causes of poor data quality, usually requiring the improvement of processes;
- **involvement of people**: specific responsibilities for data quality management exist at diverse levels of the organisation. End users have the most significant direct effect on data quality through data processing activities. In addition, data quality specialists perform the necessary intervention and control to implement and embed processes for improving data quality across the organisation. Finally, oversight by top management ensures that essential resources are made available and directs the organisation towards achieving the vision,

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

goals, and objectives for data quality.

While these principles overlap with those we discussed in the context of ISO 9001, i.e., principles applicable to the Quality Management Standard, this standard provides a more detailed definition, as shown in figure 3.2.6.

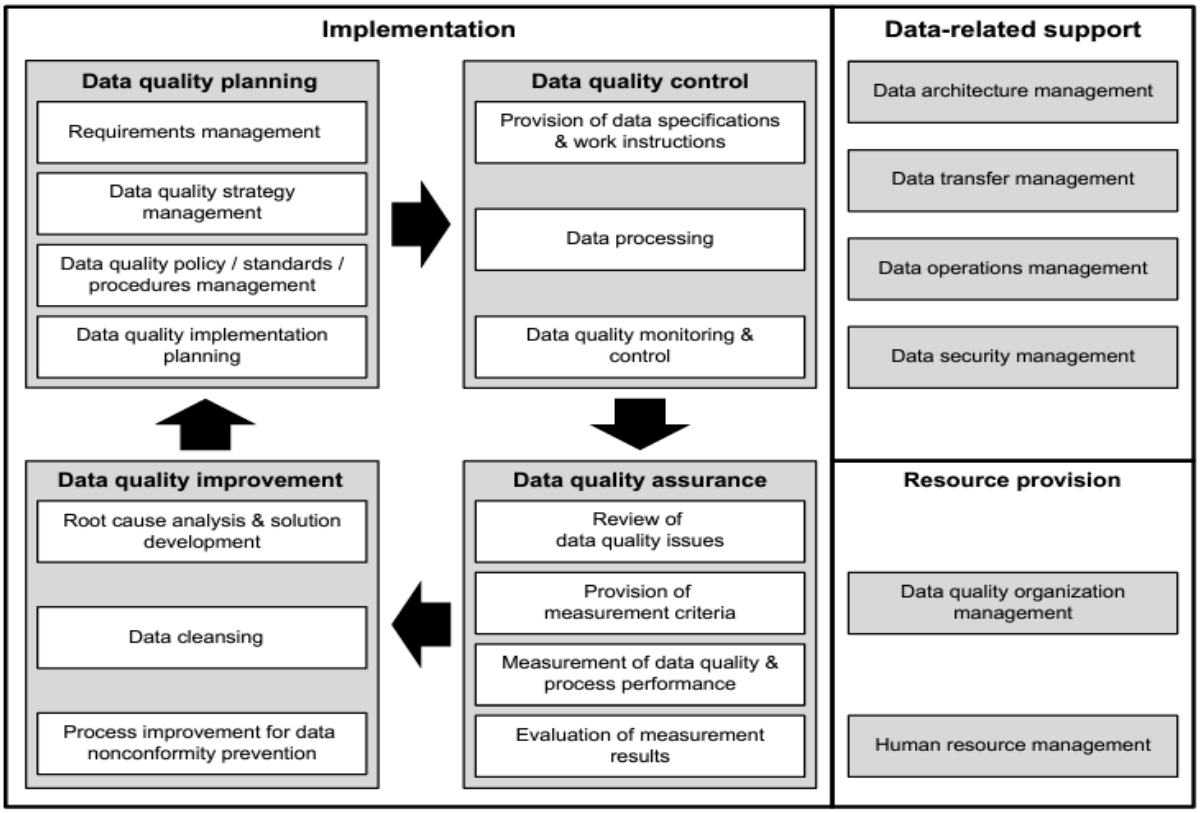

Figure 3.2.6. *Overview of the components to describe data quality (source: ISO 8000).*

Summary

ISO standards are sometimes challenging to implement and can be resource-consuming or even unnecessary in some contexts. First, on-demand access is granted after payment, which is not consistent with the principles of openness. This cost produces some resistance to adopting these standards and creates a tendency to develop and introduce some ad-hoc approaches instead. Indeed, ISO standards require not only time- and human-but also financial resources with no guarantee that they will serve as a "silver bullet."

In some cases, the process is not very intuitive (it is provided at a high level, making it difficult for people with insufficient knowledge, skills and experience to apply them to practise), or the procedures are not well-defined, so allow variations in their adoption. This is also the case for the diversity of terminology, which varies across ISOs, e.g., characteristics and sub-characteristics in ISO 250nn versus elements and sub-elements in ISO 191nn. These terms are typically referred to as dimensions in the data quality domain. However, it should be noted that in some cases, an alternative naming with reference to another standard is provided. Some ISO standards are

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

(relatively) complex, requiring a set of skills and experience, including digital and data quality literacy. The process leading to compliance with the ISO standards typically requires substantial manual reporting, often an additional barrier to standards adoption. ISO suggests and provides the help of a so-called ISO consultant, but it is a paid "offer," where the price can be relatively high (5000 - 50000 USD[34]).
Moreover, the ISO consultant provides support and help but does not write the documentation, which is expected by the organisation under scrutiny. Certification is also paid for and time-consuming depending on factors such as the size of the enterprise, the current quality system in place, etc. In addition, the certification obtained may not be permanent (see ISO 9001), which requires the applicant to restart the process and pay again to keep the certified status.

These issues are consistent with the results of our survey (details in section 3.3). The survey indicates the following reasons for standards not being actively adopted in practice: some standards are not open/proprietary formats, there is a lack of tools to be used to simplify and automate the standard implementation, there are too many standards to pick up the right one, they tend to be too complex ("it has to be seen as something that improves data quality, not just something that makes life difficult. And it must be widely adopted and the right mix of strict and flexible") or poorly defined, and they are expensive to implement. Another cluster of comments suggests that the resistance to adopting standards is related to unawareness of their existence, lack of experience, lack of incentive, and historical usage/ legacy. It should be noted that the survey's results are not purely ISO-related; respondents were consulted on the resistance to use standards in general.

In summary, ISO standards seem more appropriate for private or operational organisations rather than individuals or small research groups. This limits the direct application of the ISO standards in the context of open science. At the same time, the methods, concepts and approaches described can be adapted, with the advantage of being general and commonly accepted by a broad audience. It is a matter of fact that ISO standards are widely adapted in science, where different fields root their "ad-hoc" quality standards on one of the above mentioned ISO standards. In other words, ISO standards often become a basis for developing domain-specific data quality management strategies. For instance, ISO 9000, seen as the most general and domain-agnostic quality framework, serves as the basis for quality management in meteorology and climatology (e.g., World Meteorology Organization Quality Policy 35). Moreover, ISO 1915n, targeting geographic data, forms the basis for the INSPIRE 363738 directive. ISO 16363 defines recommended practices for assessing the trustworthiness of digital repositories and is adapted in the CoreTrustSeal certification (see section 2.10).

[34] https://the9000store.com/articles/iso-9000-cost/
[35] https://public.wmo.int/en/our-mandate/how-we-do-it/quality-management-framework
[36] https://www.ogc.org/standards/requests/151
[37]https://geonetwork-opensource.org/manuals/3.10.x/en/user-guide/describing-information/inspire-editing.html
[38]https://committee.iso.org/sites/tc211/home/standards-in-action/user-story-challenge/ecjrc---the-european-inspire-dir.html

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

The exploration of the ISO standards boosted this TF to reflect on quality-related concepts, such as continuous improvement (ISO 9000), methods to assess data quality (direct vs indirect measures in ISO 19157), assessment steps to consider (ISO 19157), principles to apply in data quality management (ISO 8000). Moreover, ISO 25012 offers an interesting categorization of the quality dimensions adapted in this document in section 2.7. ISO 38500 clarifies the relationship between data governance and quality (section 2.9). As a result, ISO standards constitute a rich body of knowledge on concepts, processes, and methods fundamental in the data quality lifecycle, even if those are not applied directly, are indirectly adopted in different scientific domains, and helped define the TF recommendations.

3.3 Task Force Survey

We developed and conducted a survey to find out the needs and expectations that EOSC should consider as its remit extends. We reached potential respondents by using the networks of the TF members, including EURAMET, NFDI, RFII, ERINHA, BY-COVID, Copernicus, FORS, LIBER. Moreover, the survey announcement was spread through mailing lists (GEANT, CODATA, ESIP, and EOSC Project newsletters) and the EOSC web portals (see figure 3.3.1). The survey was open for one month between April and May. Although there was no intention to lean towards specific research fields, the channels through which the survey was communicated might have influenced the representative nature of our results. The survey form can be found in Annex III; the main insights and conclusions are followed here.

Figure 3.3.1. *How the survey was advertised in the EOSC portal https://www.eosc.eu/news/complete-data-quality-survey*

Respondents' profile

- Geographical spread - the survey acquired responses from many countries in the World (see figure 3.3.2.). Europe contributed the most, with Germany and Portugal providing the highest share of answers;
- Discipline distribution - respondents come from all the fields of science, as classified in the

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

Frascati Manual[39]. A minimal number of answers (between 1 and 2) originated from agricultural sciences, chemistry and astronomy. The EOSC Association is invited to reflect on the little penetration in these disciplines;

- Other characteristics - respondents range from academia to private enterprises and funding agencies, usually having more than ten years of experience in their field. Respondents cover all roles in scientific data management, such as data providers, data stewards, standard makers, and publishers.

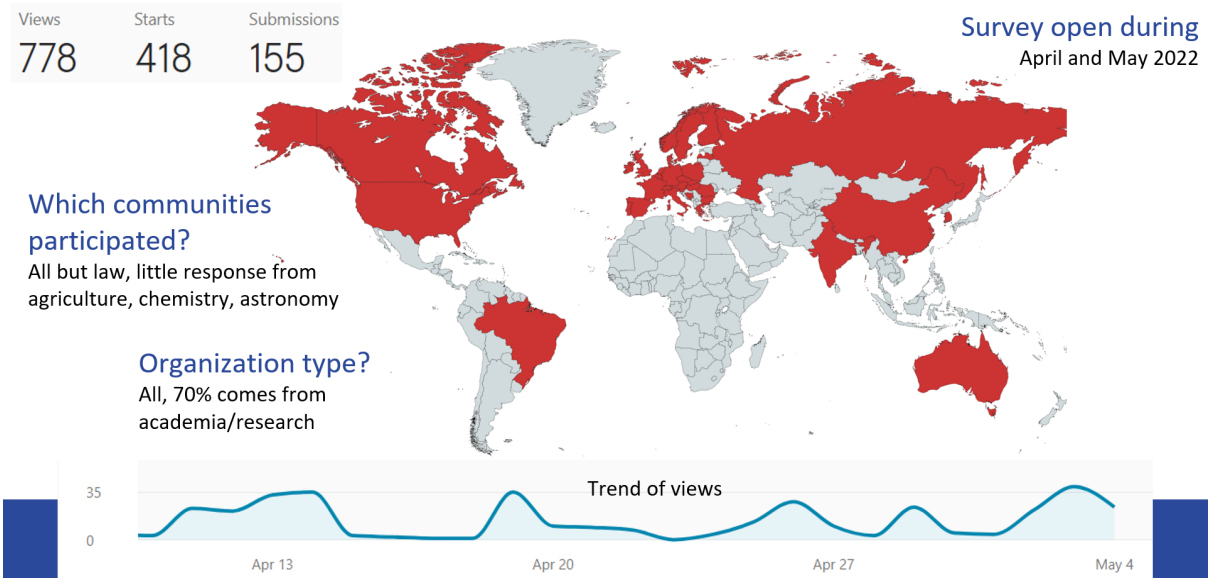

Figure 3.3.2. *Global distribution of responses. The survey had a good impact mainly in the north hemisphere, with nearly eight hundred views, of which 155 completed the survey till submission. The trend line at the bottom shows the variation of views over time.*

Insights

To the question about which information a user considers most relevant to select a dataset, the respondents highlight the importance of:

- making it clear how to cite the dataset, the terms of use, and the data provider or service to contact;
- availability of a user guide including how the data were collected, their traceability, primary purpose with examples of use and best data or software to use with the dataset,
- scientific accuracy and technical correctness of the dataset;
- versioning, i.e., the dataset needs to have a version associated, and the format should be compliant with typical community formats;
- following terms of confidentiality, disclosure agreements and any legal aspect;
- a clear statement of the datasets' strengths, limitations and known issues.

Other parts of the survey reveal that the user usually spends time processing a dataset to clean it when errors are found. On the other hand, when the providers find errors in their datasets, they tend

[39] https://www.oecd.org/sti/frascati-manual-2015-9789264239012-en.htm

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

to improve the dataset (which usually means remaking it) or transparently describe the weaknesses. Remaining on the side of the data provider, the survey asked about barriers that prevent following standards or providing quality-assured datasets. In this respect, significant factors acting as barriers are time, lack of expertise or experience, lack of dedicated tools, and lack of community-recognized methods or frameworks. Moreover, resource limitation puts quality in the second order of importance because there is no reward for providing standard-compliant datasets or no dedicated funding. Quality is seen as less important than other academic achievements. More insights come out of this answer and are reported in section 2.10.

The survey has shown misunderstandings and discrepancies between respondents due to their skills and expertise in data stewardship. One out of three respondents seems unaware of the importance of shared and community-approved vocabularies. That could be linked to the fact that vocabularies do not mean the same thing for all respondents. Metadata quality is attached to this issue: metadata schema should be based on agreed controlled vocabularies, but this aspect is neglected by most of the respondents. When the respondent is asked a similar question from a different angle, the answer surprisingly claims that most communities have metadata standards but suffer from the lack of controlled vocabulary. That seems to suggest that an effort should be made regarding data sharing literacy and FAIR recommendations. Along the same lines, data curation importance is underestimated by the respondents, with the majority understanding data curation as an error detection task, which is the primary concern for the survey's participants.

Most respondents stated that in their organisations, there is nobody officially responsible for data quality management. This calls for the need for EOSC to have a quality management function that guarantees datasets are ready for sharing. Respondents gave a striking preference for no ranking being associated with the datasets. If a ranking must be applied, then the priority should be placed on showing the FAIRness level of the datasets. No data content assessment is expected in EOSC, but a check of documentation availability for data understanding is required. Users expect EOSC to have a quality management function targeting basic curation. Thus, it must focus on (i) controlling the availability of basic metadata or documentation, providing sufficient contextualization of the dataset to make possible its proper interpretation and (ii) basic metadata compliance checks, but no enhanced curation like provision of new documentation or performance of independent assessments. This is coherent with the picture that users do not consider EOSC responsible for performing data content assessments (a.k.a. scientific quality, see section 2.7); the respondents are genuinely concerned about the possibility of EOSC performing that type of analysis. Looking at the processes the quality management function should be equipped with, the responses suggest that the assessments shall be first shown to the data provider to give a chance to improve the data and then to the users. Communication across the different actors involved (service managers, funders, data providers, users) is essential (in agreement with section 2.8). The methodology must be the same for similar datasets and guarantee that quality information is consistent across datasets for comparability. In addition, users would like to leave comments about data quality along with the datasets served through EOSC.

The survey results gave insights that are further analysed in the context of specific sections of this document and support our main conclusions.

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

3.4 Preliminary data quality requirements for EOSC

Based on the survey results and the mapping between CoreTrustSeal certification and INSPIRE directive (see section 2.10), as well as indications from the EOSC interoperability framework (EC 2021), a number of requirements valid for all datasets in EOSC can be drafted. These represent the requirements with the highest priority to guarantee a minimum potential re-usage of the datasets. Because a FAIR ecosystem is characterised by unknown specific applications re-users want to pursue, these requirements guarantee fitness-for-use but not necessarily fitness-for-purpose (see section 2.5 for details):

- documentation reporting data content accuracy (sometimes referred to as scientific quality);
- clarity about how to cite, attribute/give credits (including contributors, year, comprehensive title, persistent identifier);
- user guide documentation showing what the dataset is about (e.g., abstract summarising resource), its structure and version, its original purpose, how it was produced, examples of usage, and best combinations with software or other datasets;
- documented terms of use (licence), evidence that data were created and curated in compliance with disciplinary and ethical norms (e.g., GDPR[40], non-disclosure agreements, security, copyright - IPR);
- proof of compliance with metadata standards, enriched of controlled vocabulary;
- data format according to community standards;
- documentation describing essential provenance and traceability. The more advanced information in this regard is not considered a priority;
- contact details /information about who to contact, who is responsible for the dataset and who owns it;
- evidence that the dataset has passed sanity checks, these are simple and usually automated checks (e.g., no unexpected gaps in time series);
- strengths, limitations and known issues;
- compliance with the FAIR Metrics.

Further refinement will be necessary for the future, and specific standards to follow will need to be identified. The recommendation will be to prioritise those standards applicable across domains. An example is the Dublin Core Metadata Initiative (DCMI[41]), a standard that makes it easier to describe and find resources. Moreover, it is advisable to structure the consolidated requirements into standard templates to facilitate inter-comparison and machine-readability.

[40] General Data Protection Regulation https://ec.europa.eu/info/law/law-topic/data-protection_en

[41] www.dublincore.org

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

3.5 Gaps and ways of gauging community maturity in data quality management

Data quality assessment needs standards to check data against; unfortunately, not all communities have agreed on standards like metadata models, vocabularies, or storage formats. It is necessary to fill these gaps to guarantee data interoperability within and across communities and perform complete data quality assessments. We extracted a few examples of the gaps, but the current situation requires a more detailed and systematic evaluation in each community. It should be considered for EOSC the opportunity to assist and push each community to agree on community standards that guarantee the FAIR exchange of research data.

According to the survey (section 3.3), the most common gap communities report is the lack of agreed controlled vocabulary. For instance, Earth sciences miss a list of variable names and descriptions consistent across dataset categories. An example is the widely used variable "surface temperature," so far, it is not clear what reference to use to name and describe it; it ranges from "(surface) temperature" in GCOS[42] to "near-surface air temperature" in CMIP tables[43]. Furthermore, some communities suffer a more substantial lack of standards in format, metadata structure, and storage, as in the Bioimaging community (see case study below). Materials Engineering lacks complete community-agreed metadata standards, particularly a detailed description of the material's microstructure.

Another relevant gap stems from the lack of acknowledgement, rewarding or/and funding for data curators. Data processing, analysis, and publication have much larger visibility than data curation. Despite its relevance, this task is considered a lower priority (Wierling et al. 2021). This situation could be improved by having a quality management function in EOSC that pushes the actors involved to follow precise requirements to serve their datasets through the EOSC ecosystem.

Case study - Lack of standards and quality specifications in bioimaging

The amount of imaging data has increased dramatically in the life sciences over the last two decades due to new labelling techniques such as live cell labelling with a green fluorescent protein (GFP) and the resulting long-term experiments and time-lapse movies, as well as new methods for imaging whole organs in 3D (e.g., light sheet microscopy). Instead of a few hundred MB, data volumes are often in the range of several TB. A major problem remains the management and storage of this data - not only because the volume of data is increasing exponentially but also because there are still no standards for a consistent file format. Usually, far too little attention is paid to the metadata stored with the data. To make matters worse, it is not only the raw data that need to be stored but also the processed images. The image data is nothing more than a spreadsheet with different intensity values in the various cells. However, understanding and analysing the imaging data can only be done if meaningful metadata are stored alongside the imaging data. The most basic metadata describes the microscope hardware used with the settings (wavelength, light intensity, scaling, etc.). In addition to this hardware metadata, biologically relevant information such

[42] https://gcos.wmo.int/en/essential-climate-variables/surface-temperature
[43] https://github.com/PCMDI/cmip6-cmor-tables/blob/master/Tables/CMIP6_3hr.json

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

as sample origin, animal model, phenotype, batch number, labelling technology, etc., must also be added. The most optimal selection of the stored metadata determines the reusability of the imaging data in the life sciences.

Unfortunately, the microscopy/bioimaging community has not yet been able to agree on a standard file format, and twelve standards44 exist (FAIRsharing 2022). This is due to the microscope manufacturers, who usually optimise the file formats for their systems and the best and fastest possible local data storage. These so-called proprietary file formats have often led to locked-in situations on the respective devices, meaning that the data could only be opened with the respective software from the respective company. Globally used translation programs such as Bio-Formats[45] have decisively improved this situation but have not solved it. Only when the community agrees on uniform file formats and on standards for the metadata stored with them will it become possible to store image data with reliable hardware metadata. In addition to these challenges, the bioimage community often encounters insurmountable difficulties concerning data repositories. To date, many universities provide inadequate or no storage facilities. Furthermore, open research data repositories like the ones available in genomics and proteomics are still extremely limited but are slowly building up. The Image Data Resource[46] (IDR) is a perfect example of recent bioimage community progress towards FAIR open bioimaging data.

Due to pressure from funding agencies and the suffering of researchers who can no longer find/use the data of former colleagues because of insufficient information, FAIR data management is now becoming a pressing priority everywhere and often necessary to get research proposals accepted. Global initiatives such as QUAREP[47] (Quality Assessment and Reproducibility for Instruments & Images in Light Microscopy) and the microscopy community, led by the many microscopy core facilities, are now planning solutions to these problems. These efforts aim to define tier-levels for metadata standards as laid out in a recent special issue of Nature methods[48]. They suggested metadata tier levels are:

- tier 1: Minimum information/qualitative or basic quantification/materials and methods;
- tier 2: Advanced quantification;
- tier 3: Manufacturing / technical development / full documentation.

The other significant effort is to find an acceptable file format solution handling extensive data in cloud-like storage, based on the principle of pyramidal file format that we already know from other cloud storage solutions (only the necessary section of the image is loaded in the relevant detail at any given time, e.g., Google maps). This format could be OME-NGFF[49] (Open Microscopy Environment-Next Generation File Format). More information on the status of the bioimaging community is offered by Swedlow et al. (2021) and the Elixir RDMkit page[50]. A collection of imaging

[44] https://fairsharing.org/search?q=microscopy&fairsharingRegistry=standard&searchAnd=false
[45] https://www.openmicroscopy.org/bio-formats/
[46] https://idr.openmicroscopy.org/
[47] https://quarep.org/
[48] https://quarep.org/special-issue-in-nature-methods-released/
[49] https://www.glencoesoftware.com/blog/2022/01/31/OME-NGFF-in-action.html
[50] https://rdmkit.elixir-europe.org/bioimaging_data.html

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

data resources and their standards is available in FAIRsharing (2022).

Conclusions

FAIR bioimaging data management in the future will depend on the agreement on a standard microscopy file format accompanied by clear rules and recommendations for minimally required metadata for both hardware and biological data. Searchability and reuse of the data will only be possible once these needs are fulfilled.

Gauging community maturity in data quality management

It could be helpful to have a set of requirements to measure the maturity of a community when dealing with quality. It may help to identify gaps that need to be filled to facilitate interdisciplinary data interoperability. It is worth investigating the opportunity to generate a maturity matrix to measure a community's ability to manage data quality from ad-hoc to more managed stages. In this respect, a first step was taken by asking through a survey (details in section 3.3) for specific requirements that the communities consider most relevant. The requirements are listed in order of priority:

- Availability of metadata standards;
- Agreed vocabulary and terminology;
- Definition of a standard quality management framework;
- Identified metrics to quantify quality, which calls, with lower importance, for automatic tools for quality evaluation;
- Quality assessments are operational routine and funded;
- Established data governance;
- Recognition of the need for a capacity-building programme;
- Existence of a ranking system;
- The feedback loop between consumer and supplier;
- Quality assessments are not separate and cumbersome but are continuously integrated into all parts of the dataset lifecycle.

The future TF may find it beneficial to explore the usefulness of this maturity matrix and keep developing it.

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

4. PRINCIPLES AND RECOMMENDATIONS

4.1 Principles

This TF identified a set of principles valid for EOSC:

- **EOSC shall adopt an approach based on the principle of documented known quality.** This approach seeks to maximise the descriptive metadata to allow the user to understand how the dataset was produced and to assess its appropriateness for the intended application. This approach accommodates the data variability from different producers and supports the informed use of the same dataset for multiple applications. As frequently demanded in the survey this TF released (details in section 3.3), documentation must enable data values to be interpreted in context to prevent misuse and facilitate reuse. Format and metadata must be as close to the community standards to facilitate the usability, licence of use need to be upfront, and aspects like uncertainty or collection process are essential for some disciplines and must be well documented. More specific minimum requirements are listed in section 3.4.
- Data content accuracy (sometimes referred to as scientific quality) is in the hands of the data producer/provider and peers. EOSC does not assess it but must ensure that accuracy information is available to the users. As a result, the main point is that the data quality approach in EOSC has not to focus on assessing the data content accuracy. Still, it must **focus on guaranteeing the dataset is understandable and correctly reusable**.
- **The data producer/provider is in charge of the product's inherent quality,** e.g., data values are accurate. In contrast, the curator is in charge of guaranteeing minimum completeness of the documentation associated with the product. It is a common practice extracted in different fields (e.g., climatology (Lacagnina et al. 2022) and oceanography[51]). Therefore, errors found by the curators or users must be rectified by the data producer/provider. If not possible, errors need to be documented. Improving data quality as close to the source (i.e., producer or provider) as possible is highly recommended (Belbin et al. 2013).
- **Data and associated information are commodities that are not destroyed by their use.** Thus, if properly preserved, they can be made available at minimal cost for many uses not anticipated during data collection. It implies that sharing data is a strategic choice that EOSC can amplify. It is necessary to explore certification and conformity mechanisms to assure researchers that the infrastructures they deposit and access data conform to clear rules and criteria. If researchers feel a loss of control and visibility or have concerns about how professionally their data will be managed, additional barriers to data sharing will emerge.
- **Quality information must be objective, transparent, consistent, and communicated effectively.** A way forward is to document the assessment method and use it consistently. Several factors influence the selection of the assessment method; we stress the importance of keeping in mind the peculiarity of data: they change fast and keep increasing, so their assessment must step up in a reasonable amount of time.
- **Data quality mission**: it ensures (i) non-functional requirements are met (fitness-for-use); (ii)

[51] https://marine.copernicus.eu/sites/default/files/CMEMS-PQ-StrategicPlan-v1.6-1_0.pdf

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

information necessary to understand/use product is available; (iii) [if user's application is known] functional requirements are met (fitness-for-purpose), [if not known] information to guide user in selection process is available (self-assessment of fitness-for-purpose).

- In line with the conclusions made in the EOSC interoperability framework, **consistent and traceable harmonised dataset documentation is needed** to facilitate and enhance interoperability. Quality practitioners must guarantee that all stakeholders clearly understand the adequacy of the information they access through a centralised portal with consistent terminology.
- Beyond FAIR principles, the community requests that EOSC start **focusing on the CARE** (Collective Benefit, Authority to Control, Responsibility, and Ethics) principles**.**

4.2 Recommendations

Based on the experience gained by the members of this TF, the following recommendations can be made:

- **Identify** EOSC users, service managers, data providers, funding authorities and system **requirements**. User engagement is necessary to understand the user requirements; it may or may not be part of a quality management function. Evaluation of stakeholder needs is not a one-time requirement but a continuous and collaborative part of the service delivery process. Collection of the stakeholder needs can also help to define quality indicators for EOSC.
- The survey (details in section 3.3) suggested that users of EOSC expect tools and services designed according to a user-centric model. It appears important to explore the possibility of having a service quality function evaluating the degree to which the EOSC ecosystem fulfils the user requirements in terms of the quality of the service and products provided. It is recommended to develop a proof-of-concept quality function performing basic quality assessments tailored to the EOSC needs. The proof-of-concept cannot be a theoretical conceptualization of what is nice to have in terms of quality. Still, it must be constrained by the reality of dealing with an enormous amount of data within a reasonable time and effort. The recommendation is to **define a core of disciplines where a pre-operational quality management function should be built** and then progressively see which parts of the approach can be expanded to more disciplines. In addition, this function can support rewarding research teams most committed to providing FAIR datasets.
- It must be clear and well-advertised that quality does not refer to data content quality only, a.k.a. scientific quality. The survey demonstrated that several respondents see quality assessments as dangerous when done by external organisations like EOSC because the respondents see quality usually associated with assessing the data content. The latter must be evaluated by peer colleagues of the data provider (if not self-evaluated); other aspects of quality can be accessed by practitioners far from the data provider's expertise because they address dataset facets that are not purely scientific, e.g., archiving policies. Therefore, **EOSC should focus on checking that the dataset is structured and documented in a way that can be (re)used and understood.** EOSC should avoid checking the soundness of the data content. Things like uncertainty are also important to properly (re)use a dataset but must be

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

evaluated outside EOSC, which only checks that evidence about data content assessments is available. The survey revealed that the communities expect EOSC to be equipped with basic data quality management, i.e., it should perform tasks like controlling the availability of basic metadata or documentation and performing basic metadata compliance checks. The EOSC quality management should not change data but point to deficiencies that the data provider or producer can address. Moreover, the users should be allowed to comment on the quality of the datasets.

- Related to the point above, many EOSC stakeholders see quality usually associated with assessing the data content only. It points to a lack of data sharing and curation literacy. It is recommended to **make available training material to create a common ground understanding of what quality is.**
- This TF prepared a survey (section 3.3) to explore interest in any indicator or ranking system to apply to the datasets in EOSC. The results showed a striking preference for no ranking. If a ranking has to be used, then the priority should be placed on establishing the FAIRness level of the datasets. Assessment of the data content, a.k.a. scientific assessment, should be avoided. **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 assessments should be accompanied by the methodology applied and by the profile of the authors. The methodology must be the same for similar datasets.
- **Establish workflows for quality assessments**. The quality assessments must be a multi-actor process between the data provider, EOSC and users. The resulting content should be captured in structured, human- and machine-readable, and standard-based formats. Information across datasets must be easily comparable, which calls for providing homogeneous quality information. Furthermore, it shall be explored whether to update the assessments every certain time (e.g., every two years, depending on resources) to keep quality assigned up to date. It is also recommended to make use of the concept of minimum requirements (section 2.8). It is a widespread practice to identify a set of mandatory versus optional requirements because time, resources and workforce constraints limit the depth of the activities a data quality function can deal with.
- Whether the EOSC will be equipped with a quality management function also affects the response to survey respondents' request (section 3.3) for having **automatic tools for validation**. If a quality function is set, these tools are not necessary. Otherwise, these need to be made available. It must be recognized that it is not always possible to develop automatic validation tools because there are practices in data quality that go beyond the product itself and may regard the organisational aspects of data production. As such, documentation screening may need human intervention.
- **EOSC should support each community to agree on community standards** to guarantee the FAIR exchange of research data. Setting a quality management function can help in this direction because the function can identify which standard already in use by some initiatives can be enforced as a general requirement for that community. For instance, the bioimaging community has some format standards (section 3.5), but none has acquired the status of global acceptance so far. A specific format could be identified in EOSC, and ask the

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

bioimaging providers that only datasets complying with the identified format can be served in EOSC. We recommend that EOSC considers taking the opportunity to encourage communities to reach a consensus in using their standards. As an RDA-endorsed output, FAIRsharing (2022) helps communities to show what standards they use, which allows monitoring community buy-in and adoption.

- In setting data quality processes in EOSC**, it will be fundamental to analyse governance models of similar activities.** This includes aspects such as efficiency, effectiveness, and inclusivity, as well as questions about how generic and specific competencies are mapped in governance. For example, will it make sense to have different advisory boards or just one central council? Which actors will be sought for participation? What voice should existing networks get?
- Additional recommendations more specific for the future **evolution of this Task Force** are given in Annex II.

ACKNOWLEDGMENTS

This study is based on work conducted in the European Open Science Cloud (EOSC) Task Force (TF) "FAIR Metrics and Data Quality." We want to acknowledge the contribution and input of colleagues from several European institutions, the EOSC Association and several external-to-TF stakeholders who gave feedback based on their own experience, and the TF Support Officer Paola Ronzino (EOSC-A) for the final formatting of the document. Acknowledgements also go to Sarah Stryeck and Raed Al-Zoubi for their contribution to desk research on ISO standards, Richard Dennis and Carlo Lacagnina for final proofreading. We would also like to acknowledge all the respondents to the survey launched in April 2022, in particular the ones who took time to complete it fully, namely Lucy Bastin and Stefano Materia.

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

5. REFERENCE

Akerlof, G. A. (1970). The Market for 'Lemons': Quality Uncertainty and the Market Mechanism. The Quarterly Journal of Economics, 84(3), 488. https://doi.org/10.2307/1879431

Alexander, J. E., & Tate, M. A. (1999). Web wisdom: How to evaluate and create information quality on the Web. Lawrence Erlbaum Associates.

Amicis, F. de, & Batini, C. (2004). A methodology for data quality assessment on financial data. https://doi.org/10.5169/SEALS-790977

Baker, K. S., Duerr, R. E., & Parsons, M. A. (2016). Scientific knowledge mobilization: Co-evolution of data products and designated communities. International Journal of Digital Curation, 10(2), 110–135. https://doi.org/10.2218/ijdc.v10i2.346

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. ACM Computing Surveys, 41(3), 1–52. https://doi.org/10.1145/1541880.1541883

Batini, C., & Scannapieco, M. (2016). Data and information quality: Dimensions, principles and techniques. Springer Berlin Heidelberg.

Belbin, L., Daly, J., Hirsch, T., Hobern, D., & LaSalle, J. (2013). A specialist's audit of aggregated occurrence records: An 'aggregator's' perspective. ZooKeys, 305, 67–76. https://doi.org/10.3897/zookeys.305.5438

Bouchana, S., & Janati Idrissi, M. A. (2015). Towards an assessment model of end user satisfaction and data quality in Business Intelligence systems. 2015 10th International Conference on Intelligent Systems: Theories and Applications (SITA), 1–6. https://doi.org/10.1109/SITA.2015.7358431

Bouwknegt, M., & Havelaar, A. H. (2015). Uncertainty assessment using the NUSAP approach: A case study on the EFoNAO tool. EFSA Supporting Publications, 12(1). https://doi.org/10.2903/sp.efsa.2015.EN-663

Cai, L., & Zhu, Y. (2015). The challenges of data quality and data quality assessment in the big data era. Data Science Journal, 14(0), 2. https://doi.org/10.5334/dsj-2015-002

Cappiello, C., Di Noia, T., Marcu, B. A., & Matera, M. (2016). A quality model for linked data exploration. In A. Bozzon, P. Cudre-Maroux, & C. Pautasso (Eds.), Web Engineering (Vol. 9671, pp. 397–404). Springer International Publishing. https://doi.org/10.1007/978-3-319-38791-8_25

Chapman, A., Belbin, L., Zermoglio, P., Wieczorek, J., Morris, P., Nicholls, M., Rees, E. R., Veiga, A., Thompson, A., Saraiva, A., James, S., Gendreau, C., Benson, A., & Schigel, D. (2020). Developing standards for improved data quality and for selecting fit for use biodiversity data. Biodiversity Information Science and Standards, 4, e50889. https://doi.org/10.3897/biss.4.50889

Chen, L., Ali Babar, M., & Nuseibeh, B. (2013). Characterizing architecturally significant requirements. IEEE Software, 30(2), 38–45. https://doi.org/10.1109/MS.2012.174

Cox, S. J. D., Gonzalez-Beltran, A. N., Magagna, B., & Marinescu, M.-C. (2021). Ten simple rules for making a vocabulary FAIR. PLOS Computational Biology, 17(6), e1009041. https://doi.org/10.1371/journal.pcbi.1009041

David, R., Bouveret, L., Coché, L., Corrêa, P. P., Edmunds, R., Heredia, A., Jung, J.-L., Kondo, Y., Berre, I. L., Bras, Y. L., Lerigoleur, E., Mabile, L., Machicao, J., Madon, B., Murayama, Y., O'Brien, M., Osawa, T., Raoul, H., Richard, A., … Wyborn, L. (2021). Data dictionary cookbook for research data and

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

software interoperability at global scale. https://doi.org/10.5281/ZENODO.4683066

Debattista, J., Auer, Sö., & Lange, C. (2016). Luzzu—A methodology and framework for linked data quality assessment. Journal of Data and Information Quality, 8(1), 1–32. https://doi.org/10.1145/2992786

Dobratz, S., & Schoger, A. (2007). Trustworthy digital long-term repositories: The nestor approach in the context of international developments. In L. Kovács, N. Fuhr, & C. Meghini (Eds.), Research and Advanced Technology for Digital Libraries (Vol. 4675, pp. 210–222). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-74851-9_18

Dumbill, E. (2013). Making sense of big data. Big Data, 1(1), 1–2. https://doi.org/10.1089/big.2012.1503

English, L. P. (1999). Improving data warehouse and business information quality: Methods for reducing costs and increasing profits. Wiley.

Eppler, M. J., & Muenzenmayer, P. (2002). Measuring information quality in the web context: A survey of state-of-the-art instruments and an application methodology. In C. Fisher & B. N. Davidson (Eds.), Seventh International Conference on Information Quality (ICIQ 2002) (pp. 187–196). MIT.

European Commission, Directorate-General for Research and Innovation, EOSC Executive Board, Corcho, O., Eriksson, M., Kurowski, K., Ojsteršek, M., Choirat, C., Sanden, M. van de, & Coppens, F. (2021). EOSC interoperability framework: Report from the EOSC executive board working groups fair and architecture. Publications Office of the European Union. https://data.europa.eu/doi/10.2777/620649

Evans, B., Druken, K., Wang, J., Yang, R., Richards, C., & Wyborn, L. (2017). A data quality strategy to enable fair, programmatic access across large, diverse data collections for high performance data analysis. Informatics, 4(4), 45. https://doi.org/10.3390/informatics4040045

Fairsharing. (n.d.). Retrieved 4 October 2022, from https://fairsharing.org/

Fürber, C., & Hepp, M. (2016). Data Quality Management with semantic technologies. Springer Gabler.

Jeusfeld, M. A., Quix, C., & Jarke, M. (1998). Design and analysis of quality information for data warehouses. In T.-W. Ling, S. Ram, & M. Li Lee (Eds.), Conceptual Modeling – ER '98 (Vol. 1507, pp. 349–362). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-49524-6_28

Günther, L. C., Colangelo, E., Wiendahl, H.-H., & Bauer, C. (2019). Data quality assessment for improved decision-making: A methodology for small and medium-sized enterprises. Procedia Manufacturing, 29, 583–591. https://doi.org/10.1016/j.promfg.2019.02.114
Haiden, T., Janousek, M., Vitart, F., Ferranti, L., & Prates, F. (2019). Evaluation of ECMWF forecasts, including the 2019 upgrade. https://doi.org/10.21957/MLVAPKKE

International Energy Agency (Ed.). (2000). Energy labels & standards. OECD. Retrieved 4 October 2022, from https://www.iea.org/reports/energy-labels-standards

ISO/IEC Guide 2:2004. Standardization and related activities — General vocabulary. Geneva, Switzerland. https://www.iso.org/standard/39976.html

ISO/IEC 2382-1:1993. Information technology — Vocabulary — Part 1: Fundamental terms. Geneva, Switzerland. https://www.iso.org/standard/7229.html

ISO 8000-8:2015. Data quality - Part 8: Information and data quality: Concepts and measuring.

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

Geneva, Switzerland. https://www.iso.org/standard/60805.html

ISO 9000. Quality management. Geneva, Switzerland. https://www.iso.org/iso-9001-quality-management.html

ISO 9000:2005. Quality management systems — Fundamentals and vocabulary. Geneva, Switzerland. https://www.iso.org/standard/42180.html

ISO/IEC 16363:2012. Space data and information transfer systems — Audit and certification of trustworthy digital repositories. Geneva, Switzerland. https://www.iso.org/standard/56510.html

ISO/IEC 17000:2020. Conformity assessment - Vocabulary and general principles. Geneva, Switzerland. https://www.iso.org/standard/73029.html

ISO 19011:2018. Guidelines for auditing management systems. Geneva, Switzerland. https://www.iso.org/standard/70017.html

ISO 19135-1:2015. Geographic information — Procedures for item registration. Geneva, Switzerland. https://www.iso.org/standard/54721.html

ISO 19157:2013. Geographic information - Data quality. Geneva, Switzerland. https://www.iso.org/standard/32575.html

ISO/TS 19158:2012. Geographic information - Quality assurance of data supply. Geneva, Switzerland. https://www.iso.org/standard/32576.html

ISO/IEC 25000:2014. Systems and software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - Guide to SQuaRE. Geneva, Switzerland. https://www.iso.org/standard/64764.html

Jahn, G., Schramm, M., & Spiller, A. (2004). Differentiation of Certification Standards: The Trade-off between Generality and Effectiveness in Certification Systems.

Jahn, G., Schramm, M., & Spiller, A. (2005). The reliability of certification: Quality labels as a consumer policy tool. Journal of Consumer Policy, 28(1), 53–73. https://doi.org/10.1007/s10603-004-7298-6

Kahn, M. G., Callahan, T. J., Barnard, J., Bauck, A. E., Brown, J., Davidson, B. N., Estiri, H., Goerg, C., Holve, E., Johnson, S. G., Liaw, S.-T., Hamilton-Lopez, M., Meeker, D., Ong, T. C., Ryan, P., Shang, N., Weiskopf, N. G., Weng, C., Zozus, M. N., & Schilling, L. (2016). A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. EGEMs (Generating Evidence & Methods to Improve Patient Outcomes), 4(1), 18. https://doi.org/10.13063/2327-9214.1244

Klein, A., & Lehner, W. (2009). Representing data quality in sensor data streaming environments. Journal of Data and Information Quality, 1(2), 1–28. https://doi.org/10.1145/1577840.1577845

Knight, S.-A., & Burn, J. (2005). Developing a framework for assessing information quality on the world wide web. Informing Science: The International Journal of an Emerging Transdiscipline, 8, 159–172. https://doi.org/10.28945/493

Lacagnina, C., Doblas-Reyes, F., Larnicol, G., Buontempo, C., Obregón, A., Costa-Surós, M., San-Martín, D., Bretonnière, P.-A., Polade, S. D., Romanova, V., Putero, D., Serva, F., Llabrés-Brustenga, A., Pérez, A., Cavaliere, D., Membrive, O., Steger, C., Pérez-Zanón, N., Cristofanelli, P., … Díez, M. G. (2022). Quality management framework for climate datasets. Data Science Journal, 21(1), 10. https://doi.org/10.5334/dsj-2022-010

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

Langstaff, S. A. (2010). Sensory quality control in the wine industry. In Sensory Analysis for Food and Beverage Quality Control (pp. 236–261). Elsevier. https://doi.org/10.1533/9781845699512.3.236

Lawrence, B., Jones, C., Matthews, B., Pepler, S., & Callaghan, S. (2011). Citation and peer review of data: Moving towards formal data publication. International Journal of Digital Curation, 6(2), 4–37. https://doi.org/10.2218/ijdc.v6i2.205

Lee, Y. W., Strong, D. M., Kahn, B. K., & Wang, R. Y. (2002). AIMQ: A methodology for information quality assessment. Information & Management, 40(2), 133–146. https://doi.org/10.1016/S0378-7206(02)00043-5

Lindemann, M., Nuy, L., Briele, K., & Schmitt, R. (2019). Methodical data-driven integration of perceived quality into the product development process. Procedia CIRP, 84, 406–411. https://doi.org/10.1016/j.procir.2019.04.205

Long, J., & Seko, C. (2005). A Cyclic-Hierarchical Method for Database Data-Quality Evaluation and Improvement. In R. Wang, E. Pierce, S. Madnick, and F. C.W., editors, Advances in Management Information Systems-Information Quality Monograph (AMIS-IQ) Monograph. Sharpe, M.E. Enterprise Knowledge Management - The Data Quality Approach - Chapter 4. Morgan Kaufmann Series in Data Management Systems, 2004.

Loshin, D. (2001). Enterprise knowledge management: The data quality approach. Morgan Kaufmann.

Magagna, B., Rosati, I., Stoica, M., Schindler, S., Moncoiffe, G., Devaraju, A., Peterseil, J., & Huber, R. (2021). The i-adopt interoperability framework for fairer data descriptions of biodiversity. https://doi.org/10.48550/ARXIV.2107.06547

Meek, S., Jackson, M. J., & Leibovici, D. G. (2014). A flexible framework for assessing the quality of crowdsourced data. AGILE Digital Editions. http://repositori.uji.es/xmlui/handle/10234/98927

Naumann, F. (Ed.). (2002). Quality-driven query answering for integrated information systems (Vol. 2261). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-45921-9

Nikiforova, A. (2020). Definition and evaluation of data quality: User-oriented data object-driven approach to data quality assessment. Baltic Journal of Modern Computing, 8(3). https://doi.org/10.22364/bjmc.2020.8.3.02

Parmelli, E., Langendam, M., Piggott, T., Adolfsson, J., Akl, E. A., Armstrong, D., Braithwaite, J., Brignardello-Petersen, R., Follmann, M., Leś, Z., Meerpohl, J. J., Neamtiu, L., Qaseem, A., Rossi, P. G., Saz-Parkinson, Z., van der Wees, P. J., & Schünemann, H. J. (2021). Guideline-based quality assurance: A conceptual framework for the definition of key elements. BMC Health Services Research, 21(1), 173. https://doi.org/10.1186/s12913-021-06148-2

Peng, G., Lacagnina, C., Ivánová, I., Downs, R. R., Ramapriyan, H., Ganske, A., Jones, D., Bastin, L., Wyborn, L., Bastrakova, I., Wu, M., Shie, C.-L., Moroni, D. F., Larnicol, G., Wei, Y., Ritchey, N., Champion, S., Hou, C.-Y., Habermann, T., … le Roux, J. (2021). International community guidelines for sharing and reusing quality information of individual earth science datasets [Preprint]. Open Science Framework. https://doi.org/10.31219/osf.io/xsu4p

Peng, G. (2018). The state of assessing data stewardship maturity – an overview. Data Science Journal, 17, 7. https://doi.org/10.5334/dsj-2018-007

Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. Communications of the ACM,

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

45(4), 211–218. https://doi.org/10.1145/505248.506010

Phlips, L. (1983). The economics of price discrimination. Cambridge University Press.

CoreTrustSeal Standards And Certification Board. (2019). Coretrustseal trustworthy data repositories requirements 2020–2022. https://doi.org/10.5281/ZENODO.3638211

Redman, T. C. (1996). Data quality for the information age. Artech House.

Renear, A. H., Sacchi, S., & Wickett, K. M. (2010). Definitions of dataset in the scientific and technical literature: Definitions of Dataset in the Scientific and Technical Literature. Proceedings of the American Society for Information Science and Technology, 47(1), 1–4. https://doi.org/10.1002/meet.14504701240

Rowley, J. (2007). The wisdom hierarchy: Representations of the DIKW hierarchy. Journal of Information Science, 33(2), 163–180. https://doi.org/10.1177/0165551506070706

Scannapieco, M., Virgillito, A., Marchetti, C., Mecella, M., & Baldoni, R. (2004). The DaQuinCIS architecture: A platform for exchanging and improving data quality in cooperative information systems. Information Systems, 29(7), 551–582. https://doi.org/10.1016/j.is.2003.12.004

Sensitive Data Expert Group (2020). Sensitive data toolkit for researchers part 1: Glossary of terms for sensitive data used for research purposes. https://doi.org/10.5281/zenodo.4088946

Shanks, G., & Corbitt, B. (1999) Understanding data quality: Social and cultural aspects. Procedures of the 10th Australasian Conference on Information Systems, Wellington: MCB University Press Ltd., pp 785–797

Shreeves, S., L., & Cragin., M., H. (2008). Introduction: Institutional repositories: current state and future. Library Trends, 57(2), 89–97. https://doi.org/10.1353/lib.0.0037

Stockhause, M., Höck, H., Toussaint, F., & Lautenschlager, M. (2012). Quality assessment concept of the World Data Center for Climate and its application to CMIP5 data. Geoscientific Model Development, 5(4), 1023–1032. https://doi.org/10.5194/gmd-5-1023-2012

Strong, D. M., Lee, Y. W., & Wang, R. Y. (1997). Data quality in context. Communications of the ACM, 40(5), 103–110. https://doi.org/10.1145/253769.253804

Stylidis, K., Wickman, C., & Söderberg, R. (2020). Perceived quality of products: A framework and attributes ranking method. Journal of Engineering Design, 31(1), 37–67. https://doi.org/10.1080/09544828.2019.1669769

Su, Z., & Jin, Z. (2007). A methodology for information quality assessment in the designing and manufacturing processes of mechanical products: In L. Al-Hakim (Ed.), Information Quality Management (pp. 190–220). IGI Global. https://doi.org/10.4018/978-1-59904-024-0.ch009

Swedlow, J. R., Kankaanpää, P., Sarkans, U., Goscinski, W., Galloway, G., Malacrida, L., Sullivan, R. P., Härtel, S., Brown, C. M., Wood, C., Keppler, A., Paina, F., Loos, B., Zullino, S., Longo, D. L., Aime, S., & Onami, S. (2021). A global view of standards for open image data formats and repositories. Nature Methods, 18(12), 1440–1446. https://doi.org/10.1038/s41592-021-01113-7

Taleb, I., Dssouli, R., & Serhani, M. A. (2015). Big data pre-processing: A quality framework. 2015 IEEE International Congress on Big Data, 191–198. https://doi.org/10.1109/BigDataCongress.2015.35

Tayi, G. K., & Ballou, D. P. (1998). Examining data quality. Communications of the ACM, 41(2), 54–57. https://doi.org/10.1145/269012.269021

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

van der Sluijs, J. P., & Wardekker, J. A. (2015). Critical appraisal of assumptions in chains of model calculations used to project local climate impacts for adaptation decision support—The case of Baakse Beek. Environmental Research Letters, 10(4), 045005. https://doi.org/10.1088/1748-9326/10/4/045005

Van Der Sluijs, J. P., Craye, M., Funtowicz, S., Kloprogge, P., Ravetz, J., & Risbey, J. (2005). Combining quantitative and qualitative measures of uncertainty in model-based environmental assessment: The nusap system. Risk Analysis, 25(2), 481–492. https://doi.org/10.1111/j.1539-6924.2005.00604.x

Wallace, D. P. (2007). Knowledge management: Historical and cross-disciplinary themes. Libraries Unlimited.

Wand, Y., & Wang, R. Y. (1996). Anchoring data quality dimensions in ontological foundations. Communications of the ACM, 39(11), 86–95. https://doi.org/10.1145/240455.240479

Wang, R. Y. (1998). A product perspective on total data quality management. Communications of the ACM, 41(2), 58–65. https://doi.org/10.1145/269012.269022

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. Journal of Management Information Systems, 12(4), 5–33. https://doi.org/10.1080/07421222.1996.11518099

Whitfield, P. H. (2012). Why the provenance of data matters: Assessing fitness for purpose for environmental data. Canadian Water Resources Journal / Revue Canadienne Des Ressources Hydriques, 37(1), 23–36. https://doi.org/10.4296/cwrj3701866

Wierling, A., Schwanitz, V. J., Altinci, S., Bałazińska, M., Barber, M. J., Biresselioglu, M. E., Burger-Scheidlin, C., Celino, M., Demir, M. H., Dennis, R., Dintzner, N., el Gammal, A., Fernández-Peruchena, C. M., Gilcrease, W., Gładysz, P., Hoyer-Klick, C., Joshi, K., Kruczek, M., Lacroix, D., … Vasiljevic, N. (2021). Fair metadata standards for low carbon energy research—A review of practices and how to advance. Energies, 14(20), 6692. https://doi.org/10.3390/en14206692

World Meteorological Organization (WMO) No. 1238. (2019). Manual on the High-quality Global Data Management Framework for Climate (2019 edition).

World Meteorological Organization (WMO) No. 1221. (2018). Guidelines on Quality Management in Climate Services (2018 edition).

Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., & Auer, S. (2015). Quality assessment for Linked Data: A Survey: A systematic literature review and conceptual framework. Semantic Web, 7(1), 63–93. https://doi.org/10.3233/SW-150175

Zhu, X., & Gauch, S. (2000). Incorporating quality metrics in centralized/distributed information retrieval on the World Wide Web. Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '00, 288–295. https://doi.org/10.1145/345508.345602

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

ANNEX I: Terms and definitions

An agreed list of terminology adopted within the TF. It establishes a common ground, minimising miscommunication and ambiguity and ensures consistency when referring to specific elements.

Term	Definition	Source	Related terms
Accreditation	Third-party attestation related to a conformity assessment body, conveying formal demonstration of its competence, impartiality and consistent operation in performing specific conformity assessment activities. A designating authority provides the accreditation, an organisation established within the government or empowered by the government designating conformity assessment bodies and suspending or withdrawing their designation. The designation is an authorization of a conformity assessment body to perform specified conformity assessment activities.	ISO 17000:2020	Certification, conformity assessment, audit
Audit	Systematic, independent, and documented process for obtaining objective evidence and evaluating it objectively to determine the extent to which the audit criteria are fulfilled.	ISO 9001:2015	Certification, conformity assessment
Certification	Third-party attestation related to an object of conformity assessment, except accreditation.	ISO 17000:2020	Accreditation, conformity assessment, audit
Conformity assessment	Demonstration that specified requirements are or are not fulfilled. Conformity assessment activities include testing, inspection, validation, verification, certification, and accreditation.	ISO 17000:2020	Validation, verification, certification, accreditation
Consensus	The general agreement is characterised by the absence of sustained opposition to substantial issues by any significant part of the concerned interests and by a process that involves considering the views of all	ISO/IEC Guide 2:2004	Standard

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

	parties concerned and reconciling any conflicting arguments. Consensus does not imply unanimity.		
Curation	Data curation is the active and ongoing data management through its lifecycle of interest.	Shreeves et al. (2008)	
Dataset	A collection of organised, discrete items of related data. Data are usually presented as a group of files containing information on the same object(s) – the objects are generally variables or indicators. Datasets can have different formats. However, it shall be noted that the notion of "dataset" found in the literature cannot provide a precise formal definition, but that this general notion is characterised by an interrelated family of more specific concepts: grouping, content, relatedness, and purpose.	Renear et al. (2010); https://whatis.techtarget.com/definition/data-set; https://searchsqlserver.techtarget.com/definition/data-structure	
Dimension (quality)	An attribute represents a single aspect or constructs of quality, e.g., accuracy or completeness. Quality dimensions allow complex areas of data quality to be subdivided into groups, each with its way of being measured.	Wang and Strong (1996); https://www.sciencedirect.com/topics/computer-science/data-quality-dimension	Quality
Normative standard	A generic (general or non-specific) standard or guideline to be used as a framework by local standard-setting or certification bodies when formulating a specific standard for their certification programme. Normative standards are also referred to as Standards for Standards, e.g., the IFOAM Basic Standards and FAO/WHO Codex Alimentarius guidelines.	https://www.fao.org/3/y5136e/y5136e04.htm#TopOfPage	Standard
Pre-standard	A document adopted provisionally by a standardising body and made available to the public so that the necessary experience may be gained from its application on which to base a standard.	ISO/IEC Guide 2:2004	Specification, standard
Quality	The degree to which a set of inherent characteristics fulfils requirements.	ISO 9000, 19157;	Dimensions, requirements

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

		Fürber and Hepp (2016)	
Quality assurance	Processes for maintaining a desired level of quality in a dataset or collection. Data quality assurance is a proactive process focused on "preventing defects". In contrast, quality control is a reactive process focused on "detecting defects".	ISO 9000:2005; WMO No. 1238	Quality control, quality management
Quality control	A reactive process focused on "detecting defects". In contrast, quality assurance is a proactive process focused on "preventing defects". Quality control encompasses a set of procedures applied to detect and identify the errors made in recording, manipulating, formatting, transmitting and archiving data.	ISO 9000:2005; WMO No. 1238	Quality assurance, quality management
Quality management	A set of coordinated activities, tasks and policies are required to ensure that data maintain a required standard of excellence. Quality management involves quality planning, establishing and continued operation of a quality assurance system, including adequate quality control, quality assessment and improvement processes.	WMO No. 1238; ISO 8000	Quality control, quality assurance
Specification	A document similar to a standard that does not require complete consensus and the involvement of all stakeholders. Specifications are strategic instruments for quickly and easily establishing and disseminating innovative solutions. Any specification can be used as a basis for developing a full standard.	https://www.din.de/en/about-standards/a-brief-introduction-to-standards; https://www.iso.org/deliverables-all.html	Standard, pre-standard
Standard	A document, established by consensus and approved by a recognized body, provides, for common and repeated use, rules, guidelines or characteristics for activities or their results to achieve the optimum degree of order in each context. The broad participation of all stakeholders, a transparent development process and the consensus principle ensure the wide acceptance of community-accepted	ISO/IEC Guide 2:2004; https://www.din.de/en/about-standards/a-brief-introduction-to-standards	Specification, consensus, pre-standard

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

	standards. Experts regularly review standards to ensure they reflect state-of-the-art. Governmental standards, providing binding legislative rules, are usually called regulations.		
Usability	Usability describes how easily the data product may be understood and used by users and incorporated into a user's working environment. It includes aspects of compatibility of the publication medium with community standards and supporting documentation. The concept of usability should not be confused with fitness-for-purpose.	WMO No. 1238	
Validation	Demonstration of the alignment of data values concerning relevant external benchmarks. In validation, expectations are derived from comparisons to known or relative gold standards and external knowledge that exists as resources independent of the evaluated data source.	Kahn et al. (2016)	Verification
Verification	Demonstration of how data values match expactions with respect to metadata constraints, system assumptions, and local knowledge. It checks intrinsic consistency, such as adherence to a format or specified value range. In contrast with validation, verification does not rely on an external reference or benchmark.	Kahn et al. (2016)	Validation

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

ANNEX II: Recommendations for the future EOSC Task Force

Additional recommendations more specific for the future **evolution of this EOSC Task Force**:

- **Split the TF**: the Data Quality subgroup faces challenges different from the FAIR Metrics subgroup. In addition, Dataset Quality is a preferred name compared to Data Quality because this TF deals with more aspects than only dataset content values, but also documentation, metadata, process, and ecosystem associated with the content values are considered.
- **We need to incorporate members from the most regulated fields** (e.g., Pharma, Medical Devices, Aerospace) to expose the group to solid operational protocols in data quality. We believe that the experience accumulated by industry needs to be considered to maximise impact for researchers and their discoveries. Moreover, the survey this TF issued showed that EOSC has little penetration in the disciplines of agriculture and chemistry. Requirements from these disciplines have not been explored yet.
- **Create a catalogue of community tests/methods to apply in quality analyses**. The survey revealed that one of the most significant barriers to providing quality-assessed data is the "need to reinvent quality criteria for data generated through innovative methods."
- **Define a vocabulary of terms most relevant in the context of EOSC.** A common terminology should be shared and kept improving to (i) facilitate consistency within EOSC, (ii) give a reference for the users to consult when jargon or acronyms are found, and (iii) minimise ambiguities boosting common ground understanding across communities.
- The next TF will need to **define the exact requirements to assess data quality**; so far, we have identified a number of minimum requirements reported in section 3.4. These will need to be refined, and the exact standards to follow will need to be identified, prioritising those applicable across domains or supporting FAIR data sharing (standards structuring metadata, semantics, and provenance). Make clear the quality requirements in the form of checklists, maturity models, or automatic validator tools. There is a preference for automated tools to be made available as much as possible. It will be necessary to identify exact requirements like data format or vocabulary so that the data providers shall be asked explicitly that data is published in EOSC according to specific requirements. This conclusion is adapted from the EOSC interoperability framework.
- **Keep defining a common ground understanding of data quality** in a multidisciplinary environment. Also, explore the level of dataset granularity where users need quality information.
- We concluded that the most urgent gap the communities face in completing agreed standards is the **lack of controlled vocabularies** and terminologies. Without agreed standards, data quality cannot be appropriately demonstrated and data interoperability is undermined. Moreover, it could be interesting to keep investigating how to gauge disciplines maturity when dealing with quality management. The first steps in this direction were taken and are reported in section 3.5.

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

ANNEX III: Survey details

The survey clustered questions in three thematic blocks detailed below:

Some information about you

- Your country of residence? [droplist]
  - List of all countries
- Which best describes your community?
  - Medical and health sciences
  - Biology and life sciences
  - Metrology
  - Earth sciences
  - Astronomy
  - Chemistry
  - Engineering and technology
  - Social sciences
  - Law
  - Humanities and the arts
  - Library sciences
  - Information and computer sciences
  - Agriculture and veterinary sciences
  - Theology
  - Education
  - Other (please specify)
- In what capacity are you responding? [multiple choices]
  - Data(set) producer
  - Data(set) steward or custodian or curator
  - Data user
  - Data quality practitioner
  - Publisher
  - Service or data(set) provider
  - Policy/decision-maker
  - Standard maker
  - Funder
  - Other (please specify)
- Your career stage?
  - < 5 years of working experience
  - >= 5 years and < 10 years of working experience
  - >= 10 years and < 20 years of working experience
  - >= 20 years of working experience
- Your organisation type?
  - Academia / research
  - Government / administration / public services

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

- Policy / funding agency
- Private IT consultancy / development
- Library
- Other private enterprises

Evaluating and ensuring data quality

- What information do you consider most essential to preoperly use or select a dataset? [Buttons with “mandatory,” “truly relevant,” “somewhat relevant,” “not relevant,” and “I do not know”]
  - User guide (including a description of size, structure, abstract, typical usage, production methodology, and dictionary)
  - Compliance of metadata with community standards
  - Format according to community standards
  - Scientifically accurate (e.g., validated against a reference, plausible)
  - Technically correct / having passed sanity checks (e.g., no unexpected gaps in the time series, consistency between data and metadata)
  - Details on strengths, limitations and known issues, including the availability of information about uncertainty
  - Evidence of regard to ethical conduct, protection, and confidentiality of data
  - Data are complete in space and time, with adequate resolution (if applicable)
  - Up to date / currency, timeliness / novelty
  - Provenance and traceability information
  - Clarity about how to cite the dataset and the availability of its product locator, like URL or DOI
  - Information about the data provider and point of contact
  - The licence of use, including terms of use
  - Archiving policy
  - Version
  - Evidence of data reuse
  - Security
  - Other (please specify)
- Have you tried to reuse someone else’s data from a repository? Did you discover any quality issues? Y/N If yes, what actions did you take?
  - Used as it is
  - Cleaned/pre-processed data before use
  - Seek further information
  - Inform the provider before acting
  - Do not use
  - Other (please specify)
- If any correction of the data you provide is needed, how do you manage such issues [multiple choices]
  - Improve the data (new calculation)
  - Describe weaknesses or errors in a transparent manner

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

    - Keep such errors in mind and plan for the next generation of this dataset
    - Provide feedback to the team (usually the data producer) in charge of rectification. (It may be the case if I am a data service provider and not the original producer)
    - Discard the data
    - Other (please specify)
- What was your most significant concern/barrier to providing quality assessed data? [buttons with “biggest,” “very important,” “somewhat important,” “not important,” and “I do not know”]
    - Time-consuming
    - Missing expertise
    - Missing previous experience/guidance
    - Missing tools for evaluation
    - High complexity
    - No access to standards
    - No benefit
    - The financial cost of certification/accreditation
    - Other (please specify)
- Do you need to comply with any legal requirements (e.g., “General Data Protection Regulation” GDPR) when producing data? Y/N/I do not know, if Yes -> Which legal requirements does your discipline usually follow when producing data (e.g., GDPR compliance)?
- How do you prefer quality information to be supplied?
    - In metadata within files
    - Documents external to files
    - Dynamic documents linked in the metadata of the files
    - Dynamic documents are available on the website where the dataset files are available
    - Other (please specify)
- Would you be happy for some ranking based on quality assessments, to be applied to your dataset? Please leave comments explaining the conditions which should apply for this to be suitable.
- Rate how much you agree with the following: “Information about valid ranges of plausibility is necessary to evaluate the scientific value of the datasets.” [Button with “Highly agree,” “agree,” “not much,” “not at all,” “I do not know”] Could you list any physical quantity missing information about the valid ranges?

Landscape and processes

- Which standardisation bodies are most active in your field (e.g., ISO, OGC, W3C)?
- Which organisation do you consider a reference or leader for quality directives in your field (e.g., WHO, WMO, EFSA)?
- Does your organisation have anybody responsible for data quality? Y/N/ I do not know. If Y -> Who or what department team is responsible for data quality in your organisation?

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

- Does your discipline have community-recognized metadata standards? Y/N/I do not know. If yes -> Which ones?
- Rate how much you agree with the following: “my community has a recognized structured vocabulary that is used to refer uniquely to specific quantities. In other words, there is a widely accepted definition of the most common variables.” [Button with “Highly agree,” “agree,” “not much,” “not at all,” “not the case / not applicable,” “I do not know”]
- My community lacks the following:
    - Agreed controlled vocabularies and terminologies
    - Agreed data formats
    - Agreed metadata schemas
    - None of the above
    - I do not know
    - Other (please specify)
- Are the data (excluding metadata) quality principles/practices/standards internal to your organisation, or are community-recognized?
    - Internal to my organisation
    - Community-recognized
    - Both
    - I do not use any principle/practice/standard
- Are there data (excluding metadata) of community-recognized standards in your field? No, yes one -> which one, Yes more than one -> Why do you think some are used, and some are not?
- What level of data quality management do you expect from EOSC? [Multiple choices]
    - No need for a data quality function; content is distributed as deposited
    - Provide a flag or star system to infer the level of data quality
    - Allow (re)users to rate or leave comments on data quality
    - Basic curation: e.g., control availability of basic metadata or documentation, basic metadata compliance checks
    - Enhanced curation: e.g., enhancement of documentation, the performance of independent assessments
- What is the most critical need for data quality information from a data consumer perspective? [Ranking]
    - Data quality information should better guide the user in selecting a dataset
    - Data quality information must be available and consistent across multiple datasets and repositories
    - Data quality information should be interoperable (e.g., by following standard machine-readable schemes to capture quality information)
    - Dissemination of data quality information should be improved and made more understandable
    - Other (please specify)
- What practices/tools should a discipline have to gauge its maturity in data quality management? [Button with “Highly agree,” “agree,” “not much,” “not at all,” “I do not know”]

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

    - Existence of metadata community-recognized standards
    - The existence of agreed definitions, for instance, when describing a physical quantity
    - Availability of a standard quality management framework
    - Availability of community-recognized metrics to quantify the quality
    - Quality assessments are operational routine and funded
    - Availability of checklists
    - Availability of certifications
    - Established data governance
    - Existence of a ranking system
    - Capacity building programme
    - The feedback loop between supply and demand
    - Availability of automatic tools for quality assessment
    - Quality assessment is part of the data production and not a separate cumbersome process
    - Other (please specify)
- Do you have any expectations or requirements regarding data quality as a service manager, funder, data provider, or user of EOSC? *Some guiding questions: Which dataset lifecycle stage should EOSC consider? Which dataset characteristics are important to consider (e.g., archiving, DOI/PID, metadata) and prioritise? How do you want data quality to be measured (e.g., maturity models, checklists, statistical analyses, minimum requirements)? Any recommended standards to adopt? How shall quality be disseminated (e.g., KPIs)? Who improves dataset quality?*
- Any other comments you wish to make? Y/N followed by free text

EOSC Association AISBL
Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931