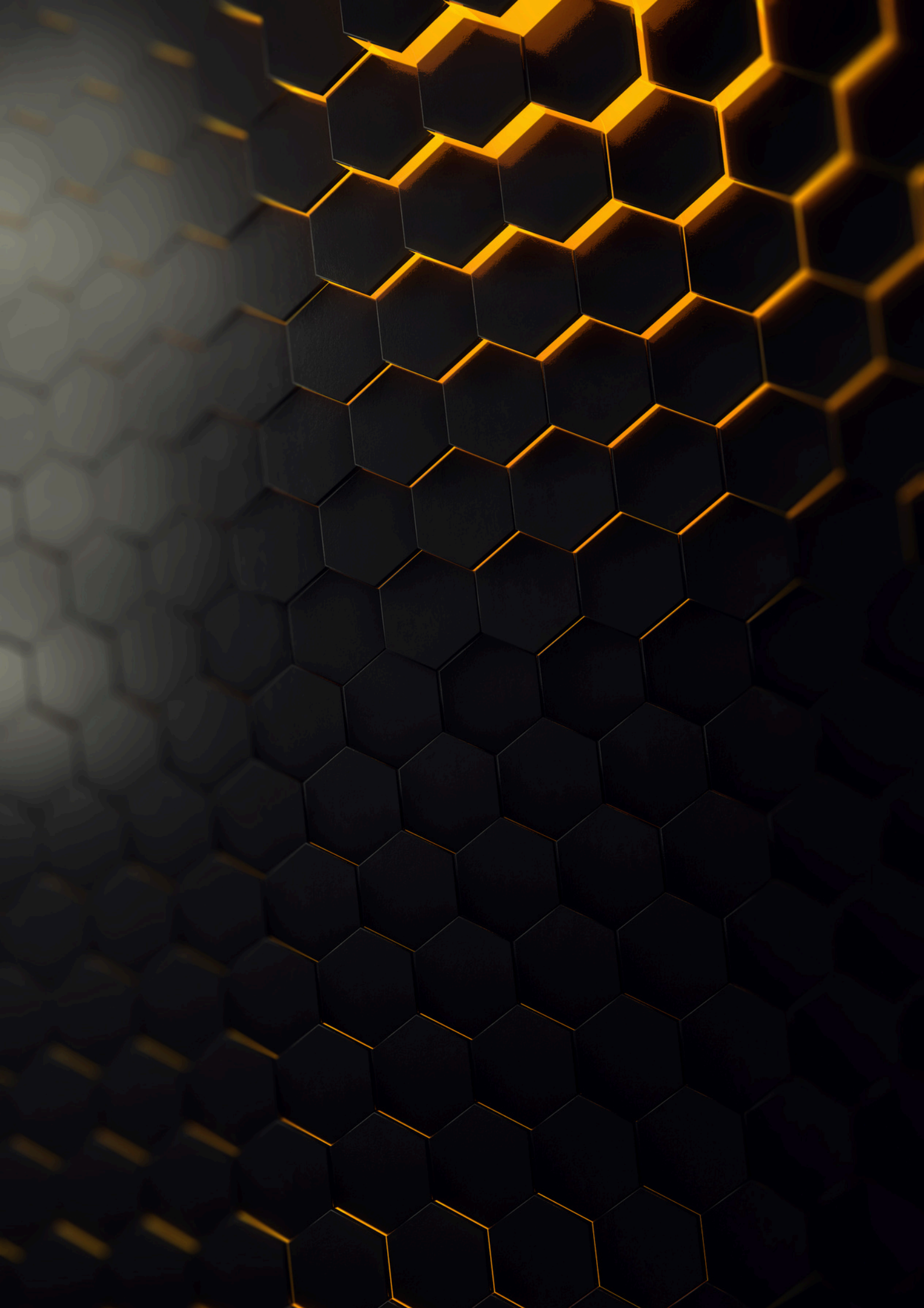
BDV BIG DATA VALUE ASSOCIATION

ELEVATING DATA QUALITY: A PARADIGM SHIFT FOR DATA SPACES AND AI NEEDS

This paper is the result of cooperative work by a group of BDVA Task Force Data Spaces members and representatives from collaboration partners. Published by BDVA in May 2024.



LIST OF FIGURES

LIST OF TABLES

CHAPTER I

INTRODUCTION, MOTIVATION AND SCOPE OF THE PAPER

In a generic sense, data quality refers to the properties of data in relation to specific standards and criteria. It reflects the extent to which data accurately represents the reality it seeks to capture. Quality affects most stages in the lifecycle of data and is a concern for most actors involved in the data value chain, coming from different domains. Ultimately, the quality of the data strongly affects the amount of value that can be generated out of it. Given its importance, data quality is a topic broadly covered in the literature, from different perspectives. There exist different approaches to data quality, each one with different dimensions and different metrics.

With the increasing importance of data sharing in the data value chain, quality has a key role in the interoperability, reliability and usability of the shared information. Moreover, the scaling-up of data sharing to data spaces, interconnected and collaborative environments where multiple participants access and share data, implies additional levels of quality and trust to ensure the adoption and success of such environments. In this sense, data quality plays a key role in ensuring that shared data is, among other features, reliable, trustworthy and fits the purposes of the collaboration in the data space. At the same time, data spaces appear as unique environments that provide the tools, mechanisms and governance framework needed to assess and ensure data quality.

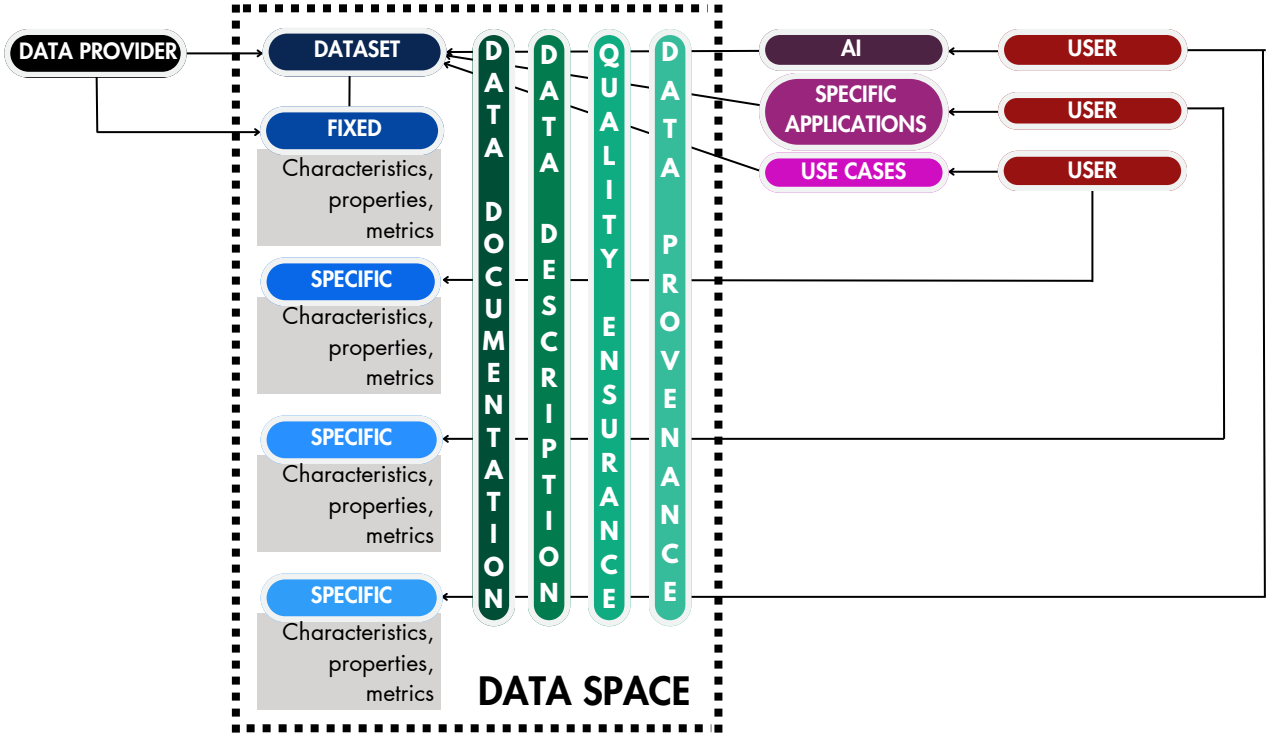


Figure 1. Data quality as addressed in this document

In this increasingly complex ecosystem data provider and data user are often several steps apart or even unknown to each other and the fit of the data for its intended purpose (beyond its original properties) is gaining more and more relevance. Driven by advancements in AI technologies, the volume of data being utilised for AI applications has been growing rapidly in recent years and its significance is likely to continue growing as AI technologies become more integrated into various aspects of society and industry, how data meets specific requirements in terms of quality imposed by those AI applications, as already stated in the AI Act, is crucial.

Based on the above, the objectives of this document are:

- Present ***data quality from its different perspectives and*** gather the information and material needed to pave the way for the subsequent discussions.
- Explore the ***symbiotic relationship between data quality and data spaces***, highlighting the importance to incorporate data quality in all aspects of data spaces (quality-by-design), but also presenting data spaces as a unique environment to ensure data quality when sharing data in a scalable way.
- Drive the ***fit-for-purpose paradigm to focus on the data quality for AI***, which introduces AI requirements that should be addressed by specific metrics and processes, also according to the AI Act.
- Provide some ***recommendations and paths to follow*** in order to fully achieve those goals and shift data quality to a new dimension.

This document is intended for all actors involved in the lifecycle of data, including data providers and data users and with special emphasis on value generation from data through use cases and from different industry sectors and domains. It also addresses data spaces designers and builders to incorporate data quality by design and reinforce the role of data spaces in keeping and even increasing data quality. Finally, the document is also relevant for AI practitioners, who bring the AI dimension into the picture and for whom the quality of data used to train, validate and test models is of paramount importance.

Authors

This paper is the result of cooperative work by a group of Big Data Value Association (BDVA)* members and representatives from collaboration partners.

Main editor: Daniel Alonso (BDVA)

Main contributors (in alphabetical order):
·Aitor Corchero Rodriguez (NTT Data)
·Amelie Gyrard (Trialog)
·Ana Isabel Torre Bastida (Tecnalia)
·Arturo Medela (Eviden)
·Elias Tragos (Insight Centre for Data Analytics, University College Dublin)
·Franck Le Gall (EGM)
·Gabriel Danciu (Siemens)
·Gerasimos Antzoulatos (CERTH)
·Ilaria Baroni (Cefriel)
·Jordi Arjona (ITI)
·Kuldar Aas (Ministry of Economic Affairs and Communications for Estonia)
·Luis Sánchez (University of Cantabria)
·Maria Jose Lopez Osa (Tecnalia)
·Mark Dietrich (EGI)
·Maroua Bahri (INRIA)
·Matteo Falsetta (GFT)
·Mihnea Tufis (Eurecat)
·Patrick van der Smagt (etami, Volkswagen Group)
·Peter Brosten (Eurecat)
·Rafael Martínez (Gradiant)
·Torben Jastrow (Fraunhofer-FOKUS)
·Tuomo Tuikka (VTT)
·Yury Glikman (Fraunhofer FOKUS)

Reviewers:
·Ana García (BDVA)
·Antonis Ramfos (ATC)
·Freek Bomhof (TNO)
·Shane O'Seasnáin (Eindhoven University of Technology)

Final public publishing: May 2024. Designed by Daniel Djamo (BDVA).

* https://www.bdva.eu

CHAPTER II

DATA QUALITY

This section presents a compendium of the different current conceptions of data quality, dimensions, metrics and descriptions. It highlights the increasing importance of the fit-for-purpose aspect of data quality, whose focus on AI and ML is further explored in Section 3. The section also examines the challenge that data sharing imposes in data quality, as an introduction to data spaces as unique environments to address this challenge as explained in Section 4.

II.1 Definition, characteristics, properties, metrics

Depending on the scope, domain and specific objectives of the organisation and application using the data, the term data quality can adopt different definitions. Table 1 shows some definitions of data quality that can be found in some standards, regulations and literature.

Table 1. Definitions of Data Quality

ISO 8000-2:2022[1]	Degree to which a set of inherent characteristics of an object fulfils requirements (need or expectation that is stated, generally implied or obligatory).
ISO/IEC 25012[2]	Degree to which the characteristics of data satisfy stated and implied needs when used under specified conditions. (Quality of a Data Product) The degree to which data satisfy the requirements defined by the product-owner organisation. Specifically, those requirements are the ones that are reflected in the Data Quality model through its characteristics (Accuracy, Completeness, Consistency, Credibility, Currentness, Accessibility...)"
AI Act[3]	Training, validation and testing data sets [for high-risk AI systems] shall be relevant, sufficiently representative and to the best extent possible, free of errors and complete in view of the intended purpose.
European Health Data Space[4]	Degree to which the elements of electronic health data are assessed and considered suitable for their intended primary and secondary use, Data quality and utility label' means a graphic diagram, including a scale, describing the data quality and conditions of use of a dataset.
Data Quality - Concepts and Problems [1]	Data Quality is the degree to which the data of interest satisfies the requirements, is free of flaws, and is suited for the intended purpose. Data Quality is usually measured utilising several criteria, which may differ in terms of assigned importance, depending on the data at hand, stakeholders, or the intended use.

1 https://www.iso.org/standard/85032.html
2 https://www.iso.org/standard/35736.html
3 https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF
4 https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space_en

Each definition of data quality in Table 1 is accompanied by a set of characteristics and properties of the data quality that varies depending on their context and purposes. For example, the standard ISO/IEC 25012 proposed the following general categories for data quality characteristics:

- "Inherent data quality" refers to the degree to which quality characteristics of data have the intrinsic potential to satisfy stated and implied needs when data is used under specified conditions.
- "System dependent data quality" refers to the degree to which data quality is reached and preserved within a computer system when data is used under specified conditions.

This differentiation of data quality into internal and external categories is the basis of the understandings of data quality which are further elaborated in Section 2.5.

For each category, the characteristics shown in Figure 2 are identified and for each characteristic, ISO/IEC 25024[5] defines several properties and a metric for each property with the objective of evaluating and quantifying the degree to which a data set meets the criteria of the quality model established in the ISO/IEC 25012.

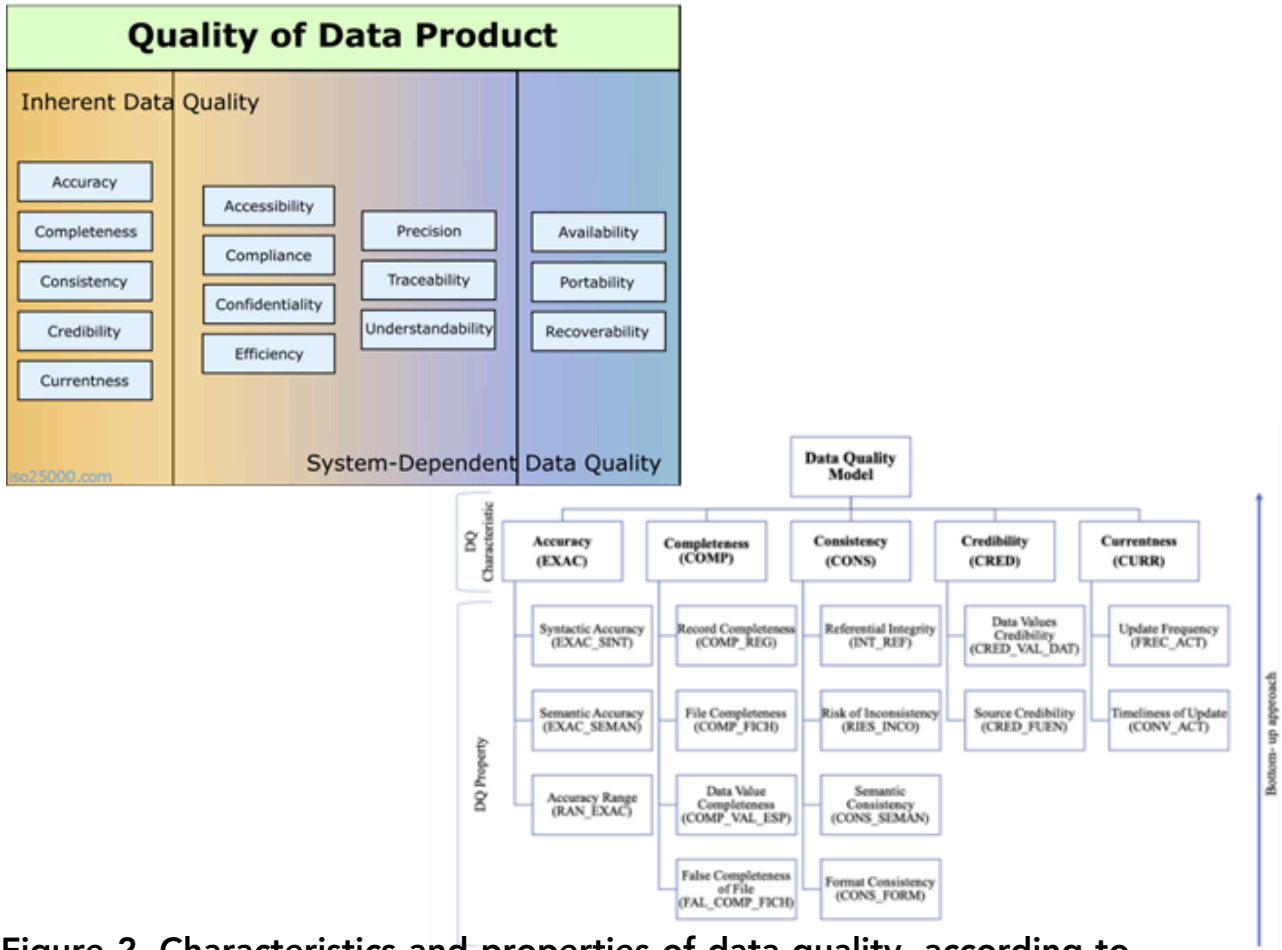


Figure 2. Characteristics and properties of data quality, according to ISO/IEC 25012 / 25024

An example of the metric for the property "data accuracy range" can be seen in Table 2 (directly taken from [2]):

5 https://www.iso.org/standard/35749.html

Table 2. Metric to measure data accuracy range

Description	Data Accuracy Range focuses on checking if data values are included in the required intervals. Its value is obtained as the ratio of records in a data file whose values for their fields are within the specific intervals.
Formula	X = A / B A = number of items having a value included in a specified interval (i.e. range from minimum to maximum) B = number of data items for which can be defined a required interval of values
Value range	[0.0, 1.0]

It is worth pointing out that characteristics such as accuracy and completeness can be applied to the dataset as a whole or to each data point individually, while certain overall characteristics, such as consistency, are computed based on calculations using the metrics per data point.

II.1.1 Data quality metrics aggregation and reporting

Data quality metrics aggregation is a relevant aspect for processes with high-dimensional outputs. For a structured or semi-structured dataset, quality assessment can be performed for each field and then aggregated for the entire dataset. However, quality assessment itself is multidimensional and can be performed in a heterogeneous way across the structure of a data set (e.g., a user can decide to measure completeness for Field-1, timeliness and accuracy for Field-2 etc.). In the context of data spaces, practitioners might soon need to reflect on further levels of aggregating data quality dimensions beyond individual datasets, to generate insights for combined datasets and eventually develop metrics to characterise, in terms of data quality, the entire data space.

The challenge is to eventually create aggregated scores incorporating all the metrics included in the assessment process and report them in a usable manner. Ideally, data quality reports should start from aggregated metrics and then offer drill-down into lower level views of data quality dimensions. This would result in a better understanding by data space users of the strengths and weaknesses of data sets and offer suggestions for data quality mitigation. Such development of reporting tools should follow a user experience design approach and consider user flows and requirements formulated by stakeholders from a variety of data spaces.

II.1.2 Data quality assessment

Data Quality assessment can be defined as the process of evaluating data according to certain criteria to find whether it meets some minimum requirements or principles. Accordingly, the assessment of the data quality is crucial and organisations can use various tools and methods to ensure that the data meets their specific requirements.
Data and metadata quality assessment can be implemented in three ways:

- ***Automatic approach***: rules and patterns are inferred from data and metadata. They are then applied to the dataset resulting in a quantification of the dataset against the inferred rules.

- ***Semi-automatic approach***: uses machine learning to automatically learn quality rules and patterns, which can then be manually validated by the user, before data quality assessment is performed.

- ***Manual approach***: a formal definition of each of the data quality characteristics allows us to measure if the data complies with them. This results in a quantification of how much the dataset is aligned with a certain characteristic. Such an approach, while allowing for a very fine-grained definition of the characteristic, also requires that an exact description of the data quality rules (via a Data Quality Language) is provided.

Some of the most common tools for Data Quality assessment and improvement include Pivau, Ydata-quality, Apache Griffin, Great Expectations, Pandera, Deequ /PyDeequ, IBM Data Quality for AI, EUT Data Quality Assessment module, Data Quality Analyzer, DataCleaner[6] or OpenRefine.[7] Detailed information about some of those tools can be found in Annex I.
These tools can help to identify, analyse and report data quality issues such as missing or inaccurate data.

6 https://datacleaner.github.io/
7 https://openrefine.org/

II.1.3 Additional situations and aspects to be considered in data quality

- There are many challenges when dealing with ***assessing or improving quality in data streams*** [9]. Normally data quality improvement methods for data streams (i.e. for anomaly detection) either operate in batches, gathering a small group of data points and processing them, or they operate separately on each data point, which isn't always optimal. In most cases, the processing of the data stream must be done extremely quickly, so that it doesn't affect the transmission rate of the stream. This means that there will be a single pass over the data points. The fast processing of the data points also requires that they will be stored in memory and on-the-fly converted into an appropriate format for processing via the data quality tools. Additionally, a major challenge is posed by concept drift, where the data distribution changes over time, making algorithms devised for static data ineffective. Key considerations and strategies for maintaining quality in streaming data are described with more detail in section 9.1.

- In the context of data quality assessment, ***ensuring data quality in motion*** from collection to sharing involves continuous monitoring, proactive alerting and the implementation of quality assessment and profiling practices. This process is essential for maintaining the reliability, accuracy and timeliness of the data as it moves through the data lifecycle. Some techniques are presented in section 9.2.

- The importance of ***precision timestamping*** is increasing in financial transactions, electricity transmission grids, computer network synchronisation, industrial automation and scientific research. It is highly important to know the order of events taking place in micro and nanoseconds to avoid confusion in stock exchange and malfunction in computer networks. For fare billing, energy efficiency and safety in smart electric grids, accurate time synchronisation is important. The proper function of automation systems and robots relies on measurement data with high quality timestamping. In climate research, time series are studied to make reliable conclusions on climate change.

- ***Reliable measurement*** data is crucial for well-functioning global trade, safety and efficiency e.g. in industrial manufacturing, traffic and energy production and decision making essential for tackling the global challenges of climate change and limited raw materials. The global network of metrology institutes and service providers ensures that globally recognised calibrations are available everywhere for providing quality information for each measurement dataset. Without this measurement uncertainty information, reliable conclusions about differences between datasets cannot be drawn, which jeopardises the usability of the datasets.

- Spatial variation of data has an important impact when data is gathered from several points along a process and which may have relatively varying degrees of sparseness and abundance. Orchestrating data into meaningful insights requires data that has a sufficient level of quality and proximity to the topic of interest, to support decision-making by downstream domain experts and systems.

II.2 Data quality descriptions

II.2.1 Data Quality Vocabulary

The Data Quality Vocabulary (DQV[8]) is a metadata model that provides a framework for expressing data quality. It was developed by the Data on the Web Best Practice (DWBP) Working Group of the World Wide Web. The DQV does not provide a formal, complete definition of quality, but it sets out a consistent means by which information can be provided to describe the quality of a dataset. It allows for the description of the quality of a dataset by the dataset publisher or by a broader community of users.

DQV is based on the Data Catalogue Vocabulary (DCAT[9]) that is extended with several additional properties and classes suitable for expressing the quality of a dataset. DQV includes constructs like quality measurements, feedback and annotation and it is designed to be used at every step of a dataset's lifecycle, making it easier to publish, exchange and consume quality metadata.

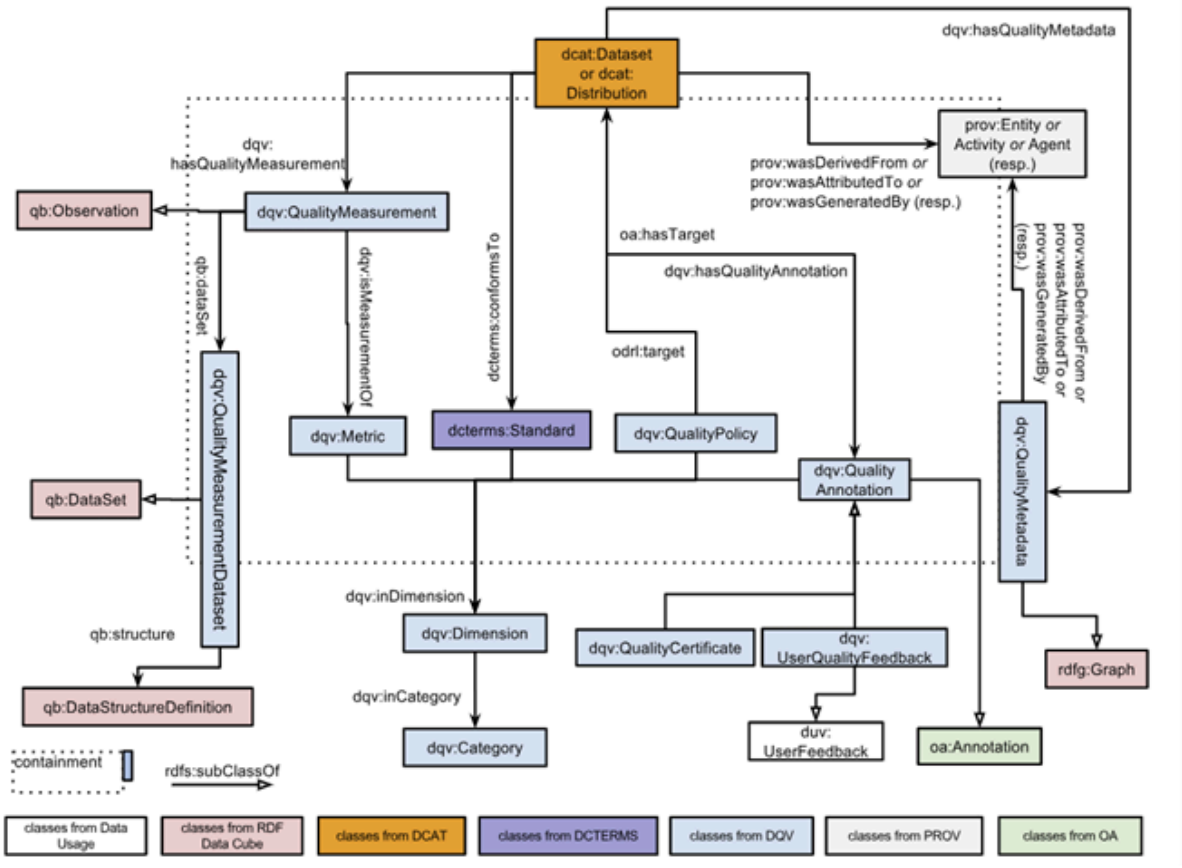


Figure 3. DQV: Data model showing the main relevant classes and their relations

8 https://www.w3.org/TR/vocab-dqv/
9 https://www.w3.org/TR/vocab-dcat-2/

The quality of a given dataset or distribution is assessed via several observed properties. To express these properties an instance of a dataset or a distribution can be related to five different types of quality information represented by the following classes:

- Quality Annotation represents feedback and quality certificates given about the dataset or its distribution.
- Standard represents a standard the dataset or its distribution conforms to.
- Quality Policy represents a policy or agreement that is chiefly governed by data quality concerns.
- ·Quality Measurement represents a metric value providing quantitative or qualitative information about the dataset or distribution.
- Entity represents an entity involved in the provenance of the dataset or distribution.

The vocabulary includes the specification of the different classes and the properties used with instances of each class. It also includes a link to the quality dimensions in ISO / IEC 25012 and to the quality dimensions defined for Linked Data [4]. Additionally, Figure 3 shows the main relevant classes and relations in the data quality model of DQV and their connections with other classes coming from DCAT and other W3C standards.

DQV offers significant potential benefits for improving data quality, but its successful implementation may require addressing challenges related to adoption, consistency, training, feedback, interoperability and scalability.

II.2.2 Data Quality Assessment within Smart Data Models

The Smart Data Models (SDM) initiative[10], aims to offer a standardised approach to data representation across different domains. Its core data representation is based on the NGSI-LD specification produced by ETSI ISG CIM[11]. It aims to enhance interoperability between diverse systems and applications, thus enabling seamless communication. The Smart Data Models are open source and are developed through constant efforts from the community. The community contributes to creating or updating the existing data models. SDM offers a customisable framework suitable for diverse domains, allowing for the creation of multiple domain-specific data models that cater to applications or datasets. The data models that are aggregated and maintained under this initiative complement the NGSI-LD standard information model[12] by providing domain-specific models that can be widely adopted (de-facto standard) for creating a global digital single market of interoperable and replicable (portable) smart solutions in multiple domains.

Among the data models that SDM initiative maintains in its catalogue, the Data Quality model[13] aims at characterising quality properties of several entities or measurements and provide information to consumers on different data quality dimensions of such entities. Currently, this data model includes the description of four Data Quality (DQ) Dimensions and the outcome of two DQ enhancement techniques. The four DQ Dimensions are:

- **accuracy**: accuracy measures the maximum systematic numerical error produced in a sensor measurement. It may take values between 0 and 1.
- **precision**: precision measures the standard deviation of a dataset. That is, it measures how close the values in the dataset are to each other. It may take values between 0 and 1.
- **timeliness**: timeliness measures the update time of sensor measurements. Default unit: minutes.
- **completeness**: completeness quantifies the number of missed measurements or observations in each time window. It may take values between 0 and 1.

More information about Data Quality Assessment withing Smart Data Models can be found in Annex II.

10 https://smartdatamodels.org/
11 ETSI CIM GS009 - "NGSI-LD API" and ETSI CIM GS006 - "NGSI-LD Information Model"
12 https://www.etsi.org/deliver/etsi_gs/CIM/001_099/006/01.02.01_60/gs_CIM006v010201p.pdf
13 https://github.com/smart-data-models/dataModel.DataQuality

II.2.3 Metadata Quality Assessment

The Metadata Quality Assessment (MQA) is a tool developed by the consortium of the European Data Portal (EDP: https://data.europa.eu/en) to study the quality of metadata harvested by the EDP. It is intended to help data providers and data portals to check their metadata quality and to receive suggestions for improvements. The results are presented via the MQA and are also available as download. In the following we describe the functionality of the MQA and the methodology it uses. The dimensions that the MQA examines to determine the quality of datasets that are available at the EDP are derived from the FAIR principles.[14]

-

The MQA Is limited by the metadata it can examine. The investigation is limited exclusively to the metadata that the EDP collects during the harvesting process. If there are errors in the source metadata, these can falsify the overall result. To limit this error potential, the MQA provides a validation service that can be used by data providers to validate their metadata for valid formats and compliant DCAT-AP before integrating it into the harvesting process.

With each harvesting, the metadata is also checked by the MQA. The MQA measures the quality of various indicators, each indicator is explained in the tables below. The results of the checks are stored as Data Quality Vocabulary (DQV).

As accessibility can be volatile, repeated checks for the accessURL and downloadURL are necessary. For this reason, the MQA regularly checks the accessibility of all distributions. In contrast to the verification of the other indicators, this has a higher runtime, since the distributions are checked via HTTP and each requested URL may have a longer response time. The MQA uses a mechanism that considers that each URL is re-examined for accessibility within a few weeks of the last review.

14 https://www.nature.com/articles/sdata201618

II.2.4 Open Data product specifications

The Open Data Product Specification is a vendor-neutral, open-source machine-readable data product metadata model. It defines the objects and attributes as well as the structure of digital data products. The work is based on existing standards (schema.org), best practices and emerging concepts like Data Mesh. Open Data Product Specification 2.1 (ODPS[15]) intends to change the data product metadata model towards a standalone model, which would help to decouple data product from the systems often directly associated with it.

Data quality is one of the attributes used to describe the data product (together with data pricing, DataOps, Data SLA, Data access, data licensing and data holder) and considers accuracy, completeness, consistency, timeliness, validity, uniqueness and quality assurance methods, with description, type and options for each.

15 https://open-data-product-initiative.github.io/open-data-product-spec-2.1/#open-data-product-specification-2-1

II.3 Data quality in the value chain

Data is being transformed from the moment it is collected to the moment it is shared or consumed at the different stages it goes through. Data can be also transformed as it moves from one step to the next (e.g. data used to train neural nets still retains the underlying epistemic characteristics of the original data but now represents its quality as predictive uncertainty and bias). On this regard, it is important to see data as an asset that is recorded and understood at each step, such that a 'data recall" can be executed easily, hence the importance to include data provenance and traceability mechanisms throughout the process.

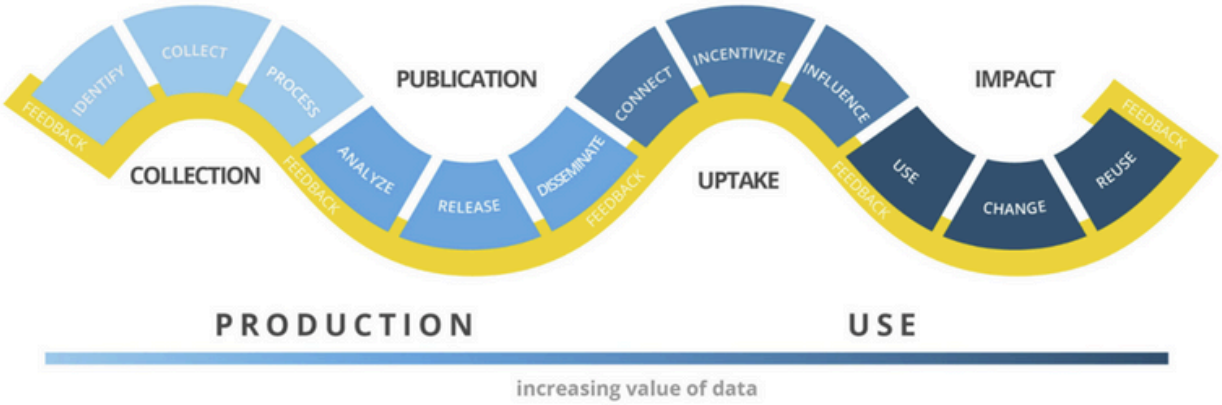


Figure 4. Data Value Chain, according to Open Data Watch

The so-called Data Value Chain (DVC) encompasses those various stages of the data lifecycle, which can differ depending on the methodology or context. While there is no unified or standard version of the data value chain, it is expected to provide a roadmap for transforming raw data into actionable insights and can be used to analyse the complex steps involved in the data lifecycle. A common framework would include, among others, identification, collection, processing, analysis and sharing, ultimately leading to the use and impact of the data (see Figure 4[16]).

In the context of the European Data Strategy[17], data value chains (DVCs) are seen as mechanisms that define a set of repeatable processes to extract data's value, involving data generation, collection, analysis and exchange. DVCs bring together data providers, end users, solution providers and digital innovation hubs to support innovation experiments and the development of trusted and secure privacy-preserving data solutions [3].

16 https://opendatawatch.com/reference/the-data-value-chain-executive-summary/
17 https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy_en

Data quality is one of the major concerns in data management since low quality data often leads to inaccurate analysis results and, eventually, to making wrong decisions. Nevertheless, developing effective and efficient data quality solutions is a challenge fraught with profound theoretical and engineering problems. The focus is usually put on the preparation and analysis phases, as cleansing the data is fundamental to perform a correct analysis and, from a good analysis it is usually assumed that good results will flow. This process may be inefficient, though, consequence of not considering data quality in that chain.

The integration of the concept of data quality into the data value chain is essential for ensuring that the data used in the various stages of the value chain is reliable, accurate and fit for its intended purpose. Relevant aspects to be considered in this integration are the following:

- ***Hybrid approach to quality evaluation***, emphasising the importance of assessing the quality of data at various stages of the value chain.

- ***Addressing data quality*** as early as possible. Businesses are encouraged to move data quality responsibilities to the left of the data value chain, addressing data quality as early as possible in the data lifecycle. This approach can be applied to data quality metrics such as accuracy, completeness and usability, highlighting the integration of data quality considerations into the early stages of the data value chain. More specifically, assessment of data quality is especially important during or before data collection, as data quality processes must inform of how good or bad is our data and this evaluation can, at least, be used with two purposes: evaluate data ingestion mechanisms and plan the data cleansing phase (or hopefully avoid it).

- ***Measuring the impact of data quality***. The data value chain describes the evolution of data from collection to analysis, dissemination and the final impact of data on decision making. It can be used as a teaching tool to show the complex set of steps from data collection to impact, emphasising the importance of ensuring data quality throughout this process.

- Importance of a ***good selection of data quality assessment metrics***. Data quality assessment metrics measure how relevant, reliable, accurate and consistent an organisation's data is, emphasising the need to evaluate data quality across the various stages of the data value chain.

These points collectively demonstrate that ensuring data quality should be integrated throughout the data value chain, from collection to impact and is essential for maximising the value derived from data. Ensuring data quality across the data value chain presents several challenges:

- ***Data proliferation.*** The increasing number of data sources can make it challenging for data teams to manage and identify data quality issues, emphasising the need for data observability tools to detect and solve such issues.

- ***Impact of poor data quality***. Poor data quality can lead to inaccurate reporting, poor decision-making and a loss of trust in data-driven insights, ultimately affecting financial outcomes and stakeholder relationships.

- ***Data standardisation in supply chains***. Data quality issues arise when data is incorrect, incomplete, or outdated, particularly in complex supply chains where standardising data quality across the entire chain is essential.

- ***Lack of priority and awareness***. In some cases, improving data quality may not be a priority for professionals, leading to a lack of awareness and investment in addressing data quality challenges.

- ***Trust and reliability***. High data quality is essential for informed decision-making, reliable reporting and accurate analysis. Without trust in the data, there can be a loss of value and credibility.

II.4 Data quality in data sharing

Data sharing is an integral part of the data value chain, as it involves the exchange and dissemination of data among various stakeholders. The value chain facilitates the understanding of how data is used and transformed into actionable information, highlighting the interconnectedness of the different stages and the constant feedback between data producers and stakeholders. Therefore, the data value chain provides a framework for analysing and optimising data sharing processes within the broader context of the data lifecycle (Figure 5, from Open DEI Design principles in Data Spaces).[18]

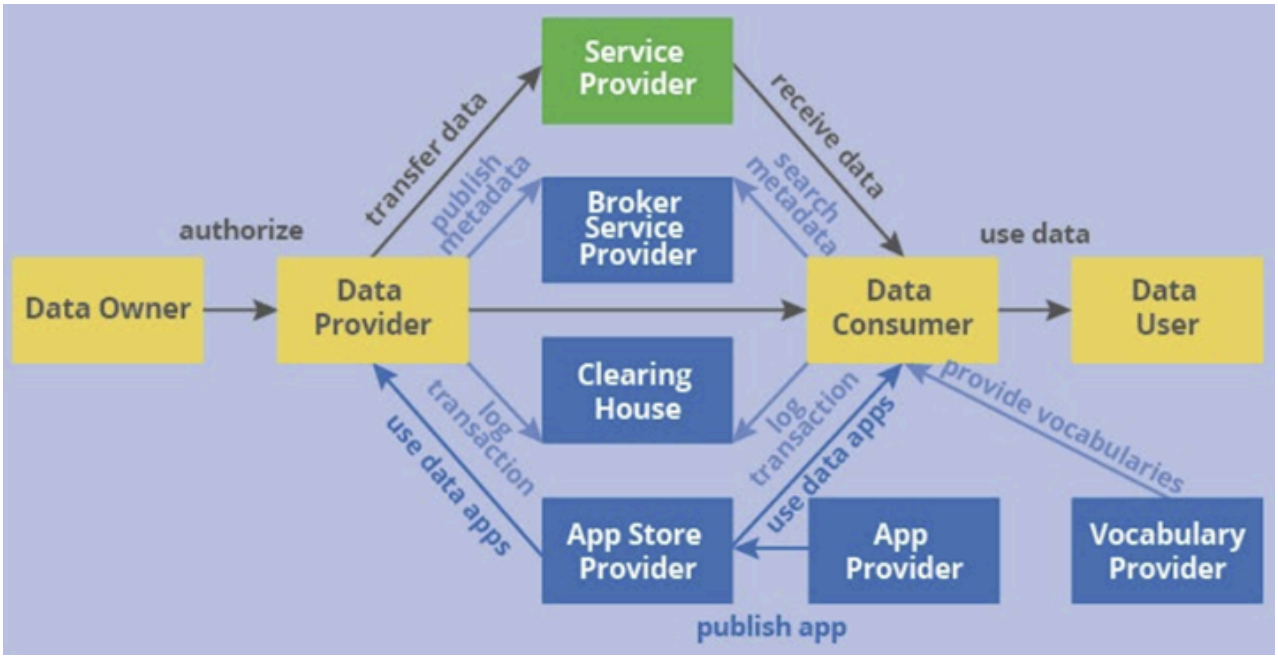


Figure 5. Data sharing in data spaces, as part of Open DEI "design principle 4" for data spaces

The relevance of data quality regarding data sharing resides in the fact that data consumers must be informed about the quality of the data products they are acquiring. Hence, to boost the acceptance of data sharing as a practice in the industry, it is vital to bring data quality into the equation, as it will increase trust on the transactions among pairs or on the different platforms.

18 https://design-principles-for-data-spaces.org/

However, this is not trivial. Some of the main roadblocks are related to how to inform about the quality of a dataset in a concise manner. Quality depends, as mentioned above and addressed in the next sections, on the purpose or the potential use of the data. This use may imply that what is relevant for one user is not for another, complicating this valuation of the data. Under this situation, the most feasible option is to provide quality results of different dimensions, informing them as well about the metrics or methods used to compute. Similarly, it is important to consider the role of interoperability in this process through the adoption of standards (e.g., Data Quality Vocabulary, see section 2.2.1) to express such information.

In addition to the data quality dimensions evaluated, it is also important to consider the various scopes that might exist in terms of the assessment of the quality of datasets and data-streams. In this sense, it is possible to identify two main scopes. On the one hand, the assessment of the dataset or data-stream (i.e. considering the features and characteristics of all the data points contained in it). On the other hand, the evaluation of each of the individual data points. In view of these different scopes, not only diverse metrics have to be put forward for each scope, but also different assessment mechanisms have to be in place in each case. Finally, the annotation of data quality assessment results and the corresponding modelling of such results' annotation, must be also adapted to each scope.

Taking into account all the above, it is clear that the inclusion of data sharing in the value chain imposes additional challenges to the quality of the shared data, how it is described and communicated to intended users, assessed and guaranteed and how a proper approach to it can be a source of trust between data provider and data user. In this context, data spaces emerge as unique environments where the right conditions to describe, assess and guarantee data quality in data sharing processes can be met. This is further explained in Section 4.

II.5 Data quality: fit-for-purpose

The introduction to this document, as well as some of the definitions of data quality presented in section 2.1 and the new dimension that data sharing introduces in the picture, already anticipate that data quality should also focus on how data are "suited for their intended purpose". As it is shown in Figure 1, more and more specific use cases and applications, not known at the time the data provider publishes its data and that eventually will make use of them, propose, on top of the generic properties of the data defined by the data provider to measure its quality, additional characteristics and metrics aligned with the specific purpose of the use case or applications. For example, in [5], data is considered of high quality only if they are fit for their intended uses, operations, decision making and planning. Since this concept could be quite subjective, user's expectations from data can differ, even when referring to the same dataset and even for similar uses.

- For example, in ***research***, the main objectives are to advance knowledge and state of the art, while applying the scientific method and allowing reproducibility. Therefore, the quality of the data will depend on the specific studies and mismatches will be identified during the peer review.
- In the ***public*** sector, main objectives are to provide citizens with common goods and new services (based on primary and secondary use of their data), ensuring transparency and trustworthiness as well as accountability and confidentiality.
- If we are dealing with ***sensitive data***, the purpose must be declared every time we want to access those data. The quality assessment should include an ethics review to ensure that the original reason for the data collection is compatible with the proposed use and the beyond this compatibility, data quality is enough for this intended purpose. Additionally, in the case that privacy enhancing technologies are used to protect the sensitive data, a new realm of data quality metrics pops up: the information leakage risk (privacy). Finally, for ***synthetic data***, the utility (closeness to the original) is also a relevant measure.
- Finally, in the case of ***business and industry***, a first use case usually motivates the initial purpose, but later use cases can be considered with proper data sovereignty mechanism in place. In this case, the main objective is to reach a competitive advantage, that sometimes can sacrifice unbiases or data transparency for the sake of the final end (which might also bring potential disinformation or collusion).

This implies that there is a parallel way of assessing data, from the perspective of where it came from and from the perspective of where it is used. In this sense, data becomes an asset that has a value which is reflected by (i) its provenance and (ii) its purpose. Taking into account the above and to address this dichotomy, data quality would differentiate between:

- *"General-purpose"* characteristics and properties (and the corresponding metrics) defined by the data provider, incorporating basically technical quality indicators. In general, those properties are the ones presented in section II.1.
- *"Fit for purpose"* properties that serve the specific goal of use cases and applications eventually using those data and metrics defined and assessed by the data users.

A paradigmatic example of this fit-for-purpose angle appears when considering that most of the applications that will use data will be AI driven. Although the concept of AI is quite broad and might include many different applications, tools, techniques, etc …, a high-level approach already helps to identify quality requirements that data should hold when used by AI driven applications. This is further explored in section III.
Finally, another example is the quality of data required in the health domain. In this sense, the recently agreed European Health Data Space regulation identifies specific quality conditions to share this type of data.

II.5.1 Data quality in the European Health Data Space

On March 15th 2023, the European Parliament and the Council of the EU reached a political agreement on the European Health Data Space (EHDS) regulation[19]. Main objectives of this regulation are:

- To put citizens at the centre of healthcare, giving them full control over their data to obtain better healthcare across the EU.
- To open data for research and public health uses.

Data quality is considered a key aspect of the regulation and the section V (Health data quality and secondary use) and more specific the article 56 is devoted to it. The data quality and utility label shall comply with:

- ***Data documentation***: meta-data, support documentation, data model, data dictionary, standards used, provenance
- ***Technical quality***, showing the completeness, uniqueness, accuracy, validity, timeliness and consistency of the data
- Data quality ***management processes***: level of maturity of the data quality management processes, including review and audit processes, biases examination
- ***Coverage***: representation of multi-disciplinary electronic health data, representativity of population sampled, average timeframe in which a natural person appears in a dataset
- Information on ***access and provision***: time between the collection of the electronic health data and their addition to the dataset, time to provide electronic health data following electronic health data access application approval
- Information on ***data enrichments***: merging and adding data to an existing dataset, including links with other datasets

Additionally, Annex II and III of the regulation collect essential requirements (general, security, interoperability) and technical documentation (including results and critical analyses of all verifications and validation tests undertaken to demonstrate conformity of the EHR system with the previous list), most of them very much aligned with some of the aspects addressed in this document.

19 https://ec.europa.eu/commission/presscorner/detail/en/ip_24_1346

CHAPTER III

DATA QUALITY IN ML AND AI

This section addresses the need to adapt the approach of data quality to AI and machine learning requirements. Although this is nowadays a hot topic under development and will be addressed with more detail in a next version of this document, this section already introduces key aspects that should be considered: data quality metrics and description for ML / AI, the relevance of data documentation, the relevance to keep the human in the loop, the symbiotic relation between data quality and generative AI and how data quality is reflected in the AI Act.

III.1 Data description for Machine Learning

Information about origin, provenance, representativity, data biases, trustworthy and equity (fairness)is becoming more and more important not only to ensure the quality of ML models trained with those data, but also added value services built on top of them. In that sense, ML communities are working to design their own tools to describe machine learning data sets (e.g. DescribeML - https://modeling-languages.com/describeml-machine-learning-datasets/, to allow data creators to describe the various aspects of the data used to train ML models in a structured manner) and collaborating to produce a "standard" high-level format of ML datasets (https://github.com/mlcommons/croissant).

Considering the above, several overall ML oriented metrics are crucial to understanding the data distribution of datasets used by ML models. Based on [9], some of the most relevant metrics are:

- ***Uniqueness*** describing the overall number of unique records in a dataset.
- ***Class parity*** reports the imbalance between classes for each feature of the given dataset. This can be measured using various metrics such as for example imbalance ratio [10] or normalised class entropy [11,12].
- ***Regularity*** is applied to timeseries data and measures the difference between timestamps of each entry in the data.
- ***Time Completeness*** is used in time series data gathering data at regular intervals. This metric gives an overview of whether the data is complete in the given time-window simply by comparing the number of actual entries with the number of expected entries.
- ***Feature Correlation*** measures how well individual data features correlate with each other. This method aids feature reduction techniques since highly correlated features can be removed from the data.
- ***Collinearity*** measures the linear alignment of variables. Variables that align well linearly can be used to predict each other.
- ***Class Overlap*** measures the overlap of features between different classes. This measure is important to understanding how strong the decision boundary is between different classes.
- ***Un-likeability*** measures the variability of values within the same feature in categorical data.
- ***Number of Outliers*** gives an insight into how noisy the given dataset is.
- ***Artificiality*** is a measure that keeps track as to the proportion of data that has been added to the dataset artificially, for example in order to fix missing values.

The assortment of various data quality metrics in the context of ML gives an important overview of the dataset, helping to build an understanding of the overall dataset and its behaviour when applied to a particular ML model. Additionally, visualisation approaches could be used in understanding the performance of the above metrics. These could include approaches such as heatmap visualisation to view the correlation matrix between individual variables of data or scatter plot displaying potential outliers of the dataset.

III.2 Data documentation

In the context of enhancing data quality for AI applications, effective data documentation plays a crucial role in ensuring that datasets are accurately used and understood throughout their lifecycle[20].
Metadata Annotation, Data Dictionaries, Data Provenance Tracking, Data Quality Frameworks and Automated Data Documentation Tools are primary methodologies for data documentation. Each method contributes to a comprehensive understanding and efficient management of data quality, but they also come with their challenges:

- Metadata Annotation and Data Dictionaries are foundational, enhancing data usability and consistency. However, they can be labor-intensive and may not fully capture complex dataset nuances.
- Data Provenance Tracking offers a detailed audit trail, critical for data integrity and error tracing, yet implementing such systems can be resource intensive.
- Data Quality Frameworks ensure adherence to quality standards but require continuous effort to maintain relevance in dynamic data environments.
- Automated Tools can streamline documentation processes but might not capture all data complexities without human intervention and can entail significant investment.

It is worth noting that regulatory frameworks such as the AI Act underscore the importance of rigorous data documentation for ethical AI practices. Documentation ensures data transparency, accuracy, bias mitigation, privacy adherence and auditability. These factors are crucial not only for compliance but for building AI systems that are reliable, fair and trustworthy.

Choosing the right documentation methodology—or a combination thereof—depends on the specific needs of the data environment and the objectives of the AI application. Effective documentation is a balance between thoroughness and practicality, aiming to enhance data quality while considering the resource and time constraints. As data environments evolve, so must the approaches to documentation, ensuring that data quality remains a central focus in the development and deployment of AI systems.

20 Datasheets for Datasets (https://arxiv.org/abs/1803.09010)

III.3 Increasing quality through "human in the loop"

How to guarantee the data quality in the applications where AI is involved, especially the ones related to knowledge, is not always an easy task. Fully automated systems often demonstrate to fail in guaranteeing results quality[21]. Furthermore, perceived quality influences also the trustworthiness in the AI: if the user perceived that an AI system could make mistakes, it is not comfortable for him to trust in the provided output. The level of trust is based also on the domain of the proposed AI application: humans are more prone to accept AI supports in cases where wrong results are not perceived as serious consequence (like support on internet research, games, etc.), while in other domain (e.g., health) the final user needs to be sure to have the right information.

From a technology point of view, we are assisting to an increasing usage of generative AI, thanks to the development of Large Language Models (LLM), AI contents creation increased even more in the last months. The involvement of human in the process where the output of an AI system is not automatically processed further but checked by a human (meaningful human control) has been proved extraordinary effective to increase the value of AI by supporting knowledge extraction. This human-in-the-loop system (HITL) scenario can be further extended to guarantee the data quality and its trustworthiness, where humans are called to collaborate with AI, for example in the case of Reinforcement Learning with Human Feedback (RLHF): LLMs are asked to provide 3 outputs (completions) to prompts for many thousands of examples, which are then ranked by humans according to pre-defined quality standards such as bias, accuracy, etc. The human output is used as a reward model by reinforcement learning agents that fine tune the original foundation LLM.

Moreover, in active learning humans annotate samples chosen by the automated systems (the technology selects the most useful data items to be delivered to the annotator, especially if there are few or missing information in some context).

21 https://www.forbes.com/sites/mzhang/2015/07/01/google-photos-tags-two-african-americans-as-gorillas-through-facial-recognition-software/

The annotation activity increases the quality of the results when more annotators are involved: the agreement among different annotators ensures a higher quality. The HTIL approach can be used also in the validation phase, for fine tuning on the AI models results. In this case, the main aspects to be considered are errors, bias identification, further model refinement and ethical concerns. Finally, considering the interaction with LLM, users are called to support and elicit the AI with prompt engineering. Prompts can be used to explore the data accuracy (e.g., data can be controlled by specific programs generated with the support of code assistance prompts), to produce consistent output (e.g. a prompt who specifies the output format for some requested data), or to work on the timeliness of the information (including specific time boundaries for the request in the prompt).

HITL allows humans to go back in the loop and to take control of the AI system, verifying the content to be suggested (as in the prompt creation) and validating the output, with the consequent improvement in the trustworthiness in AI.

III.4 Generative AI for Data Quality

Generative Artificial Intelligence is artificial intelligence capable of generating text, images or other data using generative models. These models learn the patterns and structure of their input training data and generate new data that has similar characteristics.
Nowadays, this type of AI is starting to be used to enhance the quality of our data. Many researchers believe that Generative AI will be used as a very powerful tool to improve the quality of data [13] [14].

Generative AI can be a powerful tool for improving data quality across multiple dimensions:

- Data imputation and completion, completing missing data or forecasting future data points.
- Data validation and cleansing, by anomalies detection and procuring data compliant with specific standards
- Data augmentation and feature engineering
- Data simulation and testing, by using Generative AI to create synthetic data and using them to simulate potential scenarios.
- Data governance and compliance.

In today's tech-driven business landscape, data quality is paramount as organisations leverage data analytics for actionable insights. Findings reveal the multifaceted impact of Generative AI on data quality, addressing key parameters despite the absence of specific built-in frameworks. Ethically conducted, the study envisions Generative AI simplifying data processing rules and recommending tools, particularly valuable for handling large datasets. An exciting future trend involves integrating Large Language Models (LLMs) into data quality tools, promising intuitive and efficient solutions. In summary, while Generative AI does not replace existing data quality methods, its potential augmentation is promising, offering strategic benefits for organisations and shaping the future of data quality through AI advancements.

III.5 Data quality in AI Act

Special data quality techniques should be developed to comply with current EU regulations, in particular with the recently approved AI Act[22]. As stated in the regulatory text itself[23], high data quality is essential for the performance of many AI systems, especially when techniques involving the training of models are used, aiming to guarantee that high-risk AI systems perform as intended and they do not become a source of bias or discrimination.

This fundamentally refers to the definition of appropriate data governance and management practices covering training, validation and testing of data sets related to persons or groups of persons, especially in high-risk AI systems. According to the AI Act, training, validation and testing datasets should:

- be ***relevant, representative, free of errors and complete***
- have the ***appropriate statistical properties*** (met at the level of individual data sets or a combination thereof)
- consider the characteristics or elements particular to the ***specific geographical, behavioral or functional*** setting within which the high-risk AI system is intended to be used
- appropriate ***safeguards for the fundamental rights*** and freedoms of natural persons

Besides, to ensure data accuracy and completeness, there is a need to implement appropriate risk management measures, so that possible shortcomings are duly addressed. Noticeably, these requirements do not preclude the use of privacy-preserving techniques in the context of the development and testing of AI systems.

High-risk AI systems suggest the interest of defining domain-specific quality rules according to the nature of the existing features in the data sets, in the sense that different categories (geographical, behavioral, biometric, etc.) can have specific quality indicators, constraints and therefore treatments.

22 EU AI Act: first regulation on artificial intelligence
https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence
23 EU AI Act, Recital 44: Implementing Data Governance and Management Practices (https://www.euaiact.com/recital/44)

CHAPTER IV

DATA QUALITY IN DATA SPACES

Data spaces are defined in the European Data Strategy[24] as "spaces aimed at overcoming legal and technical barriers to data sharing across organisations, by combining the necessary tools and infrastructures and addressing issues of trust, for example by way of common rules developed for the space. The spaces will include: (i) the deployment of data-sharing tools and platforms; (ii) the creation of data governance frameworks; (iii) improving the availability, quality and interoperability of data – both in domain-specific settings and across sectors". Therefore, one of the objectives of a data space is to improve the quality of the data it offers for sharing within specific domains (fit for purpose) or across sectors (general purpose).

But beyond this objective, we argue that data spaces provide the means, tools and processes to offer unique environments where to assess the quality of the data, described it properly, ensure that the quality is kept during operations and the whole process is governed properly.

Taking this into account, this section covers different aspects of data spaces addressing data quality and concludes with a set of requirements that the data space should consider to ensure the right level of quality according to different conditions.

IV.1 Data mesh, data product and quality

Data mesh concept was initially coined in 2019 [15] and implies a shift from traditional data lake architectures to a new paradigm that consider domains as the first-class concern, apply platform thinking to create self-serve data infrastructure and treat data as a product. Data mesh is conceived as a decentralisation paradigm, regarding the ownership of data, the transformation of data into information and data serving. It aims to increase the value extraction from data by removing bottlenecks in the data value stream.

24 https://digital-strategy.ec.europa.eu/en/policies/strategy-data

The data mesh paradigm is guided by the following three principles, helping to make data operations efficient at scale:

- Data and distributed domain driven architecture convergence
- Data and product thinking convergence
- Data and self-serve platform design convergence

Distribution of data ownership and data pipeline implementation into business domains raise an important concern around accessibility, usability and harmonisation of distributed datasets. These concerns can be addressed by the concept of "data product", that implies a change of mindset of providers to consider their datasets as products with an implicit structure. To provide the best possible service for their unknown customers, data products need to have the following qualities:

- ***Discoverability,*** through a catalogue of all available data products with their metadata information
- ***Addressability,*** with a unique address following a global convention that helps users to access it.
- ***Trustworthiness and truthfulness,*** applying data cleansing and automated data integrity testing when creating the data product and including information about data provenance and lineage.
- ***Self-description*** semantics and syntax
- ***Inter-operability*** and governance by global ***standards***
- ***Security*** and global ***access control***

Data spaces have different origins but also much in common with data mesh. Data mesh focuses on data management within organisations, while data spaces focus on data management across organisations, enabling data sharing and data reuse among them. Data spaces build on concepts such as data space connectors, data governance capabilities, identity and access management and data catalogues[25].

25 https://dssc.eu/space/News/blog/108199969/Data+Products+in+Mesh+and+Space

IV.2 Data quality & data governance

"Data quality can be started without formalised data governance, but it cannot be sustained without data governance [15]"

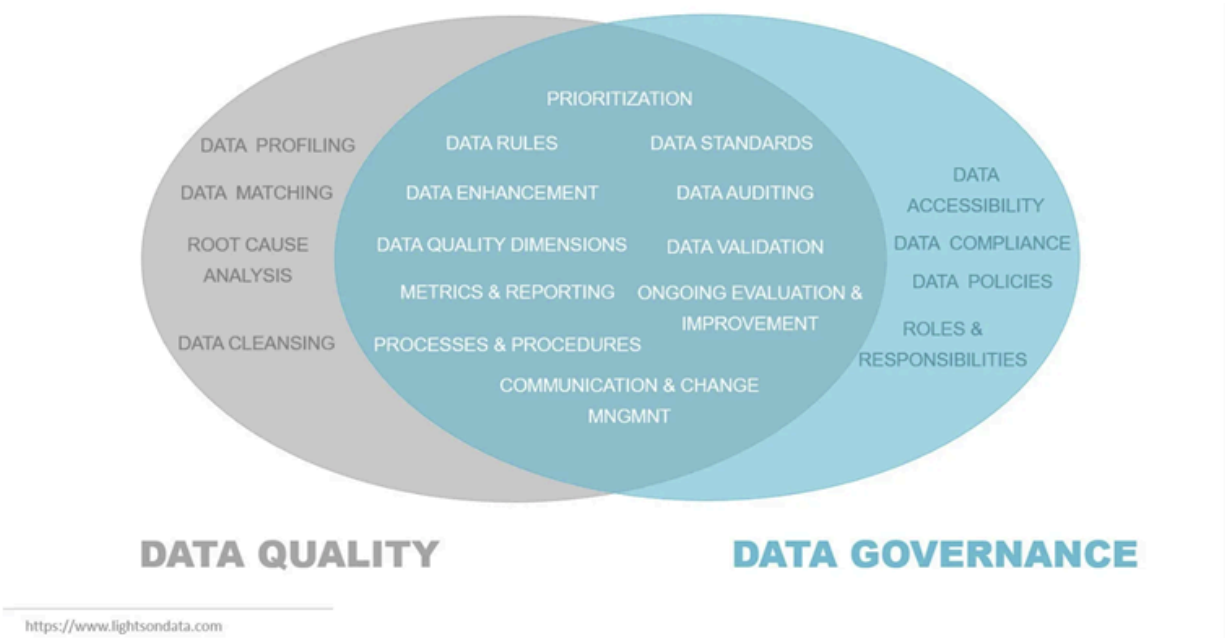


Figure 6. Symbiotic relationship between data quality and data governance**

As it is highlighted in the AI Act, data quality is quite tied to data governance, since a proper governance ensures:

- relevant design choices
- data collection
- preparation processing operations, such as annotation, labelling, cleaning, enrichment and aggregation
- formulation of relevant assumptions
- prior assessment of the availability, quantity and suitability of the data sets
- examination in view of possible biases
- identification and addressing of any possible data gaps or shortcomings

** https://www.lightsondata.com/data-quality-and-data-governance/

Data governance affects the quality of the data used to train, validate and test models. But sometimes it might also be difficult to differentiate differences between those two terms, since they coexist in symbiotic relationship[26] with some overlap (see Figure 6). Basically, data governance describes roles, types of data, procedures, permissions, conditions and tools, so it has a lot of impact on data quality, but also in other aspects. And similarly, there are aspects of data quality that might fall per se under the data governance.

26 https://www.lightsondata.com/data-quality-and-data-governance/

IV.3 Data quality in DSSC building blocks

The importance of data quality implies that it does not affect just a specific building block of a data space, but it should be seen as a key transversal aspect to be considered by design in most of the building blocks composing a data space.
By doing so, data spaces appear as a unique and controlled environments that guarantee the quality of data shared in and provided by, the data space, based on the data governance rules, the imposed trust and sovereignty, available technical tools to support this data quality, etc. …

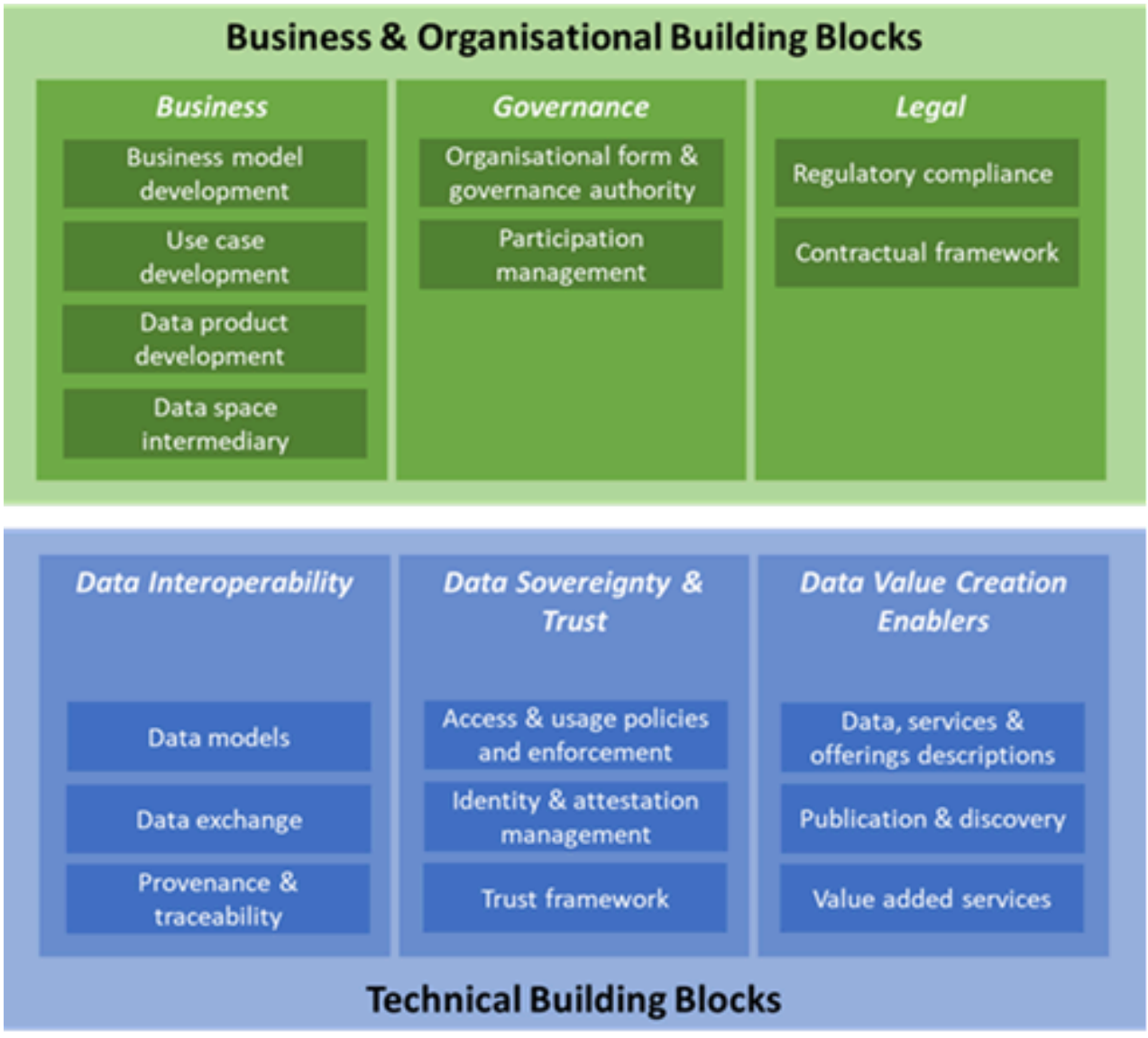


Figure 7. Data space building blocks, as presented in Data Spaces Support Centre (DSSC) blueprint

This feature appears also as a great advantage for external applications that rely on the data space to get data, since the data space can guarantee that provided data meets the quality requirements imposed by the application. In this case, data spaces can play an important role in supporting implementation of AI Act, regarding the required data quality to be compliant with the regulation.

This section takes as a reference the building blocks of a data space that the Data Spaces Support Centre project has identified in its version v1.0 of the blueprint[27] (Figure 7) and based on them, Table 3 provides this two-vision: (i) how data quality should be considered in the different building blocks of a data space and (ii) how this consideration reinforces the role of a data space as a unique environment to guarantee the quality of data.

27 https://dssc.eu/page/knowledge-base

Table 3. Data quality in data spaces: requirements and value

Building block	Quality in Data Spaces	Data Spaces for quality (value)
Business Model Development	• Definition of specific dimensions and metrics for data quality based on the scope of the business model • Specific dimensions and metrics considering the sector(s) where the data space will operate.	• Guarantee that the data available in the data space fulfills the required quality criteria from the business perspective.
Use cases development	• Definition of specific dimensions and metrics for data quality based on the specific use case that will use those data. • Definition of specific indicators for each metric, based on the expected performance of the use case. • Monitoring, assessment and validation of the quality of new data produced in the use cases.	• Guarantee that the data used in the use case matches its specific needs. • Guarantee the global quality of data in the use case (data used, data generated) meets required quality criteria.
Data Product Development	• Data quality to be included as part of the data product (resources in the data product, but how??).	• Guarantee that every data product shared in the data space includes relevant information about the quality of the datasets therein (descriptions, evidences of data source, provenance and lineage, verified and signed attestations of quality).
Organisational form and governance authority	• Definitions of global data quality requirements at data space level. Common dimensions and metrics and required minimum (or different levels of) indicators, list / classifications of data based on their quality, etc. • Provides the data governance rules to sustain data quality in the data space.	• Ensure that data in the data space fulfills general criteria for quality at data space level.
Participation management	• All participants in the data space share same criteria regarding data quality, are aware of required levels of quality.	• The whole ecosystem shares a similar approach towards data quality in the data space. Same dimensions, metrics and levels of quality. Same descriptions and how they are reflected in the data product.
Regulatory compliance	• Addressing data quality w.r.t EU general regulations (AI Act, DGA, …), specific to each sector (e.g. data quality in health data space EU regulation, etc .), or national or regional regulations.	• Ensure that data quality in the data space fulfills EC regulation requirements, at different levels and for the specific domain of the data space.

Building block	Quality in Data Spaces	Data Spaces for quality (value)
Contractual FW	• Data quality included in the contracts between providers and users. The contract should specify the quality of data to be shared: requirements from the user and description of data from the provider, taking into account all the aspects described in this table.	• Trust for users that they can rely on the contract to guarantee the quality of the shared data.
Provenance and traceability	• Information of data quality through the value chain, how different indicators have changed, if the data have been transformed by any means (cleaning, aggregation, etc …). • Information of source and lineage of data. • Data quality as a dynamic feature. What new dimensions and metrics have been incorporated.	• Knowledge about source, lineage and provenance of data. • Knowledge about changes in quality along the value chain.
Trust	• Verify claims about data quality.	• Source of trust for metrics and indicators included in the description of the quality of data.
Descriptions	• Data quality included in the description of datasets. • Exhaustive description for the different dimensions, metrics and indicators, lineage of data, source and provenance, etc … and other aspects.	• Potential data users find exhaustive, reliable and standardised information about the quality of data they are planning to use.
Value-added services	• Data quality to be defined and evaluated depending on what services and applications will use the data. • Materialise the technical needs from business and governance regarding data quality: • Services to assess data quality • Services to validate data quality	• The whole ecosystem shares a similar approach towards data quality in the data space. Same dimensions, metrics and levels of quality. Same descriptions and how they are reflected in the data product.
Data Space level	• Quality considered in (most of) the building blocks of the data space (see previous rows).	• Data space provides a controlled environment for data and data quality • Governance rules to ensure the quality of data in the data space • Knowledge about data source, lineage and provenance • Ethical and trustworthy environment • Exhaustive, standardised and reliable descriptions of data • Services to assess and validate the quality of data • Access to proprietary data (not accessible under other circumstances or other environments)

CHAPTER V

DATA QUALITY IN DIFFERENT DOMAINS AND PROJECTS

V.1 Public sector / public administration

In the public administration and public sector, main objectives of adopting and implementing data spaces are:

- Make sure that data collected by the public sector is available for all data space participants, in particular data that falls under the data categories listed within the Open Data Directive[28] (including High Value Datasets Implementing Act) and Data Governance Act (Section I). The Open Data Directive promotes fair access to information and increasing the number of available public sector datasets and at the same time motivates the public and private sector to reuse data. A High Value Dataset is a dataset holding the potential to (i) generate significant socio-economic or environmental benefits and innovate services; (ii) benefit a high number of users and SMEs; (iii) assist in generate revenues; and (iv) be combined with other datasets.
- Support private sector institutions in the reuse of public sector data for the delivery of new or improved services to citizens, based mostly on the secondary use of their data (i.e. G2B data exchange).
- Use data available from the private sector in order to improve public services (i.e. B2G data exchange, for example the use of anonymised mobile positioning data for public transportation planning purposes).
- Contribute to ensure transparency and trustworthiness for citizens from public bodies and governments

When it comes to implementing the concept of data spaces in the public sector, several challenges need to be addressed:

A change of mindset in the public sector to embrace the culture to use private sector data in the provision of public services and to provide public sector data as data products, defined according to user needs. Such a mindset change would lead to the active participation of public administrations and their service providers in the elaboration of data spaces.
Based on the previous point, public sector institutions need to monitor external user needs and work towards the required level of interoperability and quality of data to allow greater exploitation through external services.

28 https://eur-lex.europa.eu/eli/dir/2019/1024/oj

Development and generation of mechanisms to ensure the efficient implementation of governance procedures, including potential certification of applications and services before being included in public catalogues.

Automatically ensure verification and control processes for law enforcement (e.g. GDPR).

In terms of data quality, the main challenge is to shift from the current, mostly quantitative, approach to a more qualitative one in the publication of public sector data. Namely, the current best practice is to answer the question "what data does the public sector have" - public sector data catalogues and open data portals mostly allow you to find out what kind of data is being collected and stored.

For public sector data to be truly useful in a data space context, users also need an answer to "is this data usable for me". This means, that the current simple descriptions of datasets must be amended with further information about the context of data gathering and processing, the data quality regime applied (or the lack thereof) and ultimately a useful statement about the quality of the data. This simple challenge has deeper technical consequences for the registries (data to be shared). One of the initial consequences relies on the shift from data quality assessed from datasets (in other domains such industry) to the specific data point. Similar needs exist in terms of data accessibility, sharing and traceability. The main rules and changes must be adopted for each data point, incrementing the complexity of data quality, security and safety approaches.

In terms of approaches, most of the authors are focused on the application of Open Digital Rights Language (ODRL[29]) and Shapes Constraint Language (SHACL[30]) rules inside specific knowledge graphs to maintain traceability and data reliability for the public administration registries. However, this approach requires the combination of other techniques that permit us to measure and assess the quality of the atomic information during the storage and their exposition. In this regard, there is a lot of work to be done inside the public administration.

29 https://www.w3.org/TR/odrl-model/
30 https://www.w3.org/TR/shacl/

V.2 Industry

There is no question that data is a basic resource for organisations. Appropriately handled, they generate great competitive advantages, both in decision-making and in the generation of new products and services, empowering advancements such as Artificial Intelligence. This present circumstance has made numerous organisations careful about sharing their information. Be that as it may, the circumstance is evolving and more and more companies and organisations are becoming aware of the upside of this practice.

Data sharing drives efficiency in supply chains, empowering quicker and more imaginative product development. By sharing their data, organisations also benefit from access to third-party data, which can be of extraordinary use in various fields: from training machine learning systems to enriching internal analytics. Furthermore, the fact that several organisations are working in similar fields, creating progresses, implies that the market matures earlier, opening up new business opportunities, as well as decreasing the time and expenses of marketing products. There are likewise benefits in terms of transparency and reputation.

Secure and controlled environments, such as data spaces, are necessary for this data exchange to take place in a safe and secure manner, being the final result a more flexible and resilient supply chain.

Local industrial data spaces have emerged in a semi-developed form in recent years. The businesses that make up these groups are arguing about what information should be shared and how to regulate it all and they are also pressuring their software providers to adopt specific semantics and formats. It is now clear that these local areas must grow in order to attract a sufficient number of enterprises to provide a truly valuable data exchange and to draw in technology suppliers to create customised solutions.

V.3 DATAMITE

DATAMITE will build a modular stack that helps European enterprises and public administrations to improve the monetisation of their data. Understanding monetisation as the ability to use data to increase income, two types of monetisation must be considered, internal and external monetisation. Internal monetisation is the result of leveraging data to make better business decisions, to have a better understanding of customers or improve their own products, among others. External monetisation encompasses everything from selling data to sharing or exchanging it with potential partners looking for synergies, to collaborate in shared ventures or even for free, proposing challenges that may help the organisation. Data quality is key in all these situations.

Looking at the internal monetisation, it can be found that one of the main issues of European organisations is the lack of confidence in their data, resulting in the so-called data-decision-gap reported to affect 95% of companies. This is, basically, that companies do not rely on their own data to make decisions. Most organisations do not have any control or monitorisation of the quality of their data. As a result, data may be incomplete, imprecise, inaccurate, inconsistent, etc. Any analysis, AI model, or forecast built taking these data as input may be misleading or incorrect, resulting in wrong or even prejudicial decisions. Decision makers may err once, but not twice, so they will not rely on their data for a second time.

On the other hand, when looking at the external monetisation, data quality affects both the data owner and the consumer. Being aware of the quality of their data, the data owner will be capable of valuing her data more precisely and even raise its price if the data quality is superior to that of its competitors. Similarly, data consumers must know how good the data is they may be willing to purchase and whether the disbursement they will make is justified. Moreover, given the number of initiatives being undertaken to facilitate data sharing in the European context (e.g., Gaia-X, SIMPL, EU Data Spaces; mostly supported in terms of data exchange by IDSA reference architecture-based technology) it is key to find a common language to express data quality in these shared environments and, also, that it is not only feasible but easy to use by data owners publishing their data.

The goal of DATAMITE is to become the open-source framework that answers these issues. DATAMITE will be composed of five main modules: governance, quality, security, sharing and support tools.

The quality module will allow data owners not only to profile their data with an ample collection of informative indicators already provided in the framework, but also to define their own context-based quality indicators and use them on their data for an accurate and context aware quality monitoring. This information will be key for data owners willing to become data driven organisations, allowing them to know whether their data is or not reliable and improve internal monetisation. Moreover, DATAMITE will report the data quality results following industry standards, e.g., DQV, to facilitate, as well, external monetisation providing rich, detailed and standardised data quality information when publishing data into shared environments such as data spaces. Hence, in a nutshell, DATAMITE will aid data owners to manage their data from the moment it is ingested to the moment it is shared with other parties. It will do so by providing tools to assist them in the enrichment of data with metadata, through a data catalogue and using data glossaries; offering data quality tools to profile and analyse their data with informative or purposedly defined context-aware indicators to ensure that their data is good data or to allow them undertaking corrective actions otherwise; or facilitating how to share their data into a collection of initiatives (e.g., EOSC[31], AIoD[32], DataSpaces, Gaia-X[33]), defining their own terms of use by using data sovereignty tools and using a collection of plugins or connectors such as the IDS-based EDC connector[34].

31 https://eosc-portal.eu/
32 https://www.ai4europe.eu/
33 https://gaia-x.eu/
34 https://github.com/eclipse-edc/Connector

V.4 PISTIS

PISTIS stands for Promoting and Incentivising Federated, Trusted and Fair Sharing and Trading´of Interoperable Data Assets. It provides a comprehensive framework and a reference platform made up of connected deployments to unlock the full potential of the data economy. It aims to transform how data is handled, shared and monetised. By incorporating top-notch methods and solutions from various fields such as data management, analytics, finance, crypto and security, PISTIS facilitates efficient and reliable sharing and monetisation of data assets.

Data quality monitoring and assurance are important topics addressed in the project. It provides several data quality related services, which have a high potential to be used in other project and initiatives.

One of them is the Metadata Quality Assessment (MQA) service, originally developed for the European Data Portal (EDP), that is being utilised to measure and showcase the quality of the metadata in the Data Spaces. MQA is being enhanced in PISTIS so that it can evaluate not only the metadata that describes the data but also to enable the users to improve it. Unlike the EDP's rather static metric system, the MQA service within PISTIS introduces a dynamic and configurable set of metrics to fit to the need of the Data Spaces. Moreover, the MQA leverages SHACL (Shapes Constraint Language) to validate metadata against the predefined PISTIS metadata model, to indicate errors and inconsistencies that could undermine data interoperability and usability. The outcomes of the MQA's assessments are stored using the Data Quality Vocabulary (DQV).

The PISTIS metadata model is an DCAT extension inspired by Gaia-X. SHACL (Shapes Constraint Language) is used as modelling and validation language. The metadata model is the basis for the PISTIS Data Catalogue (and Marketplace) implemented with help of the Open Source and Java-based data management solution "piveau". Piveau is designed around Semantic Web technologies and uses Triplestore as its primary database. That allows to store metadata, data and data models in native RDF (Resource Description Format). The project aims to present the catalogue as an alternative implementation of the GAIA-X Federated Services Catalogue.

Additionally, PISTIS provides a service for enforcing dataset quality by mapping the tabular data to the domain-specific data models. The PISTIS data models are defined with the help of existing domain models, CSVW (CSV on the Web)[35] and ODRL (Open Digital Rights Language). The information about the data structure and types in the dataset is then recorded in the metadata stored in the Data Catalogue and used for more efficient management, discovery and usage of data.

35 https://csvw.org/standards.html

V.5 MobiSpaces

MOBISPACES provides an end-to-end mobility-aware and mobility-optimised data governance platform. Its primary point of differentiation is that mobility analytics results will be leveraged to optimise the entire data path, resulting in data processing that is efficient, dependable, secure, equitable and trustworthy. MobiSpaces offers intelligent mobility services, enforces privacy restrictions at the intended point of action and promises the decentralised extraction of actionable insights from ubiquitous mobile sensor data and Internet of Things devices. To achieve its goals MobiSpaces leverages on two main assets: the AI-based Data Operations Toolbox and the Edge Analytics Suite. Additionally, it integrates a Green & Environmental Dimensioning Workbench to ensure eco-friendly processing practices.

The AI-based Operations Toolbox includes tools for:

- **Declarative querying**, separating the definition of query execution from the actual code for optimal implementations for heterogeneous NoSQL storage systems and clearly describing abstract data operations (filter, sort, join, etc.).
- **Decentralised data management**, employing a massively distributed and adaptable architecture to drastically lower the processing requirement on a centralised node, resulting in the creation of a more capable, reliable, secure and environmentally friendly ecosystem.
- **Online data aggregation**, where the edge nodes that gather raw data locally—as opposed to the more basic sensors that can only gather and send data—will function as miniature data centers, with each one in charge of gathering, processing, analysing and mining a specific subset of the system's available data.

The Edge Analytics Suite, instead, envisions:

- **XAI prediction modeling**, providing a tool with XAI methods for explaining deep neural network explainability. It will discuss methods for determining which characteristics AI systems find most interesting and instructive, with a focus on mobility patterns and the factors that lead to their prevalence.

- **Edge-driven federated learning**, by enabling XAI techniques, edge devices can utilise and analyse locally stored and proximity-based datasets to train lightweight versions of the global model that are fully explicable and reliable.
- **Visual analytics**, providing interactive and scalable methods that can effectively manage streaming and historical spatiotemporal data from many sources at varied quality and resolution levels. Advanced visualisations are powered by data transformations, which constitute the foundation of the Visual Analytics tool.

With the latter tool, data quality seems to have a greater influence on data transformation and its ability to facilitate the manipulation of mobility data in several formats to support different visualisations and help people (domain experts) carry out certain tasks. Furthermore, the correctness, dependability and credibility of the insights obtained from visualisations are guaranteed by data quality, which is essential to the success of visual analytics. More specifically, the quality of data has cascading effects on many other aspects such as:

- the accuracy of insights, since inadequate or erroneous data can result in deceptive visualisations, which can lead to erroneous judgments and choices,
- data cleansing and integration, as improperly structured or inconsistent data might impede integration.
- timeliness: while data quality procedures should contain safeguards to guarantee the freshness of the data used in visualisations, stale or outdated data might provide useless insights and impede the capacity to react swiftly to changing circumstances.
- improved interaction, since users with high-quality underlying data are better able to drill down into specifics, filter data and carry out other interactive tasks.

V.6 SEDIMARK

SEDIMARK is a HORIZON Europe project that aims to build a secure and trusted decentralised marketplace for allowing the sharing of high-quality data and services between companies, researchers and public authorities. Data quality is of the utmost importance for a data marketplace, as users might be willing to pay higher prices for datasets or services of high quality compared to others that might have missing values or dirty and duplicate data. Thus, SEDIMARK is working on building a complete data curation and quality assessment pipeline for both streaming data and offline datasets. The pipeline consists of various tools that can work either as standalone modules, or as components of a full pipeline for optimal results.

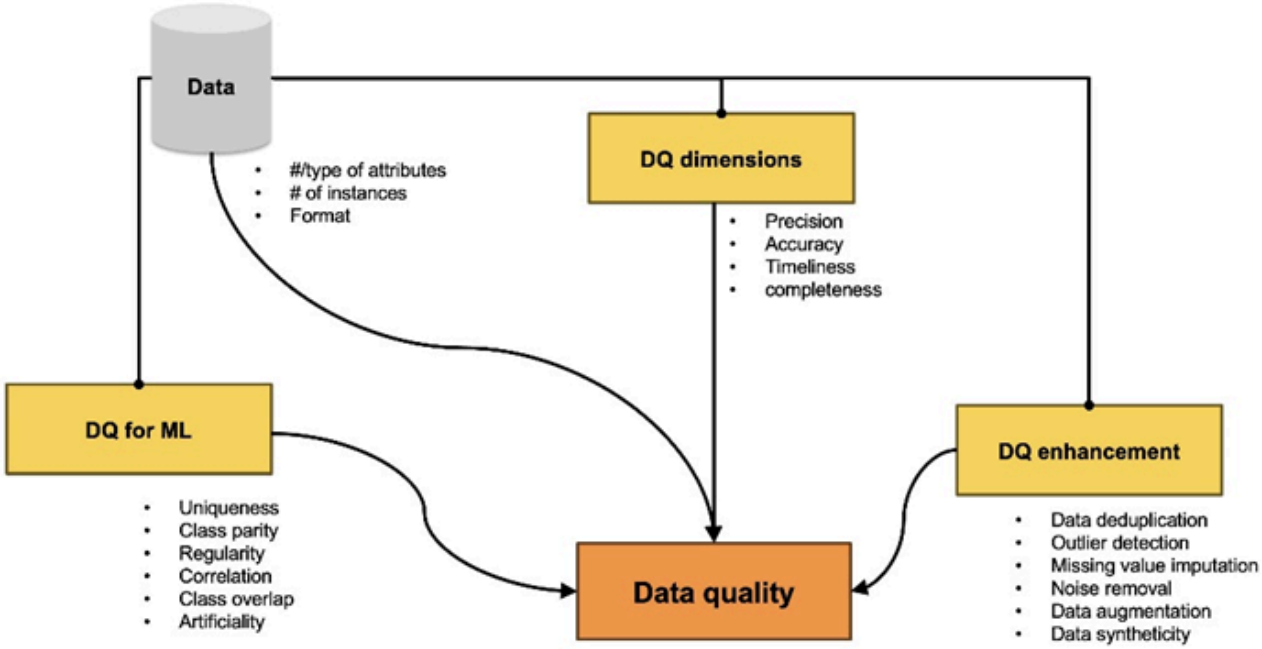


Figure 8. Data quality in SEDIMARK

As discussed in [9], the data curation and quality assessment pipeline of SEDIMARK consists of different modules, including (i) anomaly detection, (ii) deduplication, (iii) missing value imputation, (iv) augmentation and (v) profiling and quality assessment tools. SEDIMARK has built python components based on existing well-known libraries (i.e. PyOD, tods, etc.) that process both streaming and offline datasets in a simple but customisable way, so that the users of the tools can either configure the tool parameters themselves or select from predefined settings based on the task at hand.

Additionally, SEDIMARK introduces its own data models that cater for data quality, aiming to adequately describe the datasets to be shared through the marketplace. This will allow the users of the marketplace to know beforehand the quality statistics of the datasets and select those that best fit their goals. It has to be noted here that the processing of the datasets is up to the preferences of the data provider. This means that SEDIMARK allows the data providers to share both dirty and cleaned/high quality data. In this respect, the users will be able to find datasets of various quality that fit their goals. For example, if the user is a researcher working on anomaly detection, purchasing a "clean" dataset might be of little use to them, so they will be willing to purchase dirty datasets. However, SEDIMARK goes one step further and allows data providers to use the data curation and quality assessment pipeline, without removing the bad/dirty data, but with proper annotation. That way the resulting dataset will be a replica of the original, but with additional fields denoting which rows/data points are duplicates, anomalies, etc. Using the previous example of the researcher, they will be able to benefit both from the full dataset and from the SEDIMARK annotations to compare their models.

V.7 SALTED

The key objective of SALTED is to add value to existing datasets and data-streams by enriching them through the application of the principles of linked-data, semantics and Artificial Intelligence. These datasets and data-streams come from heterogenous data sources such as Internet of Things (IoT) deployments, Open Data portals and social media and are harmonised towards a standard information model, i.e. NGSI-LD, targeting the so-essential interoperability.

The Data Enrichment Toolchain (DET) architecture designed and developed comprises different microservices that, overall, address the challenges presented along the achievement of the main goal of the project. The principal microservices are:

- data discovery, i.e., the ability to discover and request the collection of sets and streams of data;
- data formatting, i.e., the transformation of raw data into well-formed and structured sets of data accordingly to data models described in terms of NGSI-LD;
- data curation, i.e., the identification (and potential correction) of data that do not reflect the expected quality (outliers, errors in values and the like);
- data linkage, i.e., the ability to relate different datasets accordingly to a well-established definition of relationships;
- data enrichment, i.e., the ability to understand and frame the data structures according to situations and contexts and the definition of functions that exploit this contextualisation.

Among the data curation and data enrichment services developed in SALTED modules for anomaly detection, missing value imputation and profiling and quality assessment, amongst others, were included. In this sense, the pipeline composed by these modules performs several functionalities for each of the to-be-curated entities so that the new information obtained is appended to the now curated entity by storing it in a new Data Quality entity linked with the original.

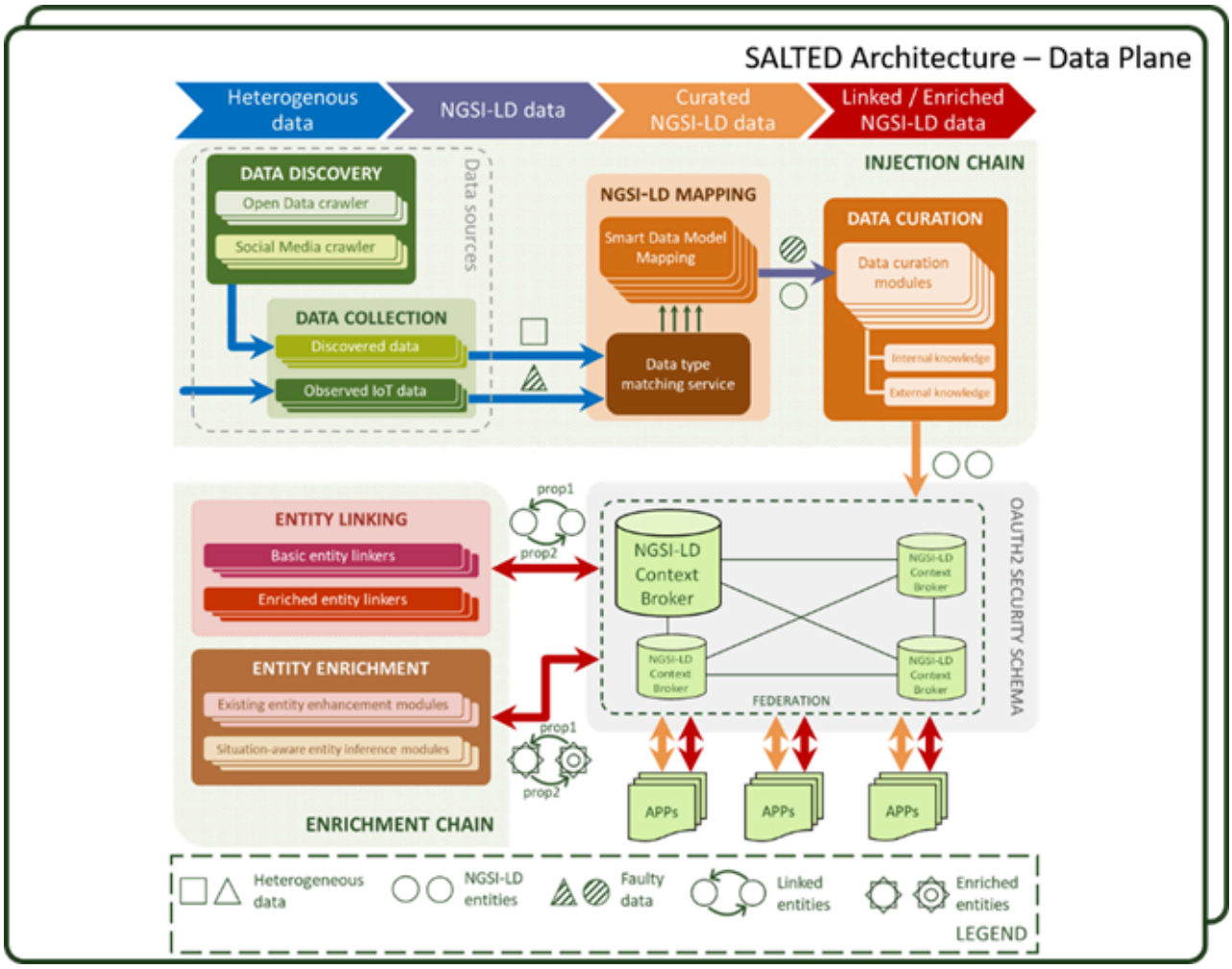


Figure 9. SALTED architecture

V.8 WATERVERSE

The WATERVERSE mission is to develop a Water Data Management Ecosystem (WDME) for making data management practices and resources in the water sector accessible, affordable, secure, fair and easy to use, improving usability of data and the interoperability of data-intensive processes, thus lowering the entry barrier to data spaces, enhancing the resilience of water utilities and boosting the perceived value of data and therefore the market opportunities behind it. In that context, several modules relevant to data spaces are being developed. This includes an anonymisation module, detecting anonymity level of a dataset and providing different strategies to anonymise a given dataset.

Another key contribution to the dataspaces relates to the development of FAIR data objects. The FAIR data principles mean an inflexion point to provide data greater value and enhance their propensity for reuse both humans and machines. It defines FAIR Digital Objects, which may represent data, software or other research resources. FAIR Digital Objects should be Findable, Accessible, Interoperable and Reusable (FAIR principles) to the greatest extent possible. Nevertheless, FAIR is not only covering the data, FAIR Digital Objects are also located in wider FAIR Ecosystem which is compound or services and infrastructures for FAIR. WATERVERSE defines three priority recommendations: (i) Define the concepts for FAIR Digital Objects and the ecosystem (ii) Implementation of the recommended FAIR principles (iii) Development of FAIR Metrics for FAIR Digital Objects and FAIR Ecosystem.

CHAPTER VI

MAIN FINDINGS, RECOMMENDATIONS & CONCLUSIONS

High quality of data collected, generated, processed, analysed and shared is crucial to guarantee that all those stages in the data lifecycle contribute to the ultimate objective ***of generating value from this data***. Following the paradigm "garbage in, garbage out", the final result and value from applications and services built on top of this data depends strongly on its quality. Hence, it can be stated that, in general, data quality drives data value and data value dimensions are either derived from or partially overlap with data quality. The complete value of data is unlocked at the intersection between its intrinsic properties (data quality included) and the requirements of the end-user.

Data quality should include general purpose properties and metrics defined by the data provider when publishing the data and not knowing yet its intended use and ***fit for purpose*** quality properties to address specifically the needs of applications and use cases. And data quality definition and properties depend very much on the domain and scope where the data is used. A paradigmatic example is the European Health Data Space regulation, that imposes specific requirements on health datasets and Electronic Health Records. But this paradigm becomes even more important when considering ***data for AI applications and*** their specific requirements in terms of data quality. The AI Act specifies how data quality should be addressed in the specific domain of AI, to ensure the proper training, validation and testing of AI models.

Data spaces represent an ideal environment to address data quality requirements and to ensure that data therein meets the quality criteria defined by stakeholders and use cases. In this way, data governance is critical to assess, validate, ensure and maintain data quality and data spaces also provide framework where to develop proper data governance. This ***symbiotic relationship between data spaces and data quality*** should be further exploited.

Finally, many ***European research projects*** are addressing data quality from different angles and developing specific blocks or components specially devoted to it, also considering specific applications including AI and coming from different domains. Some of these components and tools could be adopted by data spaces to support their quality assessment and validation.

Based on all the above, this document concludes with the following recommendations:

- Data quality is a key feature when addressing data value creation. We recommend a ***structured, methodological and standardised approach towards data quality***, focused on data value creation and considering data product as the way to bring together the datasets, quality dimensions and metrics and other aspects such as provenance, impacting the value generation from these data.

- Considering the fit-for-purpose approach and what is outlined in the AI Act about data quality, ***requirements for data quality requirements for AI need to be elaborated deeper***, specifically for new AI paradigms as Generative AI and Foundation models.

- ***Reframe data spaces with respect to data quality***, according to the different paradigms of quality identified by the paper. ***Explore further the relationship between data quality and data spaces***, taking into account both paradigms (quality of data as internal completeness and representativeness vs. measure of the suitability of the data for a particular high-level task) and identify new tools to evaluate and communicate the value of the data to other users in the data space. ***Involve key actors in the discussion***, such as, among others, Data Spaces Support Centre, EU Common Data Spaces and SIMPL and ***incorporate this aspect in the different blueprints and architectures***.

Finally, this document represents a first step from BDVA on its study of data quality, its importance from different views and its impact in most aspects of the data economy. BDVA plans to continue this work, by going deeper in some specific sections (fit for purpose, data quality for AI and ML, data value from data quality…), considering new perspectives on data quality (data quality in data mesh and data product, data quality in business, …) and bringing more experiences from European and national initiatives.

CHAPTER VII

ANNEX I. DESCRIPTION OF TOOLS FOR DATA QUALITY ASSESSMENT

Piveau Metrics[36]

This tool presents a practical and scalable solution to address the challenges of measuring and improving the quality of DCAT Application Profile (DCAT-AP) datasets, essential for Open Data publication and reuse in Europe. Based on the FAIR[37] and 5-star principles, the methodology defines concrete metrics across dimensions like Findability, Accessibility, Interoperability, Reusability and Contextuality, offering a quantitative representation of metadata quality. The microservice architecture ensures flexibility and extensibility, with a pipeline layer computing metrics and scores, a service layer for further data processing and a UI layer providing interactive visualisations and detailed reports. Piveau Metrics aims to provide an automated and scalable approach for assessing the quality of DCAT-AP datasets, enabling data providers to improve the quality of their data.

Ydata-quality[38]

An open-source Python library that evaluates data quality throughout the multiple stages of a data pipeline development. The library includes a data quality engine that performs several tests on the input data and warnings are raised depending on the data quality. It includes specific modules for each dimension, such as Bias and Fairness, Data Expectations, Data Relations, Drift Analysis, Duplicates, Labelings, Missings and Erroneous Data. By providing prioritised warnings and informative reports, YData Quality enables researchers to proactively identify and address data quality issues, enhancing the reliability and impact of AI solutions.

Apache Griffin[39]

Apache Griffin is an open-source data quality tool that provides a unified platform to measure and monitor the quality of data. The tool supports various data sources and formats, including structured, semi-structured and unstructured data. Apache Griffin uses a set of predefined data quality metrics that are extensible and customisable to assess and monitor data quality across various data sources and processing pipelines.

36 https://www.piveau.io/en/
37 https://www.go-fair.org/fair-principles/
38 https://github.com/ydataai/ydata-quality
39 https://griffin.apache.org/

Great Expectations[40]

Fully-fledged python-based data validation framework, it functions around the notion of “expectations" with respect to a target dataset. "Expectations” are very similar to unit tests and allow developers to set up individual data checks with respect to various data dimensions, including DQDs: format and domain validity, consistency etc. When an “expectation” fails, it returns a relevant sample from the database, thus helping with debugging. Besides its basic support for pandas' data frames, it comes with support for SQL databases and Spark data frames. It leverages Jupyter Notebook capabilities, which might impose development constraints.

Pandera[41]

Provides a flexible and expressive python-based API for runtime validation of data frames. Validation rules for completeness, domain and format validity, as well as hypothesis tests are defined in an accompanying validation schema, which is applied on the input data frame. Pandera was initially intended for observation-wise checks on “wide” data frames but has since been extended for column-wise checks, column-conditional checks and element-wise checks.

40 https://greatexpectations.io/
41 https://pandera.readthedocs.io/en/stable/

CHAPTER VIII

ANNEX II. DATA QUALITY ASSESSMENT WITHIN SMART DATA MODELS (EXTENDED)

The two DQ enhancement techniques are outlier/novelty detection and missing values imputation. The first one is represented in the data model as:

- **outlier: Includes information about the outlier characteristics of the measurement.**
    - isOutlier: Determines whether the measurement has been considered an outlier or not. It may take a Boolean value.
    - OutlierScore: A score indicating the degree of outlierness (anomaly score)
    - methodology: Reference (relationship) to another entity including AI methodology information.

Whereas the second enhancement technique is described as:

- **synthetic: Includes information about the origin of the measurement.**
    - isSynthetic: Determine whether the measurement has been created synthetically or not. It may take a Boolean value.
    - methodology: Reference (relationship) to another entity including AI methodology information.

Besides the DQ features, the basic fields of an NGSI-LD entity (*id* and *type*) are included, as well as *source* in order to identify the ownership or source of the information and *dateCalculated*, pointing out when this DQ assessment was performed. The remaining properties correspond to the DQ features.

Beginning with those related to the objective DQ dimensions, there is one property for each of them: *accuracy*, *completeness*, *timeliness* and *precision*. In turn, all of these are described as objects with four internal fields: *type*, *value*, *observedAt* and *unitCode*. The type property is specific to the NGSI-LD information model, value represents the numerical value (e.g. float) of the dimension, *observedAt* allows keeping a log of the time at which this property has been analysed and, lastly, *unitCode* indicates the unit code of the measurement.

As can be seen, there are two alternative ways of keeping temporal logs: with the *dateCalculated* property and with the *observedAt* subproperties. The former one allows the user or the data producer to record the DQ assessment entity as a whole, whereas the latter one belongs to each property, which is used to keep track of that property individually. Thus, including both logging methods enables the data model to be used widely, that is, by several use cases.

Furthermore, the *outlier* and *synthetic* properties relate to the applied DQ enhancement techniques employed over this piece of data, so that it is possible to trace back which pre-processing has been made over it. The first one, outlier, indicates within its internal fields whether or not the measurement corresponds to an anomalous value (*isOutlier: True/False*) and the anomaly score obtained (anomalyScore: number) whereas also including a methodology field to provide information on the AI method used to come to this conclusion. The latter, synthetic, determines through the *isSynthetic Boolean* property whether it was a missing observation in the time series which has been created synthetically using AI. A *methodology* field is included, as for the outlier property, to add traceable parameters of the application of these techniques.

In order to link the actual data value with its DQ features, the modelling that has been proposed would link a new field to the data models used for the representation of the streams. This way, the value entity (data stream) and the associated quality entity (DQ features) will be explicitly bound.

CHAPTER IX

ANNEX III. DATA QUALITY IN SPECIFIC SITUATIONS

IX.1 Data quality for streaming data

Some of the key considerations and strategies for maintaining quality in streaming data include [7]:

- ***Continuous monitoring and proactive alerting:*** Organisations should implement continuous monitoring and proactive alerting to ensure data quality in streaming data, as poor data quality can lead to flawed decisions and missed opportunities.

- ***Data quality assessment and profiling:*** Conducting comprehensive data quality assessments and profiling before ingesting streaming data is essential. This involves evaluating completeness, accuracy and other relevant quality dimensions to ensure the reliability of the incoming data.

- ***Data quality metrics and dimensions:*** Metrics such as completeness, consistency, timeliness, validity and uniqueness are crucial for assessing the quality of data in streams. These dimensions help in evaluating the structural, semantic consistency and overall usability of the streaming data.

- ***Testing and iterating data tools:*** Continuous testing and iteration of data tools, frameworks and patterns are necessary to ensure that streaming data meets evolving needs and expectations. This involves methods such as data profiling, auditing, lineage, governance and feedback to identify and resolve data quality issues.

- ***Quality monitoring for streaming data:*** Implementing quality monitoring for streaming data using tools like Spark Streaming and Delta Lake can help generate constraint suggestions based on historical ingest data and run incremental checks to ensure high-quality, high-velocity data.

- ***Comprehensive data quality solutions:*** Implementing comprehensive data quality solutions for stream and batch data processing is essential, addressing aspects such as data validation, cleansing, monitoring and remediation to ensure that the streaming data is reliable and consumable by downstream business-oriented consumers.

Ensuring data quality in streaming data requires a combination of continuous monitoring, comprehensive assessment, use of relevant metrics and dimensions and the implementation of robust data quality solutions to address the unique challenges posed by streaming data. Traditional data quality monitoring may focus more on maintaining data quality when data is at rest than on ensuring that data is of good quality when it is in motion. This can lead to interruptions and delays in operations, analytics and decision-making.

IX.2 Data quality in motion

Recurrent Neural Networks (RNN) and Ordinary Differential Equations (ODE) offer powerful tools for time series analysis, especially in contexts where data is irregularly sampled [8]. Integrating ODE into RNN models allows modeling continuous dynamics over time, which can be particularly useful for data that is not captured at regular intervals. These features can have interesting applications in data quality:

- **Continuous Data Quality Monitoring:** In environments where data is constantly generated, such as IoT systems or real-time financial transactions, it is crucial to maintain a high level of data quality to ensure the reliability of analyses and decisions based on this data. ODE-RNNs could be trained to recognise normal data quality patterns and therefore identify when data deviate from these patterns, indicating potential quality problems such as inaccuracies, inconsistencies or data corruption.

- **Anomaly Detection and Failure Prediction:** The ability of ODE-RNNs to model complex temporal dynamics makes them ideal for early anomaly detection. These models could identify atypical behavior in the data that could suggest underlying problems, such as sensor failures, data entry errors, or fraudulent activities. By detecting these anomalies early, organisations could take corrective action before problems escalate or affect subsequent analysis.

- **Imputation of Missing Data:** Missing data is a common challenge in many data analytics applications. By modeling the continuous evolution of data over time, ODE-RNNs could accurately predict missing values based not only on previous and subsequent observations, but also on the underlying trend and dynamics of the data. This could be particularly useful in time series with irregular sampling, where traditional imputation methods may not be effective.

- **Data Collection Optimisation:** Data collection is often expensive and subject to resource constraints. ODE-RNNs could help optimise this process by determining the most critical times for data collection, ensuring that the most relevant data for analysis are captured without incurring unnecessary costs due to overcollection. This is especially valuable in environmental or health monitoring applications, where the relevance and quality of the data collected is crucial.

CHAPTER X

ANNEX IV. STANDARDS

This section collects a set of standards relevant to the content of the document:

ISO/IEC DIS 5259-1 Artificial intelligence — Data quality for analytics and machine learning (ML) (https://www.iso.org/standard/81088.html), including:

- Part 1: Overview, terminology and examples
- Part 2: Data quality measures
- Part 3: Data quality management requirements and guidelines
- Part 4: Data quality process framework
- Part 5: Data quality governance framework
- Part 6: Data quality visualisation

ISO/IEC 25012:2008 "Software engineering", Software product Quality Requirements and Evaluation (SQuaRE), **Data quality model** (https://www.iso.org/standard/35736.html). This standard was last reviewed and confirmed in 2019.

ISO 8000-61:2016 – Data quality (https://www.iso.org/standard/63086.html) (This standard was last reviewed and confirmed in 2022)

ISO/IEC TR 24027:2021 Information technology — Artificial intelligence (AI) — Bias in AI systems and AI aided decision making (https://www.iso.org/standard/77607.html) (Publication date : 2021-11)

IEEE P7003 – Algorithmic bias Working Group (https://sagroups.ieee.org/7003/)

UNE 0079:2023. Data Quality Management (https://tienda.aenor.com/norma-une-especificacion-une-0079-2023-n0071118)

CEN CENELEC JTC21 – Specific Task Group (part of WG3: https://standards.cencenelec.eu/dyn/www/f?p=205:7:0::::FSP_ORG_ID:3125028&cs=1B85B1ABC4F454B7CB352839242CD7944), focused on Data Governance and Data Quality

CHAPTER XI

ANNEX V. REFERENCES

[1] Hassenstein, Max & Vanella, Patrizio. (2022). Data Quality - Concepts and Problems. Encyclopedia. 2. 498-510. 10.3390/encyclopedia2010032

[2] Fernando Gualo, Moisés Rodriguez, Javier Verdugo, Ismael Caballero, Mario Piattini, Data quality certification using ISO/IEC 25012: Industrial experiences, Journal of Systems and Software, Volume 176, 2021

[3] Curry, E. (2016). "The big data value chain: definitions, concepts and theoretical approaches. New horizons for a data-driven economy: A roadmap for usage and exploitation of big data in Europe", 29-37.

[4] Amrapali Zaveri; Anisa Rula; Andrea Maurino; Ricardo Pietrobon; Jens Lehmann; Sören Auer. Quality assessment for Linked Data: A Survey. Semantic Web, vol. 7, no. 1, pp. 63-93, 2015. URL: https://dx.doi.org/10.3233/SW-150175

[5] Fleckenstein, M.; Fellows, L. (2018). "Chapter 11: Data Quality". Modern Data Strategy. Springer. pp. 101–120. ISBN 9783319689920. Archived from the original on 31 July 2020. Retrieved 18 April 2020.

[6] Data Integrity and Data Governance, chapter on Data Quality Measurement Based on Domain-Specific Information. B. Santhosh Kumar, 26 April 2023, DOI 10.5772/intechopen.100778

[7] Mulyani, I., & Wibowo, A. (2021). "Data Quality Assurance and Governance in the Age of Big Data: Strategies, Challenges and Industry Applications". Emerging Trends in Machine Intelligence and Big Data, 13(7), 49-65.

[8] Rubanova, Yulia & Chen, Ricky & Duvenaud, David. (2019). Latent ODEs for Irregularly-Sampled Time Series. https://arxiv.org/abs/1907.0390

[9] E. Tragos, et. al. "Energy efficient AI-based toolset for improving data quality. First version", SEDIMARK Deliverable D3.1, Sept. 2023

[10] Ortigosa-Hernández, J., Inza, I., & Lozano, J. A. (2017). Measuring the class-imbalance extent of multi-class problems. Pattern Recognition Letters, 98, 32-38.

[11] Kalousis, A., Gama, J., & Hilario, M. (2004). On data and algorithms: Understanding inductive performance. Machine learning, 54, 275-312.

[12] Lorena, A. C., Costa, I. G., Spolaôr, N., & De Souto, M. C. (2012). Analysis of complexity indices for classification problems: Cancer gene expression data. Neurocomputing, 75(1), 33-42.

[13] Dhoni, Pan. (2023). Enhancing Data Quality through Generative AI: An Empirical Study with Data. 10.36227/techrxiv.24470032.v1

[14] "Foundation Data Space Models: Bridging the Artificial Intelligence and Data Ecosystems", Edward Curry

[15] "How to Move Beyond a Monolithic Data Lake to a Distributed Data Mesh". martinfowler.com.

[16] Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information (https://www.sciencedirect.com/book/9780128180150/executing-data-quality-projects)

XI.1 Additional bibliography

- JRC - Data quality requirements for inclusive, non-biased and trustworthy AI (https://publications.jrc.ec.europa.eu/repository/handle/JRC131097)
- Data.europa.eu - Data Quality Guidelines (https://op.europa.eu/en/publication-detail/-/publication/023ce8e4-50c8-11ec-91ac-01aa75ed71a1/language-en)
- High-value datasets – an overview through visualisation (https://data.europa.eu/en/publications/datastories/high-value-datasets-overview-through-visualisation)

About the Big Data Value Association

BDVA is an industry-driven international not-for-profit organisation with 250 members all over Europe and a well-balanced composition of large, small and medium-sized industries as well as research and user organisations. Our mission is to develop an innovation ecosystem that enables the data-driven digital transformation of the economy and society in Europe, delivering maximum benefit. To reach this goal, we focus on advancing areas such as big data technologies and services, data platforms and data spaces, industrial AI, datadriven value creation, standardisation and skills.

BDVA enables existing regional multi-partner cooperation, to collaborate at the European level through the provision of tools and know-how to support the cocreation, development and experimentation of pan-European data-driven and AI applications and services and know-how exchange.

Through BDVA, our members contribute to the European data and AI R&I agenda and develop guidelines and strategic roadmaps for industry and policymakers in BDVA Task Forces and our events give opportunities to build new collaborations and co-create new projects. Being part of the BDVA community, the members gain higher visibility on the European level and our services are designed to give timely updates on all the latest developments in the fields of data and AI.

BDVA believes in collaborations! BDVA has been the private side of the H2020 partnership Big Data Value PPP, it is a private member of the EuroHPC JU and it is a founder member of the AI, Data and Robotics Partnership. BDVA has developed a strong and growing cooperation with Gaia-X, IDSA and FIWARE through the Data Spaces Business Alliance (DSBA), it is a partner of the Transcontinuum Initiative (TCI) and collaborates with many industry-driven AI national initiatives and other European communities.

BDVA is open to new members!

Visit BDVA.EU to learn more about members and activities. You can contact us anytime at info@bdva.eu.

Note

This document should be referenced as follows: Elevating Data Quality: A Paradigm Shift for Data Spaces and AI Needs. BDVA. 2024.

BDVA Office
Data, AI and Robotics (DAIRO) aisbl
Avenue des Arts, 56
1000 Bruxelles
Belgium

BDVA.eu
info@bdva.eu