D4.2

Basic Dataset Computation Approach and Structured KPI Catalogue for the Valuation of Datasets

University College Cork

Funded by
the European Union

D4.2

Basic Dataset Computation Approach and Structured KPI Catalogue for the Valuation of Datasets

Revision **v0.1**

Work package	WP4
Task	T4.2
Due date	01-06-2025
Submission date	dd-mm-yyyy
Deliverable lead	University College Cork
Version	v0.1
Authors	Eduardo Vyhmeister (UCC), Andrea Visentin (UCC), Bastien Pietropauli (UCC), Alejandro Martinez Molina (ITI), Montserrat González Ferreiro (EGI), Agnieszka Raush (PSNC), Enrique Arteizaga (TEC)
Reviewers	Name Last Name (Partner name)

Abstract

This deliverable presents a structured framework for objectively valuing datasets, combining theoretical foundations and practical tools. Beginning with a systematic literature review across leading academic databases, it identifies and categorises over 100 data valuation metrics and key performance indicators (KPIs) into a coherent taxonomy aligned with the four perspectives

Funded by
the European Union

of the Balanced Scorecard. The taxonomy clusters metrics under Data Valuation Techniques, Data Monetisation, Data Quality & Governance, Operational Efficiency, Technology & Infrastructure, and Innovation & Growth, highlighting their interdependencies and strategic relevance.

For each metric and KPI, the deliverable provides clear definitions, computational formulas, and guidance on implementation, collected in appendices for easy reference. To translate these insights into action, a web-based decision support tool built on the Analytic Network Process (ANP) enables organisations to articulate their strategic profiles, perform pairwise comparisons, and prioritise the most relevant metrics and KPIs for their specific data monetisation goals.

By integrating rigorous academic research with practical computation methods and an interactive selection tool, this deliverable equips stakeholders with a comprehensive, strategy-aligned approach to quantify and optimise the value of their data assets.

Keywords

Data Monetisation, Data Valuation, KPI Catalogue, Metrics Computation, Balanced Scorecard, Taxonomy, Analytic Network Process (ANP), Systematic Literature Review, Decision-Support Tool

Document revision history

Version	Date	Description of change	Contributor(s)
v0.1	01-05-2025	1st version of deliverable template	Eduardo Vyhmeister (UCC), Bastien Pietropauli (UCC)

Disclaimer

The information, documentation and figures available in this deliverable are provided by the DATAMITE project's consortium under EC grant agreement **101092989** and do not necessarily reflect the views of the European Commission. The European Commission is not liable for any use that may be made of the information contained herein.

Copyright notice

Project co-funded by the European Commission in the Horizon Europe Programme

Nature of the deliverable	R

Dissemination level

PU	Public, fully open. e.g., website	✓
CL	Classified information as referred to in Commission Decision 2001/844/EC	
SEN	Confidential to DATAMITE project and Commission Services	

***** Deliverable types:

R: document, report (excluding periodic and final reports).
DEM: demonstrator, pilot, prototype, plan designs.
DEC: websites, patent filings, press and media actions, videos, etc.
OTHER: software, technical diagrams, etc.

Table of Contents

Funded by the European Union

List of Figures

Funded by the European Union

List of Tables

Abbreviations

AI	Artificial Intelligence
DoA	Description of Action
WP	Work Package
T	Task
BSC	Balanced Scorecard
IoT	Internet of Things
KPI	Key Performance Indicator
ANP	Analytic Network Process
AHP	Analytic Hierarchy Process

Funded by
the European Union

Executive Summary

In the context of rapid digital transformation, organisations face the challenge of assigning a fair and systematic value to their data assets, a process often referred to as data monetisation or data valuation. D4.2, "Basic Dataset Computation Approach and Structured KPI Catalogue for the Valuation of Datasets," addresses this need by defining a comprehensive catalogue of key performance indicators (KPIs) and metrics, along with methodologies for their computation, to support objective, strategy-aligned valuation of datasets.

The deliverable proceeds with a structured methodology comprising a systematic literature review (Section 3 Metrics and KPIs for Data Monetisation), taxonomy development (Section 2.3 Taxonomy), and the application of multi-criteria decision models (Section 4 ANP and AHP applied to metrics and KPIs within data Environment). The literature review follows a PICOC[1]-driven search across IEEE, ACM, and Scopus, screening over 2,000 publications to identify prevailing trends in data valuation metrics and KPIs, and answering two research questions on their coverage and interrelations. The resulting taxonomy aligns metrics and KPIs with the four perspectives of the Balanced Scorecard — Financial, Customer, Internal Processes, and Learning & Growth — grouped into subclusters.

Building on this foundation, the deliverable catalogues over 100 metrics and KPIs, organised by the taxonomy approach. Each metric is defined, contextualised within the taxonomy, and linked to computational formulas provided in Appendixes. Furthermore, to facilitate practical adoption, a web-based decision-support tool leveraging the Analytic Network Process (ANP) has been developed[2]. This tool guides users through strategy articulation, criteria weighting, and pairwise comparisons to prioritise metrics and KPIs that best align with their organisational objectives and data strategies.

[1] Carrera-Rivera, A., Ochoa, W., Larrinaga, F., & Lasa, G. (2022). How-to conduct a systematic literature review: A quick guide for computer science research. MethodsX, 9, 101895.

[2] https://datamite.insight-centre.org/

In summary, D4.2 delivers:

- A comprehensive KPI catalogue for dataset valuation.
- A structured list of underlying metrics.
- Computational approaches and formulas for all metrics/KPIs.
- A taxonomy that clusters and relates metrics/KPIs to strategic perspectives.
- A decision-support tool (ANP) for selecting and prioritising KPIs in practice.

These contributions equip stakeholders with a rigorous, strategy-aligned framework to quantify the value of their data assets, enabling informed decision-making and fostering data-driven innovation.

1 Introduction

DATAMITE is a project funded by the European Commission as part of the Horizon Europe programme and coordinated by the ITI - Technological Institute of Informatics. DATAMITE empowers European companies by delivering a modular, open-source and multi-domain Framework to improve DATA Monetizing, Interoperability, Trading and Exchange, in the form of software modules, training, and business materials.

DATAMITE unleashes the monetization potential at two levels. At internal level, users will have tools to improve quality management of their data, the adherence to FAIR principles, and will be able to upskill on technical and business aspects thanks to the multiple open-source training materials the project will generate. Therefore, data will become trustable and more reliable also in other paradigms like AI.

At external level, keeping users in control of their data will provide new sources of revenue and interaction with other stakeholders. The architecture envisioned for DATAMITE enables DIHs sandboxing, becoming a potential instructor on their onboarding of SMEs and low-tech SMEs into the data economy. Together, DATAMITE’s solutions will function as a catalyst to boost data monetization in the European productive fabric.

1.1 Deliverable Purpose and Scope

Specifically, the Grant Agreement states the following regarding this Deliverable: *The report includes a catalogue and defines the approach required to determine the value of data sets using KPIs.*

Hence, the purpose of this deliverable, **Deliverable D4.2 – Basic Dataset Computation Approach and Structured KPI Catalogue for the Valuation of Datasets**, is:

- To provide a comprehensive list of KPIs catalogue for the valuation of datasets.
- To provide a comprehensive list of metrics useful to determine the KPIs catalogue.
- To provide methods (Approach) to estimate the metrics and KPIs defined in the catalogue.

Independent of the goals defined here, we have extended the contribution of the work performed in this task by incorporating the following:

- To provide a taxonomical approach to cluster and classify metrics and KPIs
- Generate a useful tool for users to better identify the most suitable metrics and KPIs for their interest to track.

To further understand the objectives and goals of this deliverable, we can refer to the description of **Task 4.2 Definition of KPIs for a Fair Valuation of Datasets and Mechanisms for their Computation**. As stated, T4.2 seeks to "(1) *define an objective approach to determine the value of datasets. Initially,* (2) *existing approaches for the evaluation of datasets will be identified and structured regarding requirements and properties within a catalogue. Based on this catalogue,* (3) *criteria for the fair value evaluation of datasets will be defined as well as KPIs related to categories such as People, Process, Data and Technology.* (4) *Finally, T4.2 defines a series of mechanisms and proceedings to calculate these KPIs, considering the strategy profiles (T4.1).* "

As will be seen in the deliverable, these four contributions, seen as objectives in T4.2. were achieved and exceeded expectation by developing further components that enable users to define suitable metrics based on categories that facilitate the implementation in any enterprise. To be concise:

Objective (1) sought to define an objective method to determine a dataset value. In this task, a fundamental link between dataset characteristics and their overall value was created by devising an approach that combines the most crucial metrics of the dataset into a single, aggregated framework. The idea was to ensure that the perspective on "value" includes more than mere size or comprehensiveness and instead considers various factors such as data quality, relevance, richness, and applicability. By focusing on these multiple facets, the objective approach ensures that no single attribute is weighted disproportionately, allowing for a fair and balanced valuation.

Objective (2) focuses on identifying and structuring existing approaches for dataset evaluation. This requirement was fulfilled through a systematic literature review focused on metrics and KPIs in the data monetisation environment. Such a review helped uncover existing methodologies on how datasets are valued and monetised. Furthermore, this helps us to contribute in T4.3 which include defining objective valuation models. The idea is that these models can be improved by the outcomes of T4.2. Additionally, we defined a taxonomy which helps DATAMITE users correlate several metrics within similar concepts.

Objective (3) focused on defining criteria for dataset valuation and establishing KPIs linked to categories such as People, Process, Data, and Technology. These criteria were structured following the well-known Balanced Scorecard approach. By adapting the Balanced Scorecard framework, the project could methodically translate high-level strategies and goals into tangible metrics. This ensures that any organisation or stakeholder using these KPIs can holistically evaluate a dataset's worth by considering the human, procedural, technical, and data-specific elements in a balanced manner.

Objective (4) concludes by providing the mechanisms and procedures for computing the defined KPIs. To enable practical calculation, a set of equations (as outlined in appendix B) is included. These formulas and computational guidelines give stakeholders a toolset for quantifying dataset value, thereby fulfilling the objective of offering a systematic method for fair and data-driven valuation.

Additionally, in this task, we developed a web-based tool leveraging the Analytic Network Process (ANP) to help organisations systematically prioritise and refine their dataset metrics. This solution addresses a key challenge in data monetisation: navigating an abundance of potential KPIs in modern, interconnected environments, where linear scorecards often fail to capture the influence of complex, interdependent factors.

1.2 Document Structure

This deliverable is broken down into the following sections:

- **Executive Summary**: A summary of the contents and main findings is provided.
- **Section 1 Introduction** provides the deliverable general context, dependencies, and structure.

- **Section 2 Background** provides an overview of the main topics related to strategies and possible clustering and classification, used in this study, for metrics and KPIs linked to data valorisation and data monetisation.
- **Section 3 Metrics and KPIs for Data Monetisation** presents the results obtained during the literature review. More specifically, this section presents the taxonomic results within the Data Monetisation environment and the results of performing a systematic literature review that focuses on two specific research questions.
- **Section 4 ANP and AHP applied to metrics and KPIs within the data Environment** present the strategy and tools used to define the most suitable metrics and KPIs based on the different strategies involved in data monetisation.
- **Section 5 Conclusions** presents the main results and definitions of future works for new deliverables

Appendixes:

- **Appendix A** Supplementary Information regarding metrics and KPIs for Data Monetisation
- **Appendix B** KPI and Metrics Table

2 Background

This section provides an overview of strategies and possible clustering and classification for metrics and KPIs linked to data valorisation and data monetisation. Also, to facilitate the reading and understanding of the following sections, one of the main results of T4.2 is presented at the end of this section (the taxonomy). How this taxonomy was constructed is explained in greater detail in the corresponding subsection (3.6.2 RQ2 - What are the main trends and relations observed between the observed KPIs and Metrics for data monetisation/valuation?).

2.1 Strategy, Metrics and KPIs

To fully appreciate the distinctions between KPIs and metrics, it is helpful to view them in the context of a business strategy. Essentially, a business strategy is an organisation's comprehensive plan for achieving its long-term goals and establishing a sustainable competitive edge. Acting as a guiding framework, the strategy shapes how a company allocates resources, makes strategic decisions, and sets priorities aligned with its long-term vision.

Business strategies are grounded in the company’s mission, vision, and current position within the market. They clarify what success means to the organisation, highlight its unique value, and outline how it intends to navigate industry challenges and competitive pressures. For instance, a company aiming to lead in sustainability may focus its strategy on implementing eco-friendly innovations and operational practices. This focus differentiates the organisation, as well as defines its strategic goals and the kinds of performance indicators it will prioritise to measure progress.

With a well-defined strategy, an organisation can establish specific objectives that support its overarching vision. These objectives are often broken down into concrete, measurable targets, enabling the company to monitor its trajectory over time. At this stage, KPIs and metrics become essential: they are tools to quantify the strategy and translate it into actionable insights. KPIs, by emphasising critical success factors, and metrics, by capturing a broader array of operational data, create a framework through which organisations can steer and refine their strategic progress.

To be more specific, KPIs are high-level, critical measures directly related to a company's strategic objectives. They indicate whether the business is on track to reach its long-term goals, serving as benchmarks for success. KPIs are selected with strategic alignment in mind, focusing only on the aspects of the business that drive growth and impact the organisation's overall objectives. Generally speaking, any metric can become a KPI as long as it is delimited between maximum and minimum ranges. These ranges are normally linked to targets, agreements or benchmarks that facilitate the definition of the KPI (e.g. $KPI_i = \frac{M_{current} - M_{minimum}}{M_{benchmark} - M_{minimum}}$), where M is a metric.

For instance, if a company's growth strategy emphasises customer satisfaction, a relevant KPI might be "achieve a X% customer satisfaction rate by the end of the year." This KPI provides a measurable target and a clear focus area for the organisation, steering decision-making across teams toward a shared outcome.

Because KPIs are tied to strategic goals, they are typically fewer in number, highly specific, and reviewed regularly to ensure alignment with the company's evolving strategy. On the other hand, metrics are broader measurements used to track various aspects of a business's operations. Unlike KPIs, metrics do not always align directly with strategic goals, but they are essential in managing day-to-day activities and understanding different parts of the business. Metrics provide detailed insights that help companies evaluate their performance, identify trends, and make improvements in specific areas.

2.2 Data Monetisation and its Structured Facets

The Balanced Scorecard (BSC) is a practical framework that helps companies to connect their day-to-day activities with their broader strategic goals [58]. The BSC, developed in the 1990s, was meant to go beyond traditional financial performance metrics by including non-financial indicators, allowing organisations to track a more complete picture of their performance. Over time, it has become a powerful tool for aligning an organisation’s long-term vision with clear, actionable goals. Having this ability to link a company’s overall strategy to specific, measurable outcomes has made the BSC essential to many businesses aiming to achieve their strategic objectives.

Beyond aligning operations with strategy, the BSC is also highly effective for grouping and organising KPIs. The BSC’s structure of four main perspectives — Financial, Customer, Internal Processes, and Learning and Growth — naturally lends itself to categorising KPIs based on what part of the business they impact, making it easy to identify which KPIs are needed in each area. Here is how each perspective contributes to aligning strategy with measurable outcomes:

Financial: This perspective focuses on how a company’s strategic choices impact financial outcomes. Metrics such as revenue growth, profit margins, and return on investment help assess whether strategic actions are driving the expected financial benefits. By linking financial targets with strategic priorities, organisations can gauge their ability to deliver shareholder value, offering executives a way to evaluate how strategic shifts are influencing overall financial health.

Customer Perspective: Recognising that customer satisfaction and loyalty are foundational for long-term success, the BSC emphasises metrics like customer satisfaction scores, market share, and brand loyalty. These indicators help track how effectively the company meets customer needs and provides value. By linking customer-oriented objectives with broader strategy, the BSC helps organisations focus on building lasting customer relationships.

Internal Processes Perspective: This area examines the efficiency and effectiveness of essential internal processes that impact customer experience and financial performance. Typical metrics might include product development timelines, production efficiency, or quality standards. By prioritising improvements in these internal processes, the BSC encourages companies to refine operations in ways that support strategic objectives. This alignment ensures that internal operations are structured to meet strategic goals, which ultimately fosters a more streamlined and competitive organisation.

Learning and Growth Perspective: The BSC also highlights the importance of continuous development of organisational capabilities, from workforce skills to information systems and culture. Metrics in this area may involve employee training hours, knowledge-sharing practices, and skill enhancement programs. By connecting these learning-oriented objectives to strategic aims, the BSC underscores the importance of fostering an adaptable and skilled workforce, which is essential for sustaining innovation and a competitive position over time.

The Balanced Scorecard (BSC) is a flexible tool that helps organisations set up and track strategies focused on data monetisation. By using the BSC's four perspectives —Financial, Customer, Internal Processes, and Learning and Growth — companies can create a strategy that captures the value of their data assets. Breaking down each perspective into smaller, specific clusters further simplifies the process, making it easier to organise metrics that translate smoothly into KPIs. This approach not only keeps data monetisation efforts aligned with strategy but also makes tracking progress and outcomes more straightforward and consistent across different areas.

To facilitate tracking strategic achievements, we have defined different subclusters. Metrics and KPIs are categorised based on the analysis performed on the different metrics and KPIs in the literature, however, they are presented now to ease the readability and understanding of the document. Table 1 describes the clusters, subclusters, and their intrinsic definition.

BSC	Sub cluster	Definition
Financial Perspective	Data Valuation Techniques	Techniques, and its derived metrics and KPIs, that help to assess the worth of data assets. These techniques could be varied and thus their metrics and KPIs can varied that are not adequate for all techniques. Independently, they can help determine potential revenue-generating opportunities, calculate cost savings, define the relative value of data, and understand the economic impact of it. Given their

		relative impact, they can help to define KPIs and metrics for Data Monetisation (next sub cluster).
	Data Monetisation	KPIs and metrics that provide measurable evidence of the financial benefits derived from data initiatives. These metrics track the revenue generated through data-driven products or services, calculate cost reductions from improved decision-making, and gauge the return on investment (ROI) from data-focused projects. The focuse in this subcluster is price, not value (as previous cluster).
Customer Perspective	Customer Needs and Satisfaction	Evaluate the alignment of data products or insights with customer needs and expectations. These indicators include customer retention, engagement, and/or satisfaction tied to data-related services.
	Market Penetration	Indicators that assess the extent to which data solutions have penetrated target markets, measuring factors such as market share growth and the adoption rates of new data products. These KPIs allow companies to gauge the reach and impact of their data monetisation strategies within the market, ensuring that these initiatives are effectively attracting and retaining customers.
Internal Processes	Data Quality	Indicators that measure the accuracy, completeness, consistency, and other quality-related measures/indicators of data. High-quality data is essential for dependable insights and sound decision-making, serving as the backbone of successful data

		monetisation efforts. By ensuring that data meets high standards of quality, organisations can generate more accurate analyses and insights, impacting the reliability of monetisation outcomes.
	Data Governance and Compliance	Indicators that monitor adherence to data governance principles and regulatory standards, such as data privacy and security requirements. By aligning data practices with legal and ethical guidelines, these metrics reduce the risks associated with data breaches and regulatory non-compliance, thus fostering trust with clients and stakeholders.
	Operational Efficiency	Metrics that evaluate how effectively data is integrated into the company's operational processes, assessing the impact of data on productivity and resource management. These indicators help track improvements in process efficiency due to data utilisation, ensuring that data-driven operations are streamlined and resources are optimised for maximum productivity.
Learning and Growth	Technology and Infrastructure	KPIs that assess the organisation's capability to manage large-scale data initiatives, focusing on aspects such as data storage, processing power, and analytics infrastructure. This sub cluster ensures that the technical foundation for handling data monetisation is in place, enabling the organisation to store, analyse, and leverage data effectively. Investments in technology and

Funded by
the European Union

		infrastructure support scalability and enhance the organisation’s capacity for data-driven innovation.
	Innovation and Growth-Oriented	Indicators that measure the organisation’s progress in fostering a culture of data-driven innovation, including the development of new data products, R&D spending on data initiatives, and employee engagement in data-focused projects. By tracking innovation-oriented KPIs, organisations can ensure they are continuously evolving to capitalise on emerging data opportunities, promoting long-term growth and adaptation in a competitive landscape.

Table 1 BSC Components, Sub clusters, and Definitions for KPIs and Metrics Analyses

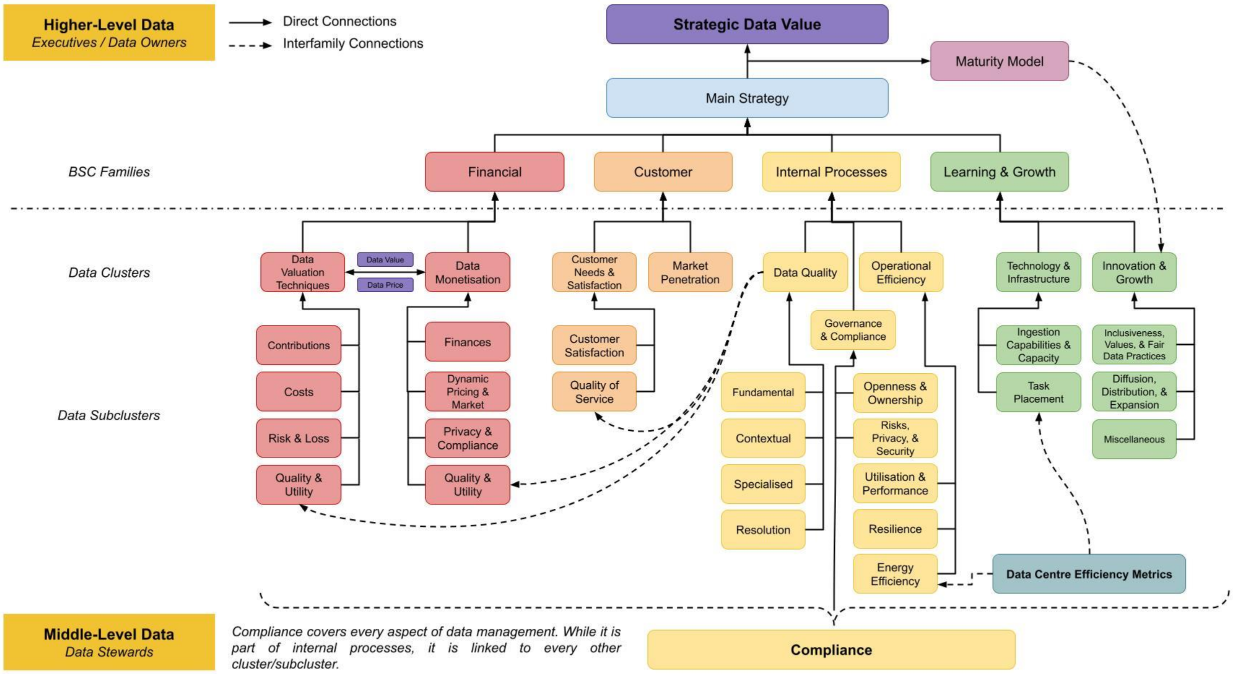
2.3 Taxonomy

Figure 1 through Figure 9 illustrate the taxonomy developed to link metrics and KPIs in a structured and coherent manner. This taxonomy is introduced here to support the understanding of the subsequent sections. The connections between metrics are discussed in detail in section 3.6.1, while section 3.6.2 elaborates on the taxonomical approach and its practical implementation. Additional information on the estimation of specific metrics is provided in the Appendix.

As shown in **Error! Reference source not found.**, different levels of relative importance for metrics and KPIs are structured following the Balanced Scorecard (BSC) approach. From top to bottom, the framework aggregates the four main BSC components into a global strategic metric (i.e., a global KPI). Each BSC component is composed of various data clusters, as defined in Table 1. These clusters are then futher clustered (e.g. Data Valuation technique subcluster is composed of four sub-subclusters) allowing for a more refined characterization of metrics. This layered structure facilitates the aggregation and technical evaluation of metrics.

Figure 2 through Figure 8 expand on each of these data subclusters. These figures illustrate how subclusters relate to specific metrics, which may belong to the same or to different BSC components.

The taxonomy is structured across three hierarchical levels to ensure a logical information progression from technical detail to high-level strategy. At the highest level, illustrated in Figure 1 the taxonomy aligns with the BSC structure and reflects the strategic and business perspective. The intermediate level, also visible in Figure 1, corresponds to the role of data stewards, who can drive or interpret and refine information from the technical domain. Finally, In the Technical level, observed from Figure 2 through Figure 8, comprises the processes of data collection, processing, and metric extraction. This level provides the foundational inputs for metric estimation, which are then integrated upward through the taxonomy.



Compliance covers every aspect of data management. While it is part of internal processes, it is linked to every other cluster/subcluster.

Figure 1 High Level Taxonomy and its Connection to the BSC Perspectives

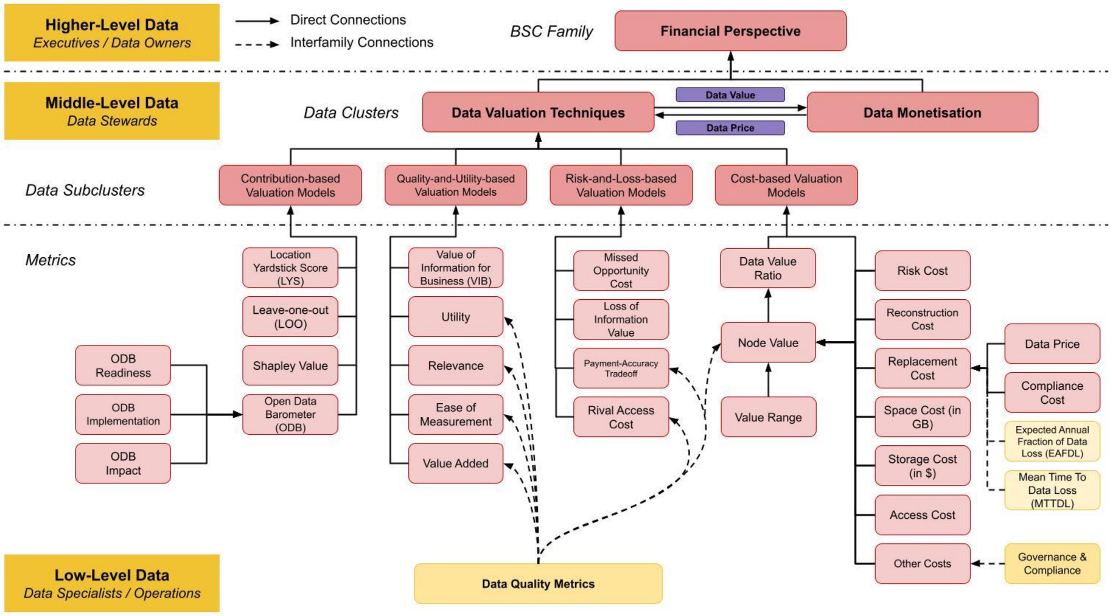


Figure 2 Data Valuation Techniques Subcluster of the Financial Taxonomical Perspective

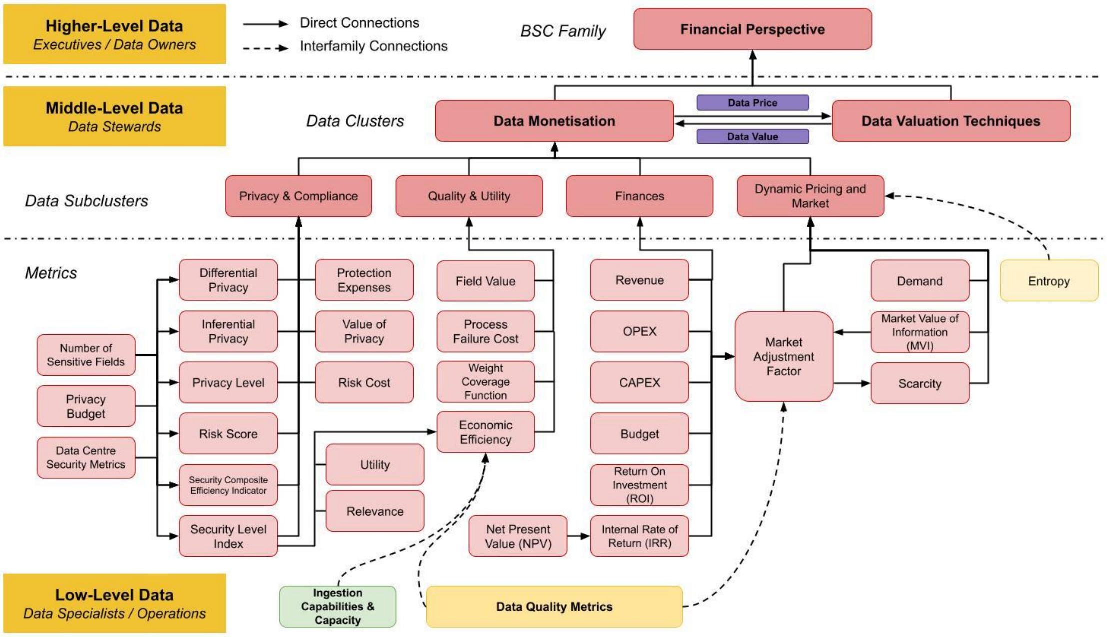


Figure 3 Data Monetisation Subcluster of the Financial Taxonomical Perspective

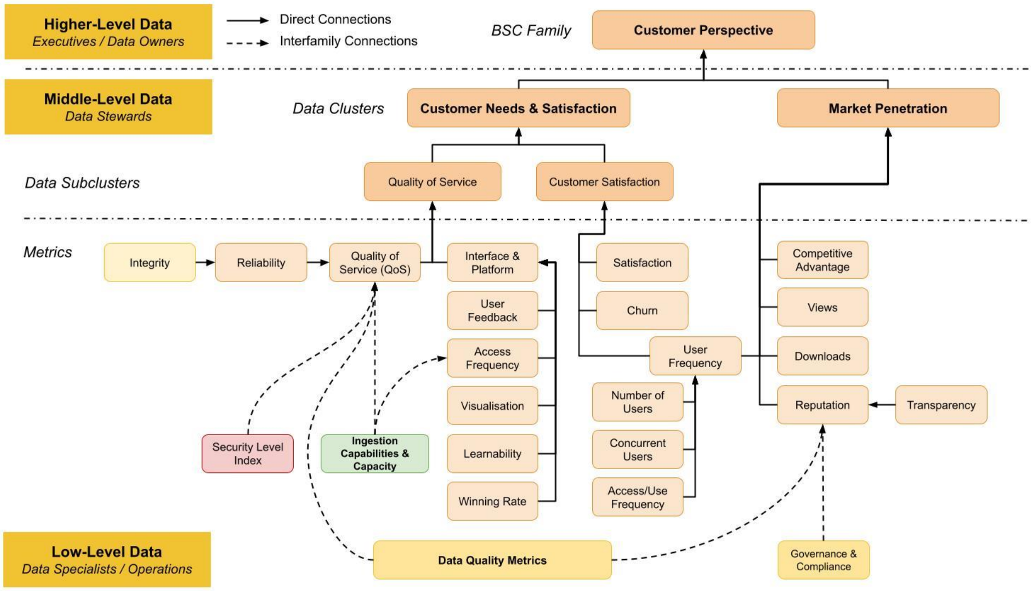


Figure 4 Customer Taxonomical Perspective

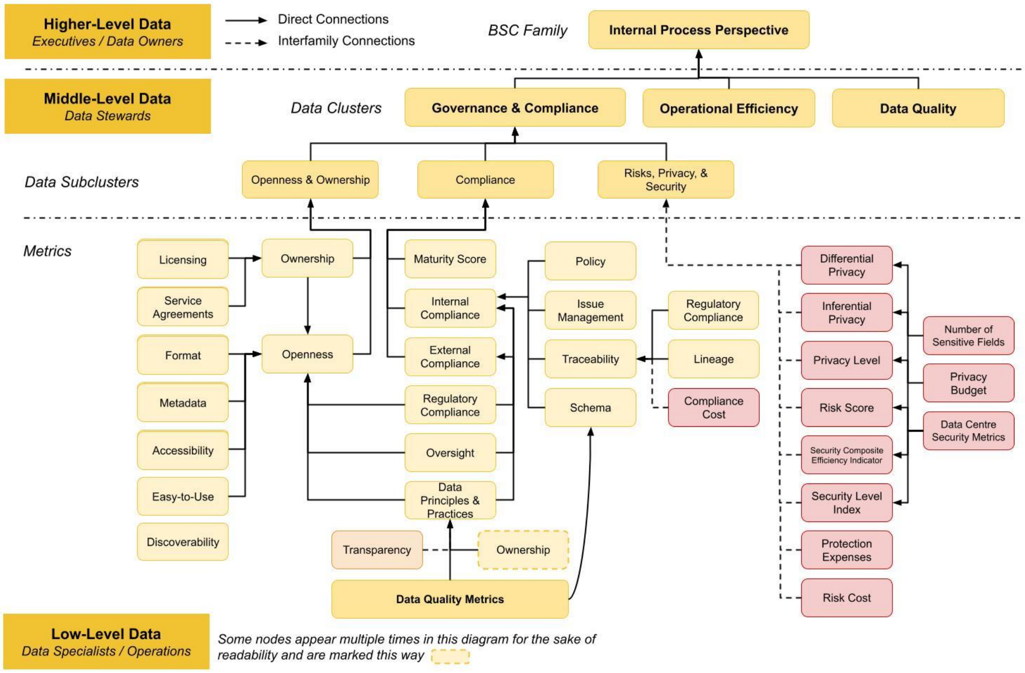


Figure 5 Governance & Compliance Subcluster of the Internal Process Taxonomical Perspective

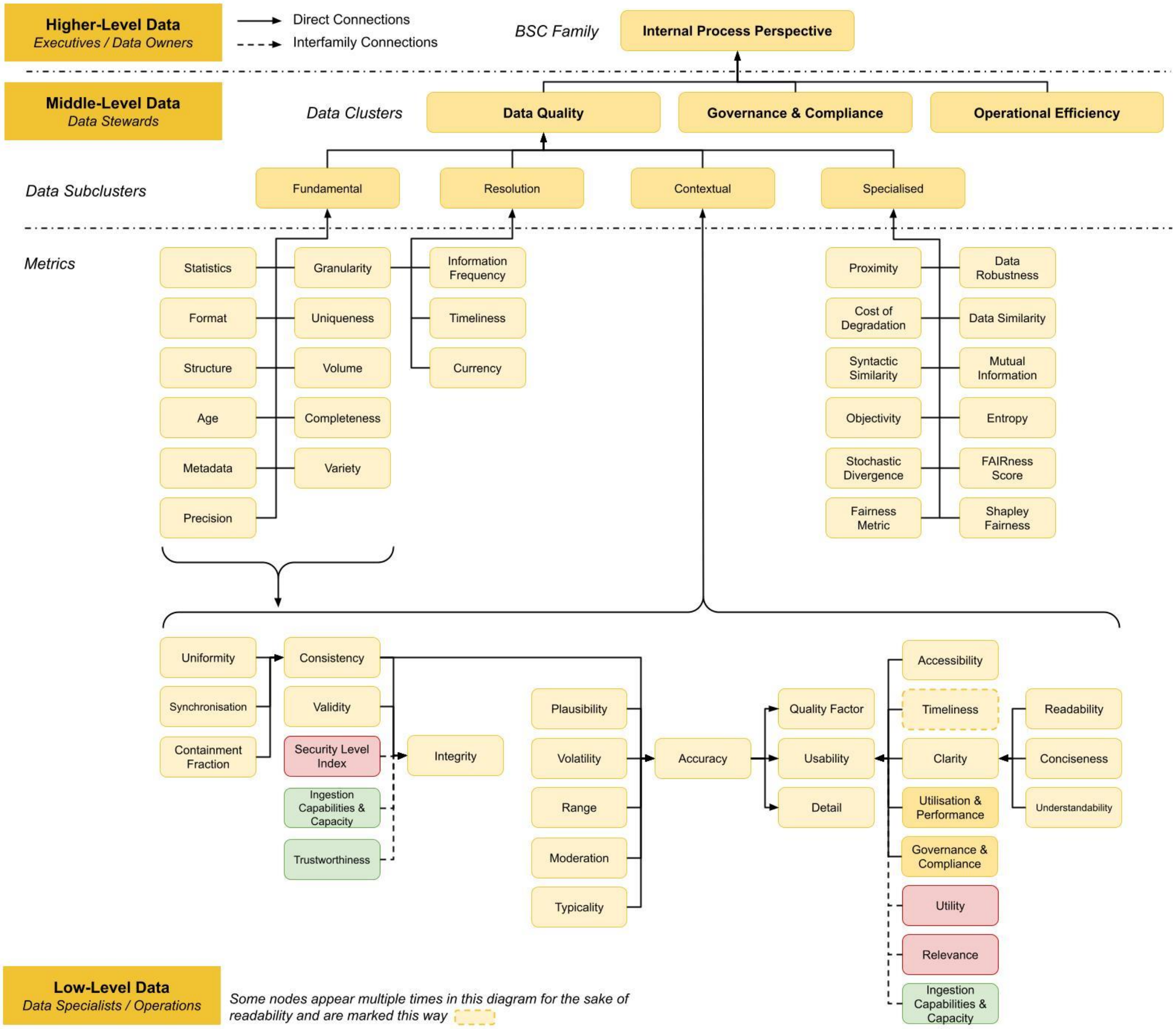


Figure 6 Data Quality Subcluster of the Internal Process Taxonomical Perspective

There is only one additional cluster in Figure 7 and Figure 8 that corresponds to Figure 9. This cluster, named Data Centre Efficiency Metrics, is specific for data centres and corresponding strategies and thus, its evaluation depends on the strategy followed by the taxonomy user.

In Figure 9, two types of metrics exist, those that a re linked to “Greennes” metrics (i.e., they have a sustainability considerations) and those that are not directly linked to it (dark green).

Funded by the European Union

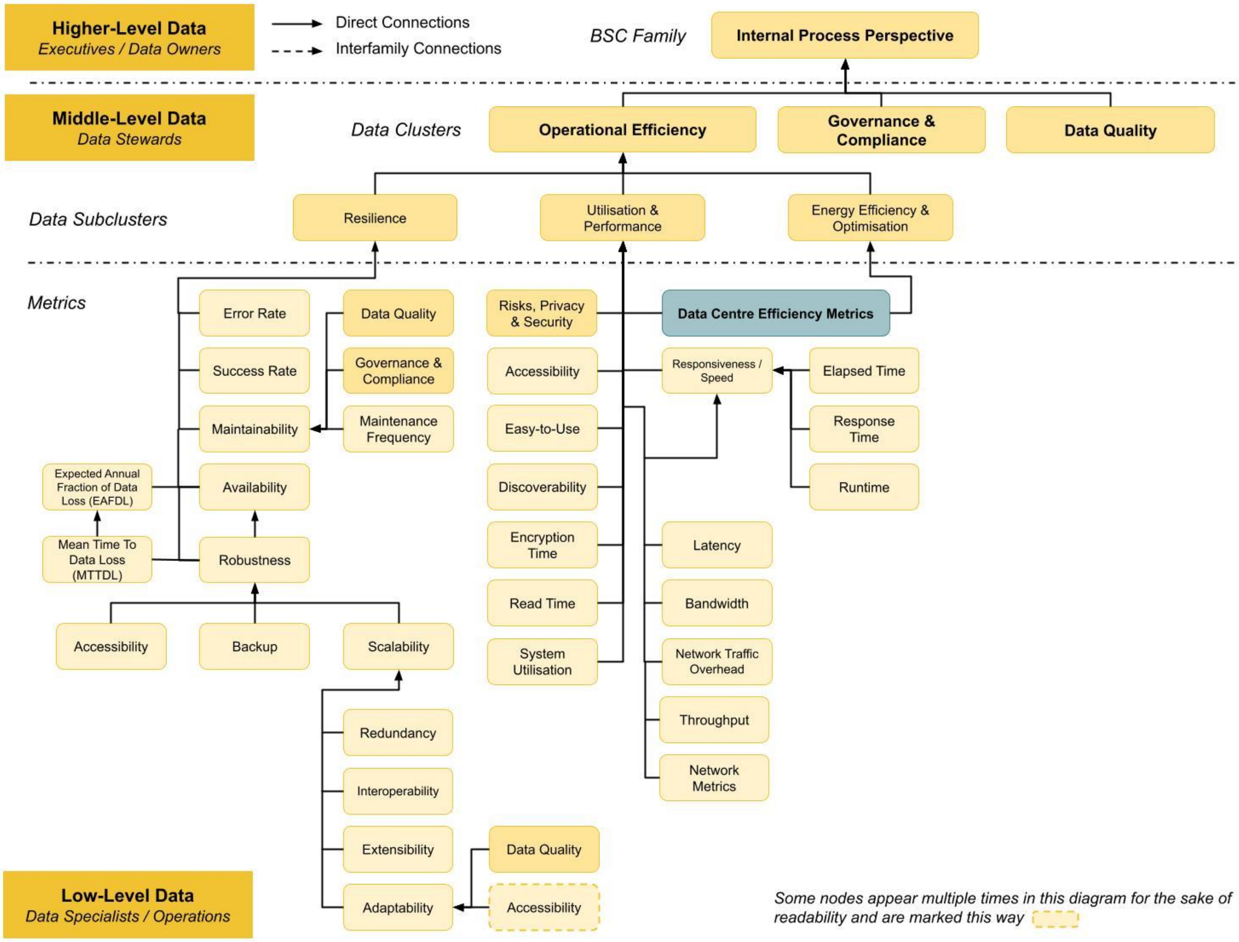


Figure 7 Operational Efficiency Subcluster of the Internal Process Taxonomical Perspective

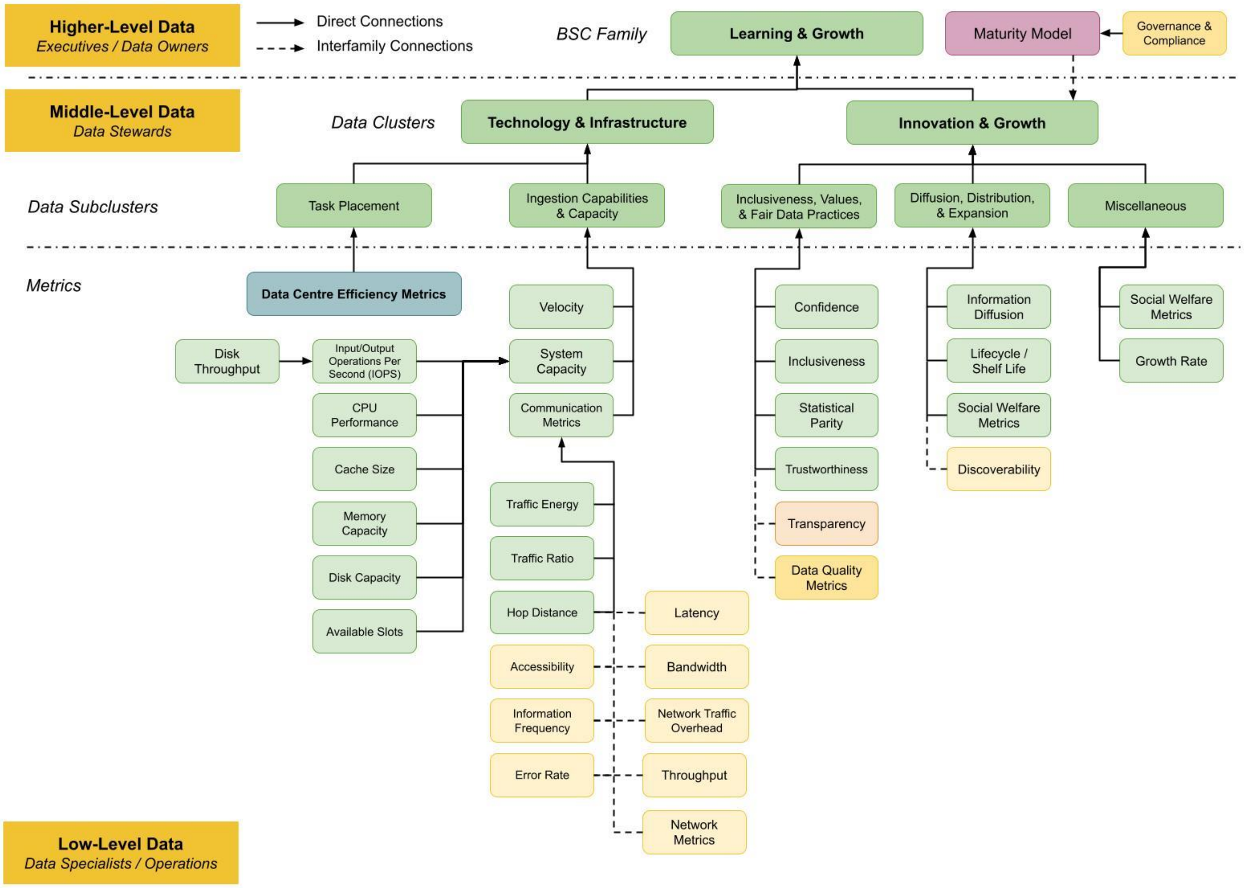


Figure 8 Learning & Growth Taxonomical Perspective

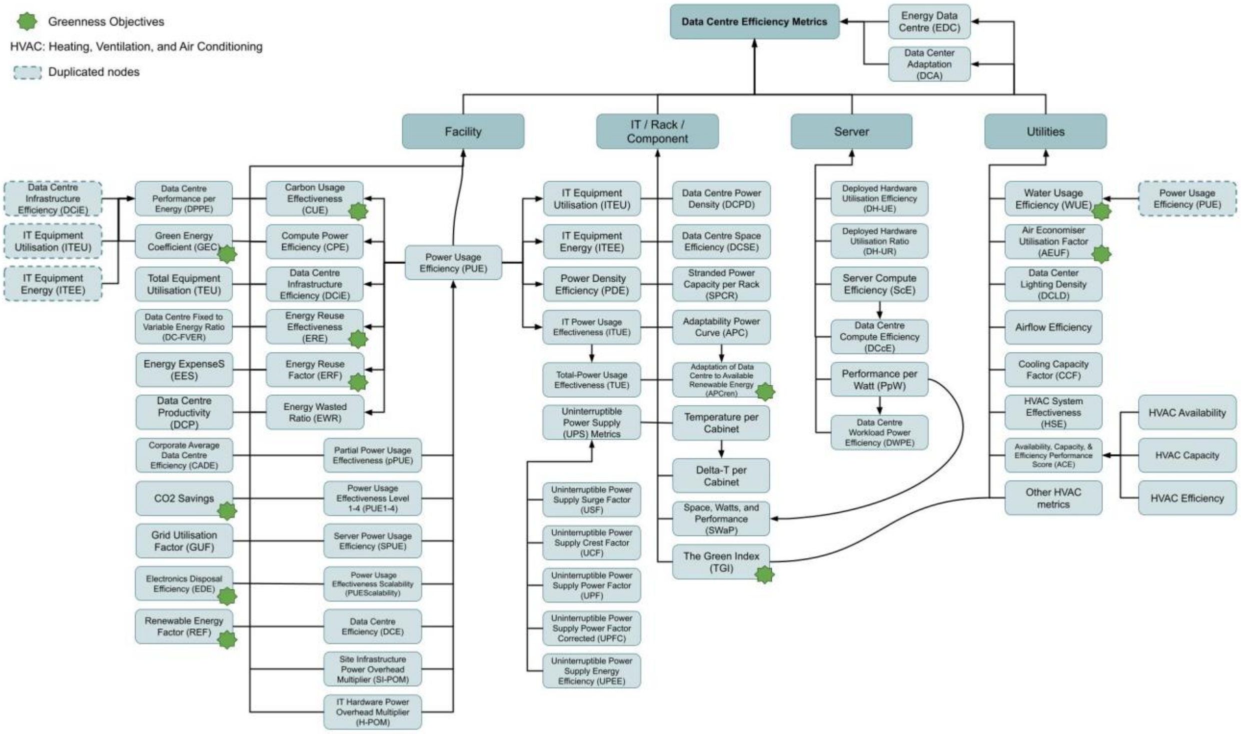

Greenness Objectives
HVAC: Heating, Ventilation, and Air Conditioning
Duplicated nodes
Data Centre Efficiency Metrics
Energy Data Centre (EDC)
Data Center Adaptation (DCA)
Facility
IT / Rack / Component
Server
Utilities
Data Centre Infrastructure Efficiency (DCiE)
IT Equipment Utilisation (ITEU)
IT Equipment Energy (ITEE)
Data Centre Performance per Energy (DPPE)
Green Energy Coefficient (GEC)
Total Equipment Utilisation (TEU)
Data Centre Fixed to Variable Energy Ratio (DC-FVER)
Energy ExpenseS (EES)
Data Centre Productivity (DCP)
Corporate Average Data Centre Efficiency (CADE)
CO2 Savings
Grid Utilisation Factor (GUF)
Electronics Disposal Efficiency (EDE)
Renewable Energy Factor (REF)
Carbon Usage Effectiveness (CUE)
Compute Power Efficiency (CPE)
Data Centre Infrastructure Efficiency (DCiE)
Energy Reuse Effectiveness (ERE)
Energy Reuse Factor (ERF)
Energy Wasted Ratio (EWR)
Power Usage Efficiency (PUE)
Partial Power Usage Effectiveness (pPUE)
Power Usage Effectiveness Level 1-4 (PUE1-4)
Server Power Usage Efficiency (SPUE)
Power Usage Effectiveness Scalability (PUEScalability)
Data Centre Efficiency (DCE)
Site Infrastructure Power Overhead Multiplier (SI-POM)
IT Hardware Power Overhead Multiplier (H-POM)
IT Equipment Utilisation (ITEU)
IT Equipment Energy (ITEE)
Power Density Efficiency (PDE)
IT Power Usage Effectiveness (ITUE)
Total-Power Usage Effectiveness (TUE)
Uninterruptible Power Supply (UPS) Metrics
Uninterruptible Power Supply Surge Factor (USF)
Uninterruptible Power Supply Crest Factor (UCF)
Uninterruptible Power Supply Power Factor (UPF)
Uninterruptible Power Supply Power Factor Corrected (UPFC)
Uninterruptible Power Supply Energy Efficiency (UPEE)
Data Centre Power Density (DCPD)
Data Centre Space Efficiency (DCSE)
Stranded Power Capacity per Rack (SPCR)
Adaptability Power Curve (APC)
Adaptation of Data Centre to Available Renewable Energy (APCren)
Temperature per Cabinet
Delta-T per Cabinet
Space, Watts, and Performance (SWaP)
The Green Index (TGI)
Deployed Hardware Utilisation Efficiency (DH-UE)
Deployed Hardware Utilisation Ratio (DH-UR)
Server Compute Efficiency (ScE)
Data Centre Compute Efficiency (DCcE)
Performance per Watt (PpW)
Data Centre Workload Power Efficiency (DWPE)
Water Usage Efficiency (WUE)
Air Economiser Utilisation Factor (AEUF)
Data Center Lighting Density (DCLD)
Airflow Efficiency
Cooling Capacity Factor (CCF)
HVAC System Effectiveness (HSE)
Availability, Capacity, & Efficiency Performance Score (ACE)
Other HVAC metrics
Power Usage Efficiency (PUE)
HVAC Availability
HVAC Capacity
HVAC Efficiency

Figure 9 Data Centre efficiency metrics connected to the Operational Efficiency and the Learning & Growth Taxonomical Perspectives (Figure 7 and Figure 8). Dependencies extracted from [95, 123, 116].

3 Metrics and KPIs for Data Monetisation

3.1 Systematic Literature Review – Research Methodology

Figure 10 illustrates the research methodology. In the top row of the figure, processes are linked to their primary outcomes in the bottom row. This includes the scope of the review (i.e. definition of research questions), conducting the search (i.e. queries definition), the bibliometric analysis, screening papers, and the classification scheme (i.e. taxonomy). The Bibliometric Analysis is not the core of the methodology, but it is used to ground a better understanding of the topic trends. They are developed in the following subsections as shown in the figure.

Funded by the European Union

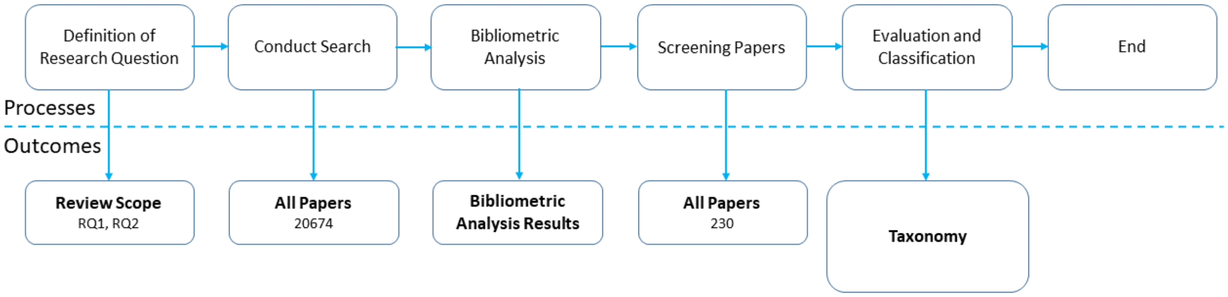


Figure 10 Systematic Mapping Approach

3.2 Definition of Research Questions

The goal of the present study is defined as follows:

To provide an understanding of different approaches, indicators, and metrics to measure the value (monetary or not) of information. The previous indicators should primarily be focused on the strategies, environments, and stakeholders that participate in the market now and in the foreseeable future.

To achieve this goal, the following research questions need to be answered:

RQ1 - What are the main KPIs and Metrics covered in the literature related to data monetisation/valuation?

RQ2 - What are the main trends and relations observed between the observed KPIs and Metrics for data monetisation/valuation?

3.3 Search Methodology

To assess the status of metrics and KPIs for data monetisation, a systematic approach was conducted. The process starts with the definition of the key pillars that drive the topic and the use of targeted keywords to conduct literature searches [87]. These keywords were used as markers, enabling efficient search across various platforms. Keywords and pillars were combined to address the research questions. Specifically, keyword selection was approached by establishing specific groups to help with the definition of the components of the keywords, taking a bottom-up approach. The methodology applied to construct queries and the definition of keywords follows the PICOC methodology – Population, Intervention, Comparison, Outcomes, and Context [88, 89, 59]. Each of these topics is defined next:

- **Population**: It refers to the specific group of individuals or subjects under the interest of the study. In the context of this work, the population is diverse and thus definition of different keywords representing the population is needed. An explanation of this process is provided later, nevertheless, we used the following keywords for representing the population of stakeholders that can trade and maintain (not necessarily use) data for different goals - data centre; data platform; data lakes; data warehouse; Information System.
- **Intervention**: The intervention refers to the approach or technique applied in the empirical study. Here, the manipulation are the intrinsic KPIs and Metrics over the data, and their calculation is the intervention process. Independent of this, to facilitate the identification on research queries, different keywords were used for the same component. In this case, these were: KPI, Key performance Indicators, Metrics, and Indicators.
- **Comparison**: The comparison component involves differentiating methods, processes, or strategies. Since KPIs are closely related to strategies, the family of KPIs could be further clustered. Nevertheless, since there is no clear classification or relation between KPIs, metrics, and strategies for data monetisation at this point, no further considerations were done for this component.
- **Outcomes**: Since empirical approaches or comparisons are not considered, no specific outcomes are defined in this component.
- **Context**: The context provides a comprehensive view of whether the study is conducted in academia or industry, the industrial segment, and the subject’s incentives. Thus, the context imposes keywords to constrain the search, for example, the use of the word KPI or Key Performance Indicator. For our case, the context is diverse, considering domains for the technological consideration of data, the market context of data, and the population relevant to the study (i.e. stakeholders).

To perform initial data monetisation queries, we used thesaurus tools to identify similar terms related to ”data,” ”value,” and ”data centres” to ensure comprehensive coverage of relevant terminology. The search was conducted in ACM, Scopus, and the IEEE journal, focusing on terms commonly associated with data monetisation. Table 2 shows the terms and their frequency.

Theme	Explanation and Values Found
Dataset	We retained core terms related to data assets and management: ”Dataset” (328,996), ”Data set” (76,056), ”Data-sets” (137,134), ”Data collection” (30,279), ”repository” (13,321), and ”data warehouse” (5,304). These terms are highly relevant for data storage and management in data monetisation contexts. Less relevant terms like ”data record” (845), ”data catalogue” (267), and ”data banks” (100) were excluded due to their low frequency and limited connection to the primary focus.
Value	Key terms related to the financial valuation of data were retained: ”Value” (303,174), ”Cost” (201,684), ”Worth” (5,231), ”Valuation” (1,280), ”Market Price” (12,218), ”Commercial Value” (5,679), and ”monetization” (198). ”Price” (26,816) was excluded as it often led to unrelated topics, detracting from metrics and KPIs related to data monetisation.
Stakeholder	Stakeholder terms related to infrastructure and platforms were included to reflect the environments where data is managed: ”data centre” (916), ”data platform” (1,904), ”data lakes” (152), ”data warehouse” (5,304), and ”Information System” (45,172). These terms capture the role of technology in supporting data monetisation efforts.

Funded by
the European Union

Table 2 Relevant terms and values found for dataset, value, and stakeholder themes in the context of data monetisation.

For the Dataset theme, we retained all major keywords except for ”data record,” ”data catalogue,” and ”data banks,” as these terms were less frequently mentioned and did not strongly align with the primary focus on data monetisation. The most significant terms, like ”Dataset,” ”Data set,” ”Data-sets,” ”Data collection,” ”repository,” and ”data warehouse,” showed high frequencies (e.g., ”Dataset” appeared 328,996 times), indicating their strong relevance in the context of data assets and storage management.

This selective approach ensured that the retained keywords were highly relevant to data monetisation metrics and KPIs, avoiding unrelated topics and improving the accuracy of initial query results.

By combining the results from the PICOC approach, and specific keywords derived from systematic surveys on data value (i.e. [12]) the different keywords for the initial queries were build. The keywords families used in this work are described in Table 3.

Root	Keywords
Context-Value	’Data value’ OR ’Data valuation’ OR ’infonomics’ OR ’data asset’ OR ’value of data’ OR ’data governance’ OR ’information value’ OR ’value of information’ OR ’information valuation’ OR ’knowledge asset’ OR ’Monetization’
Context - Data	dataset OR ”data set” OR ”data-set” OR ”Data value” OR ”data valuation” OR ”data governance” OR ’information value’ OR ’value of information’ OR ’information valuation’
Intervention	’model’ OR ’framework’ OR ’system’ OR ’theory’ OR ’dimension’ OR ’metric’ OR ”KPI” OR ”key performance indicator”
Population	”data centre” OR ”data platform” OR ”data lakes” OR ”data warehouse” OR ”information system”

Table 3 Keywords

3.4 Bibliometric Analysis

Table 4 presents the findings before the screening phase, structured on publications per year when performing a broad analysis. These Findings were obtained through the combination of queries on ‘Context’, ‘Context-data’, ‘Intervention’, and ‘Population’. The same queries were performed for each search engine:

(’Data value’ OR ’Data valuation’ OR ’Infonomics’ OR ’Data asset’ OR ’Value of data’ OR ’Data governance’ OR ’Information value’ OR ’Value of information’ OR ’Information valuation’ OR ’Knowledge asset’ OR ’Data Monetisation’)

AND

(’Assess’ OR ’Measure’ OR ’Evaluation’ OR ’Estimate’) AND (’Model’ OR ’Framework’ OR ’System’ OR ’Theory’ OR ’Dimension’ OR ’Metric’ OR ”KPI” OR ”key performance indicator”)

AND

(”Data Centre” OR ”Data platform” OR ”Data lakes” OR ”Data warehouse” OR ”Information system”)

AND

(”Data centre” OR ”Data platform” OR ”Data lakes” OR ”Data warehouse” OR ”Information system”).

For the IEEE and ACM, an additional search was performed to extract further information, given the high numbers or low numbers of manuscripts obtained with the structured search query. This was done by performing pairwise search, making sure to combine each component of the Table 3 and avoiding direct linkage to AI and Machine Learning (e.g., ”data monetisation” AND ”KPI” AND (”monetisation” OR ”data value” OR ”data price” OR ”data valuation” OR ”valuation of data”) NOT AI NOT ”machine learning”). AI was not included in the pairwise search since it showed up in too many manuscripts given the current trends of it.

According to the results table (Table 4), the last two columns indicate: the number of publications (considering the filtered years), labelled ‘Total’; as well as the percentage of queries containing terms related to the data valuation domain (i.e. obtained by using only the context search terms from the previous table) labelled ‘Perc’.Table 4 Search in IEEE, Scopus, and ACM databases between 2014 and 2025.

Engine	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024	Total	Perc.
IEEE	60	60	65	78	82	86	99	90	118	100	88	926	0.17%
Scopus	0	0	1	0	3	1	1	3	1	0	2	12	1.48%
ACM	1317	1360	1389	1652	1784	1629	1807	1950	1937	2199	2712	19736	5.85%

Table 4 Queries results for the different search engines

As observed in the table, the ACM engine shows a steady increase in publication activity. A consistent upward trend indicates that interest in how to measure the value of information is growing. Conversely, the Scopus and IEE databases exhibited low and inconsistent search volumes, which could describe preferences for authors on ACM Journals, given the high numbers of publications within the scope (e.g. a total of 336,888 publications within ACM in the domain, as estimated using the Context-Value keywords components).

3.5 Screening Papers

A systematic analysis was conducted on the collected publications with a primary focus on assessing the relevance of the titles to the key facets identified previously. In cases where uncertainty arose, a more detailed analysis was performed based on the information provided in the abstracts. This evaluation process followed a structured pipeline approach, as illustrated in Figure 11. The rigorous screening ensured that each manuscript included at least topics connected to data valuation and monetisation or topics related to metrics and KPIs. Manuscripts that were not directly connected to data valorisation and monetisation could also be considered if metrics and KPIs could be relevant to the data environment. For the high number of text from the ACM, only the first 200 of each year, ordered by relevance (i.e. appearance based on keywords), were evaluated.

Funded by the European Union

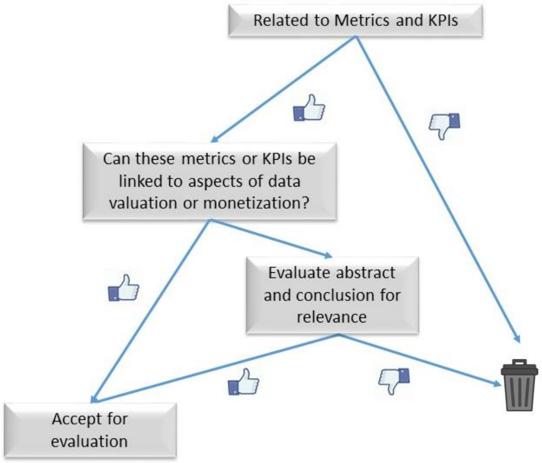

Related to Metrics and KPIs
Can these metrics or KPIs be linked to aspects of data valuation or monetization?
Evaluate abstract and conclusion for relevance
Accept for evaluation

Figure 11 Screening Process Followed to Select Manuscripts Based on Keywords

The screening process does not necessarily define if the manuscript describes an important contribution to answering our research questions. Hence, only those papers that could be linked to the facets were included in the analysis. In the appendix section, there is a table (first table) providing more details on what papers were included and the connection to the different taxonomy clusters.

3.6 Results

This section shares the findings of our systematic literature review, which was carefully conducted through a rigorous screening process. We provide a detailed look at the analysis, including the sources we used to gather the research questions and findings. Each of the following subsections in RQ1 answers is linked to the main components of the taxonomy. As a reminder, the appendix section includes the full results of the KPI links to the BSC Components and a table listing all the metrics and KPIs, which include estimation methods and citations for further reader interest.

3.6.1 RQ1 - What are the main trends observed for KPIs and Metrics within data monetisation?

3.6.1.1 Data Valuation Techniques

Data valuation methodologies provide structured approaches to assess and optimise the value of data assets. Different methods exist in the literature, for example:

Decision-Based Valuation (DBV) evaluates the intrinsic and extrinsic value of data using asset management principles, focusing on decision nodes and their impact on strategic goals [111]. The description of how to estimate a Node Value (NV) (defined as a specific point within an organisation's decision-making process where information is utilised to make decisions) is given in the appendix section. As observed in Figure 2 the NV depends on quality (and possibly cost) and can be used to estimate the Return of Investment (ROI) [111]. Furthermore, [111] motions a technique for valuation based on the type of data. Each type of data is evaluated based on its role and contribution to decision-making and classified into A, B, C, and D categories. Type A includes operational data, while Type B encompasses one-time decision data. Type C focuses on legal and safety data that do not necessarily yield value to an organisation, and Type D covers research and innovation data. The other two types of data (H and L) are related to the retention capabilities of the data value as it is processed into information. Based on this classification, each type is valued based on its specific utility and impact. Also, special emphasis is placed on the distinction between data and information, which refers to the comprehensible output based on data that informs decisions, strategies, and actions.

Capability Maturity Model (CMM) could indirectly be used as a valuation technique. A CMM enhances data value through a maturity-based framework that measures and improves data processes along dimensions like cost, quality, and utility [15, 37]. The purpose of CMMs is to identify how an organisation manages its processes to facilitate its goals (i.e. strategies), providing a qualitative approach that seeks to identify areas for improvement progressively.

As observed in Figure 1, a link with maturity models is created in general with valuation techniques and with governance (as would be specified later), providing an approach to connect metrics and Maturity Models.

Geo-Distributed Data Market Valuation is a useful valuation method for distributed systems which minimise costs by optimising data purchase and strategic equipment placement across multiple locations, balancing bandwidth and latency efficiency [97]. This type of evaluation can be seen as a hierarchically higher levels of infonomics, since the latter focuses only on the data, and not on aspects that facilitate its management. In fact, infonomics, as a general approach, recognises data as a tangible economic asset, advocating its integration into financial and strategic planning to maximise its economic potential [39].

Systematic Approach to Quantify Data as an Intangible Asset, presented in [41], offers a data valuation framework that evaluates datasets across key dimensions, including ownership, cost, usage (e.g., mission criticality, diminishing value), age, privacy, data quality, and volume & variety. Each dimension is assessed through weighted metrics (i.e. the same as merging KPIs), enabling organisations to compare datasets, evaluate the integration of new data, and optimise operational decisions. This framework, amongst others, contains KPIs and metrics linked and addresses cost-efficiency, operational utility, and economic value, offering a comprehensive toolkit to understand or relate data's economic potential.

Payment–Accuracy Trade-off described in [127] is another framework focusing on balancing the cost of incentivising individuals to share private data with the level of accuracy required for effective decision-making. It models the relationship between payments made to individuals, their chosen privacy levels, and the resulting data quality, which directly impacts the accuracy of the learning process. By deriving bounds on payments, the framework ensures cost-effective data collection strategies while meeting predefined accuracy targets. This trade-off (called in this document directly as the Payment-Accuracy Trade-off) is critical in environments where improving data accuracy (reducing error) requires higher payments to compensate for reduced privacy. The framework enables organisations to design incentive mechanisms that optimise the balance between financial cost and analytical precision, especially under privacy-preserving constraints.

After performing the analysis across the provided manuscripts, we have classified the KPIs and Metrics within this group into four key topics: *Cost-Based*, *Contribution*, *Quality and Utility-Based*, and *Risk and Loss*. Each category offers a structured approach to data valuation, with specific methods and corresponding metrics or KPIs to capture data's multi-dimensional economic and operational value. These metrics can be linked depending on the framework employed. For example, the Data Valuation framework interlinks contribution-based and cost-based metrics and KPIs for estimation.

Data Value can be seen as the main focus on evaluating the data characteristics. The metric or approaches to estimate it can be considered the ultimate objective of all the previous approaches. In fact, as observed in the literature, its name varies, but with the same intrinsic characterisation - assigns value to data based on its ability to inform critical decisions.

As observed in Figure 2, Data Value (DV, also referred to as Data Criticality, Value of Information, Fixed Record Value, IP Value, Intrinsic Record Value, Application Characteristic Index, or Value of Information [12, 15, 74, 111]) is consistently included as an output within this category. An asterisk has been used to emphasise that these values correspond to estimates derived from specific methodologies and strategies. They do not represent precise or absolute information but rather are context-dependent assessments.

Regarding DV, in [111], Data Value is examined through the **Decision-Based Valuation (DBV)** framework, which conceptualises data as a measurable asset that supports decision-making and organisational efficiency. Similarly, [15] places data value at the forefront, focusing on its management and monitoring within data value chains. Metrics such as cost, usage, and intrinsic characteristics are explicitly discussed. The concept of Intrinsic Record Value is incorporated within a capability maturity model, providing a detailed view of how data attributes contribute to valuation. In [45], the data value is implicitly addressed through metrics like Age of Information (AoI), linking data freshness to its utility in real-time applications. Furthermore, the authors expand the use of AoI by a joint data sampling and pricing optimisation framework to maximise platform profits. This explicitly considers the relationship between data freshness, user demand, and price sensitivity.

The link between these metrics and others related to quality and utility is represented by arrow as seen in Figure 2. DV is also explicitly highlighted, showcasing its relevance in dynamic and time-sensitive scenarios. For example, [81] explores the DV based on location using a Wasserstein Distance for Data Quality. [132] explores data value in the context of machine learning and mobile health, employing metrics like entropy to quantify data contributions. In [56], the data value is explicitly linked to its role in prioritising and managing IoT data, complemented by discussions on DV as a metric for assessing data freshness and relevance. Document [61] emphasises data value within security contexts, tying it to data quality metrics. In [5], Data Value is portrayed as integral to decision-making and resource optimisation in IoT systems, with DV explicitly mentioned for assessing data usefulness and prioritisation. Finally, [12] provides an extensive discussion on DV models, metrics, and applications, identifying DV as a central focus and including mentions of DV within data valuation frameworks. The methodologies to calculate DV are diverse, and only the DBV is briefly presented in the appendix. Nevertheless, readers are encouraged to visit the references for further information on the DV estimation through models.

3.6.1.1.1 *Cost-Based*

These metrics and KPIs related to valuation methods focus on data storage, access, and maintenance costs. This approach is foundational for organisations in data-intensive sectors, such as cloud computing and telecommunications, where cost-efficiency plays a critical role in data management strategies. These are the metrics within this category:

Access Cost (linked to Bandwidth Cost) assesses the financial impact of retrieving/accessing data, particularly relevant in environments with high data transfer volumes, such as geo-distributed analytics in cloud structures [97, 105].

Storage Cost (It can be referred to as part of operational costs) captures the long-term expenses of storing vast datasets, making it a pivotal metric for cost-sensitive sectors that continuously manage large-scale data, such as telecommunications [97, 111].

Reconstruction Cost evaluates the financial and computational requirements to restore lost or corrupted data. It emphasises its importance in systems where data loss could disrupt essential operations.

Packet Recovery Score measures the effectiveness of recovering lost or corrupted packets in data transmission, underscoring the critical need for resilient communication systems in maintaining data integrity [56, 105], given the similarity Reconstruction Cost and Packet Recovery Score metrics were merged as Reconstruction Cost but they can be divided if needed by the users.

Replacement Cost is a metric often associated with the financial implications of unusable data, measures the potential cost of replacing critical data assets. This metric underscores the strategic significance of robust data management practices, particularly in sectors like healthcare and finance. In [39], the importance of managing data as an economic asset is emphasised, illustrating how data loss can disrupt market competitiveness. Similarly, [23] highlights use cases where timely access to accurate and reliable data mitigates risks and maximises operational efficiency in high-stake industries. Additionally, [12] discusses Replacement Cost in combination with dimensions such as **Timeliness**, **Utility**, **Legislative Risk**, and **Competitive Advantage**, presenting a holistic framework for assessing data value in organisational contexts.

These cost-based metrics enable organisations to evaluate the financial impact of data management practices and make informed investments in data storage, accessibility, and recovery solutions [16]. Given the financial nature of these cost-based metrics and KPIs used to evaluate data value, they can directly or indirectly be linked to the Data Monetisation category for metrics and KPIs. (Which is represented by the horizontal connection between both categories in Figure 2 and Figure 3).

3.6.1.1.2 Contributional Valuation Models

Contribution-based Valuation Models emphasise data’s market value by focusing on its revenue-generating potential and economic impact through the synergetic or contribution considerations of its components. For example:

Shapley Value metric - A method derived from cooperative game theory. This method quantifies data's individual contribution to a larger dataset, ensuring fair compensation for stakeholders in shared or collaborative environments [8, 12, 40, 74, 130, 132, 133]. In spatio-temporal data marketplaces, the Shapley value helps evaluate how each data source uniquely improves forecast accuracy, enabling fair revenue-sharing amongst contributors. For example, in vehicle-hiring services, combining datasets can improve demand predictions and travel time estimates, with the Shapley value assigning value proportionate to each dataset's contribution. Extensions like S-Shapley address challenges in shared datasets, ensuring fairness in data valuation even with structural constraints [130]. Importantly, the Shapley value also supports privacy-preserving systems by integrating Differential Privacy into data valuation, crucial for sensitive applications like health data marketplaces [74]. Overall, the Shapley value could enhance trust, ensure transparent revenue-sharing, and align data pricing with utility, making it a key tool in data monetisation.

Open Data Barometer (ODB) further broadens the view of data valorisation by evaluating the societal impact of data within the public sector. It captures data's contribution to transparency, accessibility, and socio-economic impact [4]. The Open Data Barometer (ODB) is a global metric that measures how governments publish and use open data to promote accountability, innovation, and social impact. Developed by the World Wide Web Foundation, the ODB assesses the maturity of open data initiatives by evaluating the Openness, Accessibility, and Usability of government data. The ODB is divided into three main sub-indexes: *Readiness*, which includes factors such as the legal and policy environment, data infrastructure, and engagement with stakeholders; *Implementation*, which looks at the availability of machine-readable data, the types of datasets published, the level of open licensing, and whether the data is up-to-date and accessible in formats that are useful for developers and the public; and *Impact* which measures the real-world effects of open data, such as uses in policy-making, business innovation, and civic engagement. Information on how to estimate the parameter is included in the appendix section. Finally, although the subcomponents of ODB could be defined and scored independently, they were not included as individual metrics or KPIs in our study, nonetheless, for understanding, they are included within Figure 2.

Leave-one-out (LOO) is a metric which, as defined in [8], can be used to estimate the value of a data source by measuring the difference in performance of a model/system when the data source is excluded from the training set. As a result, this definition does not exclude the use of information from a single source, and can therefore be applied as a valuation technique to a variety of different features and data sources if needed.

Location Yardstick Score (LYS), defined in [57], is a metric introduced in the document to quantify the value of a single point of location data based on its contribution to improving the performance of a specific analytical task. The calculation of LYS is context-aware, taking into account factors such as the amount of historical location data already available and the specific task being analysed.

Data Value Ratio (VR), described in manuscript [111], is a key metric within the Decision-Based Valuation (DBV) framework. It quantifies the relative contribution of each data source to a decision node’s overall value. Expressed as a fraction between 0 and 1, VR determines what portion of the node’s value is attributable to a specific data input. By applying VR, organisations can identify which data sources are most critical to their decisions, optimise data collection processes, and ensure that resources are allocated in proportion to the value they generate.

Together, these market-oriented models allow organisations to identify and maximise the revenue and competitive benefits that data assets can offer [37].

3.6.1.1.3 Quality and Utility-Based Valuation Models

These models are designed to assess data’s value based on its quality, reliability, and alignment with specific application requirements. They are essential in high-stakes environments, such as predictive analytics, where data accuracy and quality directly influence outcomes. Key metrics within this category include:

Current Quality, Anticipated Quality, and Required Quality help to measure the level of data that meets the standards needed for effective decision making [10, 12]. These metrics align data’s quality with its projected utility, supporting valuation techniques that prioritise reliability in data-driven applications.

Ease of Measurement, defined in [16] as a metric, refers to the degree to which a particular data value dimension can be quantified or assessed effectively using established methods, tools, or frameworks. It serves as an indicator of how practical or straightforward it is to evaluate a data value dimension, considering the availability of standardised metrics, models, or processes.

Value-Added, Diminishing Value, and Value degradation metrics reflect the data relevance and economic utility over time. While Value-Added emphasises the contribution of data to decision-making or operational efficiency, Diminishing Value highlights the temporal decline in data utility, particularly critical in sectors like IoT and real-time monitoring systems, where up-to-date information is essential to maintain a competitive advantage [11, 17, 41]. Value degradation (coined here but referred to as value attenuation and related to data entropy in [91]) extends this understanding by quantifying the temporal loss of data value as uncertainty and irrelevance increase over time. Inspired by the thermodynamic concept of entropy, Data Entropy measures the degree to which data becomes less useful, accurate, or actionable as it ages.

Value of Information for Business (VIB) – Even though we have already defined and described the Data Value or Value of Information as a metric, VIB is set apart given its specificity. VIB, as described in [39], explicitly connects data quality attributes (accuracy, completeness) and their relevance to how useful and actionable the data is in real-world business processes. It helps determine whether the data enables efficient and effective decision-making.

Utility within the context of data valuation refers to the effectiveness of a resource (e.g., datasets or systems) in achieving specific organisational objectives. Utility can be confused with Usability (a data quality metric defined later on). The latter focuses on ease of interaction and other user-friendly aspects. Utility, on the other hand, evaluates the functional contribution of the resource to desired outcomes.

Utility is inherently outcome-focused, quantifying the degree to which a resource supports operational goals, such as decision-making, cost optimisation, or performance improvement. It is context-specific, meaning that its evaluation depends on the relevance and impact of the resource within a defined operational or organisational scenario. Utility can be expressed quantitatively, as a KPI, allowing direct comparisons and trade-off analyses [9]. Decision-based models use Utility to calculate data importance within critical decision nodes. Furthermore, Utility is instrumental in optimisation tasks, where resources are allocated to maximise value derived from high-utility datasets or processes.

Relevance (also referred to as Decision Support Capabilities, Relevance Factor, Priority Score, Importance, Existence) [5, 11, 12, 17, 20, 25, 39, 50, 55, 61, 75, 113, 118, 136] can be seen as the usefulness of data for one or more processes. Even though it is highly linked to Utility, the difference lies in the broadness of applicability. In simple terms, Utility looks at the broad intrinsic potential of the data, while Relevance depends on the specific user, use case, or context. A dataset can have high utility (applicable in many scenarios) but low relevance for a particular user if it doesn’t meet their specific needs. Conversely, highly relevant data for a niche use case may have limited utility beyond that application. This consideration defines the possible formulation of utility as a function of its Relevance in different domains (as treated in this work). For more information on how to estimate Utility and Relevance, see the appendix section.

Even though some could argue that all of the metrics merged here have some key differences, the overall concept is generally the same. For instance, in [5], the metric Importance is defined as a measure of how critical or valuable a piece of information is for a specific application. In [25], Existence is a metric that checks whether essential concepts, ideas, or entities are represented within information assets. It ensures that the dataset includes all necessary components or key ideas required for it to be complete and relevant for its intended purpose.

A more complex but similar consideration is for the Decision Support Capabilities (coined in this work based on the approach in [111] and as a pragmatic aspect defined in [75]). This metric can be understood as the ability of the asset to deliver actionable insights that directly influence decision-making processes. Given the closeness to a node contribution, as set in [111], the metric was extended to the specific contribution of a Node, data set, or data cluster, specifying the unique purpose of it, similar to how an electric motor has a specific function in a conveyor system. This attribute clarifies the node's role, removing ambiguity by providing context for its other parameters. A clear and precise description of the support capabilities ensures that the data function is well-defined and reduces the need for interpretation.

3.6.1.1.4 Risk and Loss Valuation Models

These models focus on assessing the value of the data through the possible economic risks associated with data loss or mismanagement. These models are particularly significant in sectors where data continuity and security are critical, such as finance, healthcare, and critical infrastructures.

Loss of Information Value (LIV) and Replacement Costs are key metrics in this category, estimating the financial impacts of data loss or compromise [39, 23]. In environments where data access is essential for maintaining operational stability, these metrics support proactive data governance strategies to minimise risks associated with data loss. LIV estimation is directly related to economic concepts; as defined in [39], it is composed of the cost required to acquire or replace the information plus the cumulative loss of income incurred over a period (t) if the data is lost. Important, even though not specified in the text, liability and other compliance costs could be incurred.

Missed Opportunity Cost and Root Cause Remediation address the economic impacts of unused or inaccessible data. These concepts are especially critical in IoT and data-driven sectors where real-time data utilisation significantly affects revenue and operational efficiency. The document [11] discusses opportunity costs as part of poor data quality's economic consequences, highlighting lost or missed revenues due to data inaccessibility or errors. Similarly, [129] introduces Root Cause Remediation as a metric to identify and mitigate the origins of data errors, which is integral to ensuring effective data utilisation.

Payment–Accuracy Trade-off refers to balancing the costs of acquiring data with the accuracy required to achieve a desired level of decision-making efficacy. It allows organisations to weigh the value of improved data quality against additional expenses incurred. This trade-off is especially relevant in settings like predictive analytics and decision-making frameworks, where

Funded by
the European Union

achieving high precision often necessitates higher investments in data acquisition mechanisms [127]. The framework allows organisations to evaluate whether the benefit of higher data quality (and thus improved decision-making effectiveness) justifies the additional expense incurred in data collection.

Processing Value Ratio (VP) is described as a measure used to determine the value added through processing data into information [111].

Rival access cost is a metric associated with data valuation models that include dimensions such as utility, content uniqueness, and financial implications [12]. It is used in the assessment of data value to estimate the impact of restricted or competing access to the data, thereby affecting its perceived or actual value (i.e. putting a direct differentiation with open data).

In general, Risk and Loss Valuation Methods provide a comprehensive framework for organisations to assess the economic consequences of data unavailability or compromise, fostering proactive risk management and data protection strategies in high-stakes sectors. Furthermore, as defined before, some metrics depend on the data type (described in [111] and explicitly mentioned in the appendix section).

3.6.1.2 Data Monetisation

After analysing the manuscripts on data monetisation KPIs and metrics, we have defined four clusters related to methods, and their linked metrics and KPIs, within the concepts of data monetisation: Dynamic Pricing and Market, Financial, Privacy and Compliance, and Data Quality and Utility.

Data prices, as a general metric, can be calculated using various methodologies tailored to specific contexts and applications (e.g. [2, 71, 81, 91]). For example, entropy-based approaches provide a way to estimate the inherent value of data by measuring the reduction in uncertainty that the data contributes to a given model [71]. Elastic pricing models, as proposed in [2, 91], adapt prices dynamically based on supply, demand, and utility. These models ensure that pricing remains responsive to changing market conditions, providing a flexible framework for data valuation in competitive environments. Additionally, practical frameworks for data monetisation, like the one discussed in [81], integrate these theoretical methodologies into operational systems. These frameworks not only calculate data prices but also facilitate the incorporation of implementation challenges, such as user incentives, privacy, and computational efficiency. This ensures applicability in real-world scenarios. In relation to Figure 2 and Figure 3, these estimations are general outputs of data monetisation strategies and, as shuch, their values are described as approximation of their real counterparts.

3.6.1.2.1 Dynamic Pricing and Market

Dynamic Pricing and Market focuses on assigning value to data assets through real-time pricing models and market value indicators. Dynamic pricing mechanisms are foundational, as they allow organisations to respond to shifting market demands by adjusting data prices based on factors like usage patterns and time sensitivity. Elastic pricing models, such as those described in [2, 91], complement dynamic considerations by providing adaptable structures that enable data holders to refine prices based on evolving market conditions, moving from static to value-based pricing to maximise financial gains.

Demand, as a metric, is a dynamic measure that looks for the interest and value of specific data categories in the information market. It combines buyer preferences (scores or ratings), usage metrics (e.g., views or downloads), and overall market interest, reflecting both the quantitative and qualitative desirability of the data. This metric evolves with changing buyer needs and data utility, making it essential for pricing and value assessment.

Demand is measured by consume or score indexes. In data monetisation, it reflects the interest and willingness of buyers to acquire data. Buyer ratings or scores serve as proxies for demand, influencing pricing by indicating preferences for data categories [94]. Demand is also quantified by the number of data consumers, measured through metrics like dataset views, downloads, and usage frequency [15]. But this trend depends on the type of data (as previously discussed in data valuation techniques). High demand aligns with datasets offering significant economic or strategic value, particularly those critical for business operations or decision-making. Contextual relevance further enhances demand by aligning data with specific organisational needs [15, 94].

Market Value of Information (MVI), and its understanding as a metric, is explored through frameworks that assess the competitive impact of data within organisations and marketplaces. For example, the manuscript on ”Infonomics” discusses the measurable economic value of information assets and emphasises the need for businesses to leverage information for creating new products and services, enhancing competitiveness, and addressing market demands [39]. Additionally, manuscripts focusing on data value chains and management highlight the critical role of data valuation in strategic decision-making, with metrics such as operational impact and competitive advantage providing a basis for understanding pricing and market relevance [15, 16, 37]. These studies collectively underscore the influence of data scarcity and market dynamics on shaping pricing strategies and ensuring financial sustainability.

Market adjustment Factor, in the context of data trading markets, as discussed in [45] and [131], plays a critical role in pricing mechanisms, particularly in auction-based transaction models. The factor dynamically adjusts the market-assessed price of a data component to reflect its perceived value in the market (see appendix section for more information). Additionally, as highlighted in [45], data valuation is sensitive to parameters such as freshness, measured by the Age of Information (AoI), (and other quality metrics). The AoI represents the time that has elapsed since the data was last updated and directly affects user demand and pricing. Platforms often use pricing policies that incorporate AoI, such as uniform, dual, or dynamic pricing strategies, to maximise profits. These strategies account for how users’ valuations decrease with increasing AoI and how data sampling policies impact both data freshness and availability. Finally, entropy can impact dynamic components, even if it is considered a quality metric. Entropy in the context of data pricing is related to a dynamic because it represents the degradation or uncertainty of data value over time. This connection arises from the nature of information and markets.

Scarcity refers to the likelihood that specific data is not accessible to other organisations, making it unique and increasing its value due to limited availability [12, 25]. Scarcity is a dynamic metric influenced by technological advancements and changes in data availability. Over time, initially exclusive datasets may lose their scarcity as broader collection and sharing capabilities develop [25]. This dynamic aspect underscores its temporal variability, particularly in fast-evolving digital ecosystems [12]. Scarcity is also market-dependent, shaped by supply-demand dynamics and contextual relevance. Rare and highly demanded data within one domain may not hold the same value in another domain where similar data is abundant [124, 131]. Economically, scarcity significantly affects pricing mechanisms, particularly in auction-based markets, where limited availability contributes to higher data value [131]. Strategically, scarce data offers a competitive advantage, enabling organisations to make exclusive decisions while limiting competitors' access to critical resources [25]. However, perceptions of scarcity can be subjective and vary across regions or contexts, further emphasising its dependency on specific market conditions [12, 124].

3.6.1.2.2 Financial

Financial metrics and KPIs encapsulate the financial foundation necessary for data monetisation by examining both capital and operational expenditures, alongside direct financial returns.

Capital Expenditures (CAPEX), also referred to as Capital Costs or directly as Cost, with a context connotation, together with Operational Expenditures (OPEX) are key metrics in this category, covering initial and recurring expenses tied to data infrastructure and operational maintenance. Importantly, we have separated the expenditures within OPEX and CAPEX, nevertheless they can be context dependent (e.g. Data Acquisition Cost). As observed in Figure 3, there are several cost metrics which are not detailed here but more information is provided in the appendix section.

CAPEX is central to building estimates of the infrastructure needed to transform data into valuable assets. CAPEX in data centres includes cloud computing resources, and storage systems. These are essential for managing data at scale and unlocking its economic potential. The lifecycle costs associated with such investments, including depreciation of hardware and software, are critical for assessing their long-term value and ensuring sustainable data monetisation efforts.

CAPEX plays a pivotal role in setting up essential infrastructure cost estimations for data centres, cloud services, and storage systems. These Investments in infrastructure are foundational for processing and managing vast amounts of data [3, 15, 131]. Furthermore, funding for machine learning tools and analytics platforms is critical. These platforms enable organisations to derive

Funded by
the European Union

actionable insights and monetise data through applications like customer segmentation, predictive analysis, and targeted marketing [50, 131]. CAPEX supports the creation of APIs, data pipelines, and other infrastructure necessary for offering Data-as-a-Service (DaaS), allowing businesses to unlock revenue streams by selling curated datasets and analytical services [131]. As data volumes increase, scalability becomes essential. Investments in upgradeable infrastructure ensure operational efficiency and the capacity to handle growing datasets [3]. Beyond physical investments, CAPEX also encompasses data licenses and intellectual property, and governance considerations, which are critical for competitive advantages and data-driven strategies [12, 111]. Considering hardware and software depreciation is essential for evaluating long-term value and ensuring sustainability in monetisation efforts [3].

Complementing CAPEX, OPEX (next defined metric) plays a critical role in the ongoing operationalisation and management of data infrastructure. Operational expenses include maintaining servers, cloud subscriptions, and software updates. In addition, operation expenses include the costs associated with ensuring data quality, security, and compliance. These recurring costs are crucial for enabling scalable and reliable data pipelines, as highlighted in [41].

Operational Expenditure (**OPEX**; also referred to as Operating Expenditure, Maintenance Cost, Storage Cost, Contractual Costs, Labour Costs, Publishing Cost, Cost, Application Cost, Service Cost) refers to the recurring costs incurred by organisations to sustain and manage their operational activities, particularly in data systems and services. OPEX involves a variety of cost components tied to daily operations. For example, Maintenance Cost refers to the recurring costs for maintaining systems and operations, such as the upkeep of data pipelines and optimisation processes to prevent over redundancy [12, 105]. These costs include regular updates, adjustments, and system optimisation. Storage Cost covers the operational expenses (such as energy, cooling, etc.) associated with storing data, which may involve cloud subscriptions or physical storage requirements. Managing large datasets effectively can reduce costs through redundancy optimisation strategies [15, 81, 105, 111]. Contractual Costs include operational expenses stemming from agreements with vendors or service providers. For instance, in geo-distributed systems, costs may arise from contracts tied to data access or placement agreements

[96]. Transaction Fees represent a critical operational cost within OPEX. These fees occur during interactions or data exchanges facilitated by block chain or similar technologies [6]. Application Cost, such as machine learning platforms and APIs, entails costs for computational resources, software updates, and third-party platform fees. These costs are essential for maintaining ongoing analytics and data services [12, 81].

These expenditures are essential for maintaining operational efficiency and scalability in data management systems. Properly addressing these costs allows organisations to optimise resource utilisation and adapt to growing data demands. This is highlighted in the need for scalable infrastructure and redundancy reduction [15, 105].

Return of Investment - The literautre stress that differentiating between various costs is crucial, as it provides a clear understanding of the total expenditure associated with data-driven projects. By evaluating these costs relative to financial returns, companies can better determine their Return on Investment (ROI), a critical KPI for assessing the profitability of data initiatives. ROI serves as a measurable gauge of financial returns from data-driven projects, enabling stakeholders to understand the success of their investments in data infrastructure. For instance, [3] highlights ROI as a key metric for quantifying financial returns from information systems investments, emphasising its role in strategic decision-making and profitability analysis. Additionally, [132] and [133] explore how efficient data valuation and pricing mechanisms contribute to maximising profits, thus enhancing the ROI of data-driven initiatives.

Net Present Value (NPV), Budget, Revenue, and Internal Rate of Return (IRR) metrics - Net Present Value (NPV, [3, 111]), Budget [42], Revenue (or Economic Benefit) [2, 42, 50], and Internal Rate of Return (IRR, [3]) are financial metrics that help to estimate the future revenue flows generated by data assets, providing a forward-looking perspective on data's economic potential over time, which is especially relevant in predictive analytics and strategic decision-making contexts.

Funded by
the European Union

3.6.1.2.3 Privacy and Compliance

Privacy and Compliance recognise privacy as an increasingly monetisable attribute, as consumers and regulators alike demand higher data protection levels.

Value of Privacy and Privacy Costs, and other privacy-related metrics have gained prominence as companies identify the economic potential in offering privacy-compliant data. Manuscripts [61, 115, 127] approach privacy from a strategic standpoint, highlighting that privacy valuation can directly affect data pricing and consumer engagement. These texts tend to depend on a metric, such as Privacy Level [115, 127] that reflects the amount of privacy loss incurred when sharing data. It determines how much an individual's private data can influence the reported data. By attaching financial value to privacy, organisations can meet regulatory standards and also capitalise on market trends, as protected data commands premium pricing in sectors sensitive to privacy. Privacy-adjusted valuation metrics are particularly relevant in industries with stringent regulations, like healthcare and finance, where compliance ensures data utility without compromising consumer trust or market value [94]. Additionally, manuscripts such as [61] provide insights into how privacy considerations can be integrated into overall monetisation strategies, emphasising the balance between consumer trust and profitability. As these metrics align with evolving regulations, companies can balance data utility with privacy protections, enhancing competitive advantage through consumer trust and regulatory compliance.

Protection Expense and Risk Cost metrics (also named regulatory risk index) [16,69] help to account for the costs associated with the monetary and reputational impact that arises from loss, compromise, or misuse of data. It encompasses tangible expenses, such as regulatory fines for non-compliance and the costs of replacing or reproducing lost data, as well as intangible effects like reputational damage and competitive disadvantages if sensitive business information is accessed by competitors. These metrics, and others, are described in appendix section.

3.6.1.2.4 Data Quality and Utility

Data Quality and Utility metrics emphasise the importance of maintaining data's value, by ensuring conditions, processes, and evalautions for its usefulness. These metrics are sometimes directly linked to modelling strategies that define prices of data based on its quality or utility.

Manuscripts [71, 134] explore the use of Information Entropy (quality metric) as a metric to quantify the richness and informational content of data, with applications in data pricing and quality assessment. Document [71] highlights how higher entropy levels correlate with greater data value, leading to more precise pricing strategies in data trading. It introduces pricing models based on entropy. This is where the price of data is mapped to its informational content through functions that ensure fairness and arbitrage-free pricing. These models include entropy-based pricing functions designed for query-based pricing scenarios, emphasising rationality and flexibility in pricing datasets or their subsets. Meanwhile, [134] addresses broader aspects of data quality, emphasising the importance of metrics like accuracy, completeness, and consistency in ensuring the usability and reliability of data in competitive markets. Given the direct linkage of different quality metrics for pricing approaches, they are directly linked to the Quality and Utility cluster of Data Monetisation (see Figure 3).

Weighted Coverage Function [27] is a method used to determine the cost of answering queries in data pricing systems. It assigns a weight to each part of a dataset and calculates the price of a query based on how many of these parts are affected by the query. Essentially, the more a query changes or reveals about the data, the higher its price. This approach ensures fair pricing by preventing situations where someone could combine cheaper queries to obtain more expensive results. It also avoids charging for the same information twice by keeping track of what has already been purchased. Moreover, the method can be tailored to give higher value to certain parts of the data, allowing for flexible pricing that reflects the data's importance. The weighted coverage function can be used as a metric or KPI in data pricing by quantifying the value and impact of queries on a dataset. It measures the proportion of the dataset affected or revealed by a query, providing a clear indication of how much information is being disclosed. As a KPI, it offers insights into how well a pricing model balances revenue generation and customer satisfaction. For instance, it can help evaluate whether high-value parts of a dataset are appropriately priced or if pricing structures inadvertently encourage arbitrage or inefficiency.

Field Value is described as an objective metric representing the financial amount recorded in a specific field of a data record. This can be utilised as a weight to determine the overall value of that record. Even though there is no clear description of how to calculate the Field Value in the manuscripts, a description could be found in [12]. Estimation approaches are suggested in the appendix section. Additionally, according to the framework outlined in [124], the evaluation of digital information assets incorporates Utility as a core dimension of their value. This approach emphasises assessing the data usefulness, considering factors such as financial value, pertinence, and transaction costs. By integrating these dimensions, organisations can balance the data monetisation with compliance and value preservation, even when privacy protections are in place.

Process Failure Costs, as described in [11], are explained as costs incurred when poor-quality data causes a process not to perform properly. For example, inaccurate mailing addresses may result in correspondence being misdelivered. These costs include Rollback Costs, Rework Costs, and Prevention Costs.

Economic Efficiency, also directly related to the Data Business Characteristic Index, represents the optimal use of data resources to maximise economic benefits while minimising costs. It reflects the ability to leverage high-quality data for decision-making, ensuring that business outcomes are enhanced by considering Data Value, Accuracy, and Relevance. In decision-oriented frameworks, Economic Efficiency quantifies the incremental payoff achieved by integrating data quality into the decision-making process [20, 46].

3.6.1.3 Customer Needs & Satisfaction

Within this perspective, we identified two primary areas: Customer Satisfaction and Quality of Service (QoS). These classifications emphasise how data products meet customer needs, enhance engagement, and secure economic value through effective alignment with user expectations and market demands.

3.6.1.3.1 Customer Satisfaction

Satisfaction, also named as to feedback, user's attitude, degree of satisfaction, among other user's focuse degree of satisfaction are pivotal in data monetisation, serving as indicators of how well data products align with customer expectations to drive economic value from data. Satisfaction metrics (also referred to as Feedback, User's Satisfaction, User Attitude, Business User Satisfaction, and Degree of Satisfaction) [15, 23, 35, 37, 44, 52, 77, 83, 84, 106, 137] is essential for evaluating how well a product, service, or system meets user expectations, needs, and preferences. It is recognised as a key determinant of success in information systems, closely linked to usability, effectiveness, and user perceptions. Satisfaction is frequently discussed in the context of its role in improving user engagement and feedback mechanisms, particularly in usability-focused evaluations such as those for open data platforms and disaster management systems [52, 83, 137]. In open government data platforms, satisfaction is emphasised as a tool for enhancing accessibility and usability, providing valuable insights for system refinement [23, 106]. Similarly, studies on logistics information systems and university service evaluation systems incorporate satisfaction as a measure of service quality and operational efficiency, highlighting its role in aligning systems with user needs [52, 137]. Additionally, the ”degree of satisfaction” is explicitly addressed in contexts where satisfaction metrics are tied to specific, measurable outcomes, such as task performance or system usability improvements [35, 118].

Churn [49] refers to the discontinuation or cancellation of a service by a customer, leading to a loss of revenue for the service provider. Furthermore, churn prediction models leverage customer data — such as service usage patterns, engagement levels, and historical behaviour — to identify customers at risk of discontinuing services. These insights enable targeted interventions to reduce churn, retain customers, and preserve revenue streams.

User Frequency, Concurrent Users, and Users Number measure the regularity with which individual users engage with a platform or service within a specific timeframe. This metric reflects individual behaviour patterns, such as daily, weekly, or monthly activity, and is crucial for gauging user engagement and loyalty [12, 28]. High user frequency indicates active and returning users,

Funded by the European Union

which is vital for monetisation models based on subscriptions, premium features, or targeted advertising.

User Frequency, can be tracked by measuring Unique and Returning Users, provides insight into market reach and user loyalty. The number of Unique Users offers a snapshot of the dataset's reach, while high returning (measured as Concurrent User [28]) counts reveal the dataset's sustained value over time. These indicators are crucial for understanding both the breadth and depth of data product engagement. Steady growth in unique users signals successful market expansion, while strong retention rates imply that the dataset continues to meet users' needs effectively [12, 28]. Within the concept of data monetisation, different definitions emerge when referring to frequency. User Frequency (already covered), Data Frequency, System Frequency, and Access Frequency represent distinct yet interconnected metrics that offer valuable insights into platform usage and performance.

Access Frequency refers to the rate or count of interactions with datasets, services, or features over a specific timeframe. Data Frequency and Granularity refer to the rate at which data is generated, updated, accessed, or transmitted within a system. Finally, system frequency, which can be tracked by components such as speed and velocity, ensures that the technical infrastructure can accommodate high user and access activity with minimal performance issues.

3.6.1.3.2 Quality of Service

Quality of Service (QoS) encompasses essential metrics for tracking customer trust; a cornerstone for data monetisation. QoS metrics are dependent on services (including Interface and Platform – see Figure 4) and those that directly affect user satisfaction and the likelihood of customers paying for data services (i.e. Data Quality by itself).

Quality of Service (also referred to as Data Connected to Service Levels, Service Characteristic Index, And Service Level Agreement) as a metric refers to the level of service performance experienced by customers [105], ensuring that it aligns with predefined standards, such as Service Level Agreements (SLAs). [105] highlights the importance of maintaining high QoS to ensure reliable service delivery and compliance with customer expectations.

Funded by
the European Union

Similarly, [129] associates QoS with data quality dimensions, such as timeliness and reliability, demonstrating its relevance in ensuring optimal service delivery. In [20], the concept of a Service Characteristic Index is introduced, a component closely related to QoS. This index evaluates service efficiency through metrics such as Access Frequency (related to Communication Metrics, form the Ingestion Capabilities & Capacity sublcuster, in Figure 4) and Security, prioritising critical data to enhance performance and reliability. Although not explicitly termed QoS, the Service Characteristic Index aligns with QoS objectives by ensuring that essential services and data are handled efficiently to meet user requirements.

In relation to the Interface and Platform, different metrics can be used to track customer engagement and experience. Visualisation, Usability, and platform performance are critical for customer engagement, directly supporting data monetisation by making insights accessible and actionable.

Visualisation (as a metric that can be evaluated by customer satisfaction, checklist-based scores, and expert judgment, and can be tracked as a metric/KPI) simplifies data interpretation, enabling users to derive value from data products more effectively, often justifying a premium price [23, 106].

Learnability, defined as the ease with which users can quickly acquire the skills and knowledge needed to use a system effectively, are supported by features like simplicity, self-descriptiveness, and consistency [83].

Winning Rate, together with Access Frequency [37, 92, 112], helps to indicate that a data provider’s offerings are valued highly, as they are selected more often to provide answers to queries [42]. Winning Rate can be described as the percentage of queries a data provider successfully serves, relative to the queries they are eligible to serve.

Access Frequency refers to the rate or count of interactions with datasets, services, or features over a specific timeframe. It encompasses terms such as Number of Requests, Views, Arrival Rate, and Usage Over Time. This metric reflects resource popularity, usage trends, and system demand, playing a vital role in performance optimization and monetisation strategies like pay-per-use or licensing agreements [92, 20, 112]. High access frequency indicates strong demand, as seen in Open Government Data usage or hierarchical storage optimization in high-speed railway maintenance [92, 20]. It also informs system workload management and user engagement analysis, enabling platforms to prioritize valuable resources [37, 15, 12, 111]. Even though these metrics could be linked to specific server characteristics like latency, traffic ratio, and bandwidth, the relation of the latest one to system performance is set aside.

3.6.1.4 Market Penetration Metrics

Market Penetration Metrics are essential tools for measuring data monetisation success in target markets. Given the relatively low numbers of text covering these types of metrics, there is no further clustering or classification and thus, we report only measurable indicators such as **Downloads**, **Views** (As Users frequency), **Competitive advantage**, and **Reputation**.

Downloads and Views - By measuring how frequently datasets (or parts of them) are accessed or used, organisations can gauge initial user interest and ongoing relevance. High Download and Views counts suggest that a dataset holds considerable appeal and reaches a broad audience. For example, on public sector platforms, datasets such as transportation or environmental statistics demonstrate a high level of engagement, helping organisations prioritise these resources for maximum impact [15, 23, 37, 92].

User Frequency, measured also by tracking Unique and Returning Users, provides insight into market reach and user loyalty. The number of Unique Users offers a snapshot of the dataset’s reach, while high returning (measured as Concurrent User [27]) counts reveal the dataset’s sustained value over time. These indicators are crucial for understanding both the breadth and depth of data product engagement. Steady growth in unique users signals successful market expansion, while strong retention rates imply that the dataset continues to meet users’ needs effectively [27, 11].

Reputation can be defined as the perceived quality, reliability, and trustworthiness of data, impacting decision-making and governance. Key aspects include quality, governance, transparency, and ethical use (see Figure 4) [5, 11, 17, 55, 100, 107, 124].

Funded by
the European Union

Competitive Advantage as a data value dimension refers to the degree to which an organisation’s data provides a unique strategic edge over competitors. According to the referenced work [16], competitive advantage is assessed by examining the impact of competitors gaining access to the same data. If competitors possess the data, the potential consequences range from negligible effects to severe impacts on business operations and market position. Metrics include insights into critical business processes, the ability to replicate operational advantages, and potential for competitors to gain significant leverage or strategic benefits.

3.6.1.5 Data Quality

When an analysis was performed on KPIs and metrics crucial for evaluating data quality in the context of data monetisation, we defined four core dimensions: (1) Fundamental, which are related to context-less metrics and KPIs that provide basic information about the datasets and thus, provide basic information of the data assets; (2) Contextual, which are similar to fundamental metrics with a considerable context-dependent nature or combination of fundamental knowledge; (3) Resolution, which encompasses key data quality dimensions that focus on the temporal accuracy and level of detail within datasets; and (4) Specialised, which provide deeper insights into the structure and content of datasets. These categories encompass both context-independent and context-dependent measures, forming a robust taxonomy for data quality assessment and organisation. It is imperative that we recognise that the names and definitions for similar concepts within the quality dimension is considerable. In fact, as described by [11] only, there are 76 different data quality dimensions which include:

Accessibility, Adaptability, Agreement of Usage, Applicability, Appropriateness, Believability, Clarity, Collection and Capture, Completeness, Completability, Comparability, Comprehensiveness, Conciseness, Concise Representation, Contextual Clarity, Convenience, Correct Interpretation, Correctness, Credibility, Currency, Derivation Integrity, Documentation, Ease of Manipulation, Ease of Operation, Edit and Imputation, Equivalence, Equivalence of Redundant Data, Estimation, Flexibility, Format Precision, Free of Error, Homogeneity, Integration, Interactivity, Interpretability, Maintainability, Metadata Evolution, Minimality, Non-duplication, Null Values, Objectivity, Over-Coverage, Pertinence, Portability, Precision, Presentation Appropriateness, Privacy, Processing, Redundancy, Relevance, Relevancy, Reliability, Reputation, Responsiveness, Rightness, Schema Clarity, Security, Simple Response Variance, Speed, Standardisation, Stewardship, Timeliness, Traceability, Ubiquity, Under-Coverage, Understandability, Unit/Item Non-Response, Use of Storage, Usability, Usefulness, Value Added, and Volatility. A short explanation of these metrics has been included in the appendix section in a separate table.

As seen in this list, different terms have similar meanings with current trends in their use (e.g. Credibility and Trustworthiness), others are not directly related to quality (e.g. Privacy), and others have been aggregated in dissimilar processes (e.g. Format Precision) making the present work more significant in order to reorder the metrics within their respective importance from a business point of view.

3.6.1.5.1 Fundamental

Within the fundamental cluster, several metrics can be used to provide statistics on the data within data assets. These metrics (like Average, Standard Deviation, Minimum, Maximum, etc.) are not discussed within this text, but are included as Statistics within Figure 6.

Age refers to the elapsed time since a data instance was created, updated, or last observed [5, 41, 45]. It serves as a critical metric for assessing the Timeliness and Relevance of various applications. Independently of being classified within the fundamental cluster, there could be different contexts for its estimation: (1) Static, where Age reflects the time elapsed since the generation of an entire dataset or information record. This static measure is used to evaluate the historical value of data. It influences hierarchical storage strategies, where older datasets may be relegated to lower-priority storage tiers due to reduced relevance. (2) Dynamic, where it represents the dataset freshness by calculating the time elapsed since the most recent data update was generated and received by a user or system.

Granularity (also referred to as Abundance and Data Frequency) in data management and analysis is the degree of detail or precision at which data are captured, stored, and analysed (not refreshed). It defines the smallest unit of information available in a dataset and reflects how finely data is decomposed into its constituent elements. Granularity directly influences the utility, interpretability, and application of data across various domains, as higher granularity provides more specific insights but may increase storage, processing, and complexity requirements [75]. Conversely, coarser granularity aggregates data, potentially reducing detail but enhancing efficiency and broader pattern recognition [30]. Optimal granularity is determined by the objectives of the analysis, the nature of the dataset, and the balance between precision and manageability [78].

Granularity is essential for understanding the temporal characteristics of data flow and data quality, especially in applications like real-time analytics, IoT systems, or streaming services. For instance, high data frequency may arise in scenarios like stock market data streams, where updates occur within milliseconds, or in IoT sensor networks that generate continuous streams of telemetry data [23, 30, 75, 78, 120]. Importantly, while granularity could be linked to Velocity [20, 48, 119, 134], the two concepts differ: Velocity captures the speed of data generation and flow, emphasizing timeliness and performance requirements for real-time processing, whereas Data Frequency (or granularity) concerns the rate and temporal detail of data events. Given these characteristics, Velocity is more closely associated with the Technology and Infrastructure cluster rather than Quality (see Figure 8).

Precision refers to the level of detail with which data is captured, measured, and represented. It plays a critical role in ensuring reliable analyses and outcomes [11, 25, 111, 118]. Precision can be defined in three subsets: numerical, consistency and repeatability.

Numerical precision describes the degree of detail in numerical data, often indicated by the number of significant digits or decimal places (e.g. 23.456 is more precise than 23.5, as it captures finer details). In systems that involve repeated measurements, precision also reflects the consistency and repeatability of results under the same conditions, regardless of how close these results are to the true value. Precision is not limited to numerical data, for example, ”Scarlet”, ”Azure”, and ”Emerald” is more precise than broad groups such as ”Red,” ”Blue,” and ”Green”.

Uniqueness (or Redundancy in the case of data services) refers to the degree to which data entries within a dataset are distinct and free from duplicates. It ensures that each record or entity is represented only once, maintaining the integrity and reliability of the dataset. Uniqueness is critical to avoid redundant data, improving data accuracy, and ensuring consistent data-driven decisions across systems [9, 11, 38, 39, 38].

Variety, also referred to as Multifacetedness, refers to the diversity in the types, formats, and sources of data, as well as the range of features or attributes within a dataset. This characteristic encompasses the heterogeneity of big data, including structured, semi-structured, and unstructured data, and highlights the challenges of managing and integrating datasets with varying dimensions and complexity. A higher variety often includes an increased number of features, which can impact both the analytical potential and computational demands of the system [13, 24, 49, 114, 134].

Volume, also named Quantity, Entries, Total Amount of Data, Number of Available Datasets, and Information Quantity, refers to the volume or amount of data available for analysis, processing, or decision-making [8, 13, 15, 17, 37, 41, 49, 51, 55, 75, 106, 107, 120, 134]. It encompasses the total number of data points, records, or entries within a dataset. Data quantity is a crucial aspect of data management and analytics, as it can influence the reliability and robustness of insights derived from the data. Depending on the process, volume impacts (1) Statistical Significance: A larger dataset can provide more reliable results and reduce analysis error margin. (2) Pattern Recognition: More data points can help in identifying trends, patterns, and correlations that may not be evident in smaller datasets. (3) AI/Machine Learning: many machine learning algorithms require substantial amounts of data to train effectively and produce accurate models. Adequate entries and a suitable amount of data ensure that the dataset satisfies the coverage needs for its intended applications, while a broader selection of datasets enhances versatility and reuse potential, ultimately improving marketability [13, 17, 49, 55, 106].

Metadata, directly related to Profiling, refers to descriptive information about data, such as its structure, provenance, content, and context. It allows users to interpret, discover, and utilise data effectively. Various frameworks emphasise metadata quality as a primary focus, assessing it across dimensions such as semantic consistency, vocabulary usage, and up-to-date documentation [63, 60, 44]. As a metric, it can be considered as a binary descriptor of the existence of relevant fields in the data. These fields can include, among others, (1) Discoverability: Metadata serves as the foundation for identifying and accessing relevant datasets, ensuring they are easily searchable and contextually meaningful (e.g. format). (2) Quality Assessment: helps determine dataset accuracy, completeness, and reliability. It establishes trustworthiness and usability. (3) Contextual Relevance: Allows users to understand the data's origin, scope, maintenance, and intended use, which is essential for aligning datasets with specific applications. High-quality metadata can also provide essential context, including details about data origins (useful for traceability), collection methods, and intended use cases. This transparency improves user confidence and facilitates better decision-making, making the data more attractive to potential users.

Format (or Format Compliance, codification, conformity, available formats) can be used to define both the structure of the data and the percentage of compliant cells in a column that adhere to the specified format. It ensures that the dataset follows the required data structure (e.g., geographical information) [106, 63, 118, 23, 124, 77, 80, 14].

Structure in data asset valuation refers to the organisation and format of data, influencing its usability, quality, and value. It determines whether data is structured, semi-structured, or unstructured, affecting how easily it can be processed and analysed. Syntactic and semantic structures define formats, schemas, relationships, and meaning, ensuring consistency and interoperability. Well-structured data enhances efficiency and analytical value, while poorly structured data requires costly transformation. As a core metric, structure plays a crucial role in assessing data quality and economic potential [75, 11, 124, 24, 15].

Completeness, also referred to as Appropriate Amount of Data, refers to the extent to which all required and expected data is present within a dataset. It is widely covered in the literature [39, 11, 10, 65, 70, 46, 60, 98, 103, 136, 41, 63, 118, 38, 124, 113, 100, 114, 61, 102, 17, 80, 43, 25, 48, 15, 37]. Furthermore, by linking it to metadata, it can assess whether critical data fields are populated. It can also assess whether the dataset includes all necessary records for the intended purpose.

3.6.1.5.2 Contextual

Range is a metric used to quantify the proportion of data values that fall within predefined lower and upper bounds. This reflects the validity of data within expected limits and can impact directly the Utility metric of the data. Range is an essential component of data quality, particularly for numerical variables. The metric ensures that the data conforms to acceptable thresholds, which are often determined by domain-specific knowledge or statistical methods (e.g. non-contextual based on maximum and minimums or quartiles) [80, 46].

Moderation and Typicality - Moderation, if applicable to the dataset, can be defined as a metric that measures measuring the range within which X% (the confidence interval) of the data lies. For instance, assuming a normal distribution, 99.7% of the data lies within three standard deviations of the mean. It is used to evaluate the stability and reliability of data points, focusing on how closely the values adhere to expected norms under stringent confidence levels [80]. Typicality is a data quality metric that measures how well a data point aligns with expected or ”typical” patterns within a dataset. It indicates the degree to which data values conform to common or usual characteristics, helping to identify outliers or unusual events that may require further investigation (i.e. could be more relevant to dynamic types of data) [80].

Volatility is another contextually dependent metric that can be calculated almost directly without the intervention of fundamental quality metrics. The metric measures the frequency at which data values change over time; volatility can also refer to the magnitude of changes over time for a given variable, especially in financial analysis and statistical modelling [11, 102]. Given this dual representation, Volatility can be estimated differently depending on its objective, as described in the appendix section. In both cases, it describes the degree of instability of a data value and how often it undergoes updates. Nevertheless, the relative importance of different timeframes, depending on the feature’s characteristics, makes this metric be considered contextual.

Consistency, highly correlated and treated equally here as Veracity (from compliance) and Reliability of data with the context and needs of the organisation [134], refers to how well data conform to predefined rules. The rules include those association in the context of relational databases and those defined by the metadata and standards. The metric for consistency, as defined by [7] (here defined as intra-consistency), evaluates the consistency of a tuple t based on whether it fulfils or violates certain rules r from a set R of association rules. Consistency, in the context of data management and information systems (here defined as inter-consistency), also refers to the uniformity and reliability of data across different datasets, systems, and applications. It ensures that data remains accurate, coherent, and free from contradictions, which is essential for maintaining data integrity and trustworthiness. Given that connotation, Consistency can be measured in various ways (as defined in the appendix section).

Furthermore, it can be evaluated directly based on Compliance metrics or by those defined for data management (as seen in Figure 6). For the latest (with the exception of Integrity) key aspects of consistency include: (1) Integrity - Ensuring data is accurate and reliable so that the same data yields consistent results regardless of where or how it is accessed or processed; (2) Uniformity - Ensuring data is formatted and represented consistently across different systems, using standardised naming conventions, units of measurement, and data types; (3) Synchronisation - Reflecting data updates across all relevant systems to prevent discrepancies, particularly relevant in environments where multiple systems interact with the same data; (4) Validation Rules - Implementing validation rules and constraints, such as checks for data types, ranges, and relationships, to maintain consistency and prevent invalid entries; (5) Error Prevention - Identifying and mitigating errors that arise from conflicting data entries or updates, which is crucial for decision-making processes relying on accurate data; (6) Data Governance - Establishing governance policies and practices to define standards for data management, including entry, storage, and retrieval processes [75, 11, 10, 70, 46, 98, 103, 136, 128, 63, 38, 129, 113, 34, 9, 102, 17, 80, 43, 25, 28].

Even though all these concepts are relevant, the connection to metrics has been simplified as seen in the taxonomy figures. Furthermore, the relationship between Integrity, Validity, Uniformity, and Consistency can be viewed as complementary rather than hierarchical. However, one can argue that Consistency, Uniformity, and Validity are often a component of integrity (as managed in this work and represented in the figure).

Containment Fraction helps to measure the extent to which one dataset is contained within another. This is considerably relevant in distributed systems, helping to evaluate the data consistency. Additionally, this helps to identify Redundancy and optimise storage usage [105].

Integrity refers to the correctness, trustworthiness, and adherence of data to predefined standards or rules. It ensures that data is complete, unaltered, and reliable for its intended purpose [39, 11, 46,137, 23, 50, 76, 129, 124, 100, 134, 55, 6, 43, 5, 25]. Consistency is one of the dimensions of integrity. It ensures that data does not conflict within a dataset or across datasets. Another way to recall Integrity, indirectly, is **Reliability** (grouped together in the appendix section but for better understanding, separated in Figure 4 and Figure 6). Reliability is strongly dependent of the data integrity.

Uniformity promotes consistency by standardizing the format in which data (or parts of it) is represented—such as using a specific date or number format. It is focused solely on consistent representation, without considering the logical correctness of the data (which is addressed under the concept of Validity, e.g., whether a value is a valid integer). This metric is introduced in this work, but it could be excluded if it overlaps with or is already addressed by Validity.

Validity refers to the compliance of data values individually with predefined rules or criteria that ensure data entries are logically sound and meet the intended business or operational requirements. For example, a person’s age must be an integer, or a cost must always be a positive number [46, 38, 134, 55, 43, 25].

For a better understanding, as an analogy for a puzzle set, Integrity ensures all the pieces are present, untampered, and from the same set. Validity ensures each piece individually conforms to the expected standards (e.g., no piece is warped, double-cut, or has an extra or missing tab). The shape and design of the pieces must be logical and align with the intended use, while Consistency ensures the pieces fit together without conflicts. Finally, Uniformity secures that all pieces must be made of the same material (e.g., cardboard or wood) and have the same thickness, finish, and printing style

Accuracy - Having understood integrity and consistency, Accuracy in scientific analysis is a comprehensive measure of how closely data or results align with true, intended, or expected values. It represents the degree to which data reliably and correctly reflects real-world entities or phenomena, ensuring its applicability and validity to specific objectives. Accuracy is not only about correctness but also about the data suitability for decision-making, analytics, and operational processes [111, 39, 11, 10, 70, 83, 46, 60, 127, 107, 82, 40, 68, 2, 35, 137, 118, 23, 124, 77, 113, 9, 125, 55, 13, 102, 17, 80, 43, 48].

Accuracy is evaluated through a combination of metrics such as range (proportion of data points within predefined valid intervals), consistency, typicality, and moderation [80]. These metrics, along with correctness, proximity to ground truth, reproducibility, and adherence to specified standards, collectively determine the reliability and usability of the data.

Contextual requirements and application-specific thresholds define acceptable accuracy levels, as minor deviations may be tolerable depending on the domain and purpose. Accurate data ensures meaningful insights, reduces risks, and enhances usability across diverse scientific and operational domains [80, 13, 125, 113, 124, 68, 82, 46, 70, 11].

Quality Factor - As outlined in [111], to assess the usefulness of Quality in driving business innovation and growth, the Quality Factor metric can be used as an estimate. As defined, it considers Accuracy and Frequency; nevertheless, this concept could be extended to include other quality metrics (as described in appendix section). The frequency mentioned before is called Information Frequency (IF) and corresponds to a Resolution Metric (Discussed later on).

Detail and Plausibility are concepts both related to Accuracy. Detail, as defined in [39], quantifies whether data is written accurately, thus linking accuracy to the Data Valuation cluster (Figure 6). Plausibilityrefers to the ”degree to which data values match knowledge of the real”, nevertheless this definition has been extended as mentioned in the European Central Bank (ECB) framework. Plausibility (here linked also to Credibility, Believability, Match Between the System and the Real World [39, 11, 60, 136, 107, 137, 44, 124, 134, 114, 17, 25]) has been linked to the process of detecting outliers in the reported data. As reported in [114], ”the assessment carried out by the ECB focuses on those data points that present extreme values that are a possible consequence of errors in the compilation of the reporting obligations”. As a KPI, Plausibility could be considered as a combination of Range - the data should fall within expected ranges or follow patterns that are logically based on historical data or business rules, Consistency - already defined; Domain Rules Adherence - Data should follow known domain-specific constraints.; Lack of Outliers - Implausible outliers or anomalies should be flagged for review, as they may indicate errors in data collection or processing. These links have not been embedded in the taxonomy but could be considered.

Usability is another relevant, widely discussed, and with considerable context-dependent connotation metric. As observed in Figure 6, Usability can be defined as a hierarchical high metric (i.e. a combination of several components). In fact, Usability according to ISO 9241: 11 (2018), is a benchmarking tool that can be used to determine the extent to which a system, product, or service can be used by specific users to achieve the goals determined by the effectiveness, efficiency, and satisfaction of its users.” [44]. Building on the definition of variable indexes from [44] (see Table 1) and insights from other sources [11, 65, 136, 41, 110, 118, 23, 50, 124, 77, 25] that explore the concept of Usability, we have expanded the term to mean: “The ease and efficiency with which quality data—defined by Data-Value, Accuracy, Integrity, and Completeness—can be accessed (through Communication, Accessibility, and Timeliness), understood (Clarity), and effectively used (Ease-of-use, linked to Operational Efficiency, Openness & Performance, Relevance, or Utility) by users to complete specific tasks”. Importantly, not all these components are relevant at the same time and others could be further linked since they depend on the strategies involved. Usability ensures that data not only meets Technical Standards (highly related to Quality and Governance) but also aligns with the practical needs of users, enabling them to extract meaningful insights, make informed decisions, and complete tasks efficiently [124, 25]. Based on the previous description, Usability can be treated as a KPI, rather than a metric, in which a combination of metrics (or other KPIs) are used in its calculation.

Clarity is a composite metric representing the degree to which information is presented in a way that is unambiguous, easily interpretable, concise, and readable. This ensures that it meets user needs effectively. It encapsulates attributes such as understandability, lack of confusion, unambiguity, conciseness, and readability, which collectively enhance information usability and accessibility [39, 11, 70, 83, 107, 128, 63, 10, 4, 44, 17, 25]. Since clarity is a combination of factors, it can be treated as a KPI where the values are agglomerated and weighted. Clarity serves as an integrative measure reflecting several metrics (see the following ones), emphasising the presentation of high-quality, actionable, and user-centred information across diverse applications, from open data platforms to governance and compliance systems.

Conciseness refers to the principle of presenting data or information in a way that is both brief and clear, without unnecessary detail or redundancy. As a metric, Conciseness helps measure clarity (see Figure 6). Furthermore, it can be used as an efficient representation of data quality, fostering minimisation of redundancy while preserving functionality. Conciseness (or Unambiguity) ensures that information is free from unnecessary complexity and includes only relevant data. It also refers to the removal of ambiguity. For example, unambiguous schemas and standardised metadata and presentation of data play key roles in improving data quality and ensuring consistency across systems [39, 11, 83,107, 100].

Understandability, or Ease of Understanding, refers to how easily users can interpret information. Explicit references to understandable data include clear field names, precise definitions, and avoidance of ambiguous units so that users can comprehend datasets without extensive technical expertise [70, 11, 63, 83, 63, 44].

Readability enhances clarity by structuring information for intuitive use. This includes proper formatting, well-designed user interfaces, and visual representations, making it easier for users to interact with and extract meaning from data [10, 83].

3.6.1.5.3 Resolution

As previously described, within this group we agglomerate those metrics that are time, dynamic, or strongly granulometry-dependent. Within this group, Timeliness (Data Freshness) and

Funded by
the European Union

Currency are considerably recognised 39, 11, 10, 4, 70, 46, 60, 98, 103, 136, 107, 118, 23, 38, 124, 77, 113, 131, 134, 93, 114, 55, 45, 61, 17, 80, 14, 43, 5, 25, 48, 16].

Timeliness and Currency - Timeliness is a measurement of the delay between an event occurring and the data being available to the business, therefore is strongly dependent on Age. On the other hand, Currency is whether the data has lost its value due to its processing, modification or elapased time and thus can be linked to change in data value with respect to time change or an event, thus it can be considered to be dependent on data valuation techniques.

Information Frequency (IF) is defined as the rate at which information is updated, accessed, or utilised within a system or process. It reflects the temporal characteristics of information flow and interaction, ensuring that information remains relevant and aligned with the operational or decision-making needs. This metric is linked to operational efficiencies, but given its relative connection to quality concepts, it has been incorporated into this cluster. IF is integrated into multiple dimensions, including:

Raw Frequency Component (IFr): A baseline measure of how often information or data becomes available.

True Frequency Component (IF): A refined measure of frequency adjusted for external factors influencing the information flow.

Frequency Tolerance (FT): The range of acceptable frequency deviations that still supports effective decision-making.

Node Frequency Requirement (FN): The specific frequency threshold necessary to satisfy the operational or decision-making needs of a given system.

This aggregation captures both quantitative and qualitative aspects of frequency, ensuring that information meets the availability requirements of the baseline and is provided with the precision and timeliness necessary for optimised decision-making processes [111]. Even though each of these frequencies could be defined as individual metrics (or linked to others described in this work), the direct connection to IF made us establish it separately and be considered only as part of the calculation of IF.

Funded by the European Union

3.6.1.5.4 Specialised

Entropy and other advanced metrics such as Mutual Information, and Shapley Values expressions (e.g. Shapley Fairness and Shapley Robust) provide deeper insights into the structure and content of datasets. Entropy is a measure of uncertainty, randomness, or information content within a system, dataset, or process. It quantifies the degree of unpredictability or heterogeneity in data and is used to assess information richness, uncertainty, or the effectiveness of data representation. Commonly rooted in information theory, entropy plays a crucial role in evaluating data quality [120, 131], value [71, 131, 132, 133], and informativeness [71, 131, 120, 48, 133, 132]. Although Entropy has extensively been used for valuation, its incorporation as a metric is rooted through its connection to quality and thus, has been established in this work as a Quality metric connected to valuation techniques (as described in Figure 3). The various forms of Entropy (as clustered in this work under the same metric name), including Shannon's Entropy, Heterogeneity, Information Entropy, Additional Information Value (AIV), Joint Entropy, Individual Entropy, Information Score metric (included in the appendix section) are context dependent and thus the decision of which one to implement depends on the strategy involved within business models.

Mutual Information, as defined in information theory, focuses on quantifying the amount of information shared between two random variables. It measures how much knowing one variable reduces uncertainty about the other. It is expressed mathematically as the Kullback-Leibler divergence between their joint probability distribution and the product of their marginal distributions. This makes mutual information a powerful tool for assessing dependencies or relationships between datasets or variables, often applied in data science, machine learning, and signal processing [5, 22]. Mutual information plays a critical role in evaluating Relevance, Consistency, and Completeness. For instance, low mutual information may highlight a lack of coherence in the data, pointing to potential gaps or errors. Furthermore, it serves as a key metric in feature selection and dimensionality reduction, ensuring that only the most relevant and informative attributes are preserved, thereby enhancing the quality of data used in downstream processes [5, 12, 22, 71].

Shapley Fairness and Fairness Metric, as covered in [2], builds on the concept of Shapley Value. The latter quantifies the contribution of individual participants in a cooperative system. It ensures that the allocation of rewards or resources respects specific fairness principles, such as balance (distributing the total value amongst participants), symmetry (equal rewards for equal contributions), zero element (no reward for no contribution), and additivity (consistent allocation across combined tasks).

This framework is particularly relevant in data-centric contexts, as it ties closely to data quality. Shapley Values help assess the importance of individual data points or datasets to the performance of predictive models, reflecting the impact of high-quality data. They can also highlight redundancy or noise by identifying data that adds little value, encouraging improvements in data quality.

Based on the content of [53, 82], the Fairness Metric (linked to Shapley fairness) is a measure designed to quantify the equitable allocation of resources, contributions, or compliance with specific principles in diverse contexts. For example, the Fairness Metric is used in a Weight-based Fair Share Algorithm for allocating cache space amongst virtual machines (VMs) [53]. It ensures proportional allocation based on pre-assigned weights.

FAIRness Score, not to be confused with the previous metrics, evaluates dataset's compliance with the FAIR principles (Findable, Accessible, Interoperable, Reusable) [73]. Automated tools like CkanFAIR compute this score to assess how well datasets meet these criteria, aiming to improve data quality and promote effective sharing.

Data Similarity, named in this work, encapsulates measures such as Euclidean Distance, Projection Similarity, Similarity Score, Cosine Similarity, Average Distance, Kolmogorov-Smirnov (KS), Jaccard Similarity, Mann-Whitney (MW), Mood's Median (MD), and Levene (LE)) [5, 24, 31, 120, 125]. These metrics are used to evaluate similarities or differences between sets, vectors and properties. For example, Jaccard Similarity is ideal for determining the overlap between datasets or sets; Syntactic Similarity (similarly linked to Levenshtein Distance, edit distance, cosine similarity, Q-gram distance, semantic similarity, Jaccard coefficient, MinHash-Based

Distance, Overlap Set Similarity, String-based Measures) focuses on assessing how similar data values are in terms of their syntax and are widely used in text analyses [12, 120, 128]; Stochastic Divergence (named by us) corresponds to metrics that measure the similarity between probability distribution (e.g. T-test scores, Identity-based Exact Match, Jensen-Shannon Divergence, Wasserstein distance).

Objectivity is a measure of the degree to which data or a data source is believed to be free from biases, ensuring that the information presented is impartial and unaffected by subjective influences (during collection, evaluation, and use). It is categorised as an intrinsic or subjective data quality dimension and is critical for assessing the reliability and credibility of data used for decision-making [11, 17, 25]. Since Objectivity is linked to the measure of bias, metrics already known to be used to measure it can be directly linked (e.g. Non-parametric cohort analysis, Statistical Parity, Distributional Skewness, Equalised Odds, and others).

Cost of Degradation, Information Content (IC), and Proximity – Cost of Degradation quantifies the loss in data quality resulting from data transformation [109]. IC is related to the data acquisition process. Some of that data might be very common or predictable, while other data might be rare or surprising. IC gives higher importance to the surprising, less predictable data because it's often more useful for understanding new or important events [56]. Finally, Proximity is described as a factor related to the physical distance of the source of an event. This metric is relevant when there exists a correlation between the quality of the data and the distance of the sensor from the event [5].

Data Robustness and System Robustness - Robustness refers to the ability of a system, model, or process to remain stable and perform well despite disturbances, faults, or unexpected inputs. It is a measure of how well a system can handle variability or adversity while still functioning as intended. Thus, Robustness for systems can be tracked as a combination of Stability - The ability of a system to maintain functionality despite errors or faults; Resilience to Adversarial Attacks and Change - Particularly in machine learning and Data Management, robustness refers to the ability of a model to resist manipulation by adversarial inputs.

Concerning robustness in data, it can be seen as a separate concept not related to statistics. It is related to the concept that data is constructed, acquired, manipulated to survive and function in multiple settings (i.e. resilience to Change but not necessarily to Stability, which is related to the system).

Given these two concepts, we have defined two metrics: Data Robustness and System Robustness. For example, in [2] the emphasis is on data robustness in the context of creating a marketplace for data. This ensures that datasets maintain their value and usability across various prediction tasks. Alternatively, in [68], the focus is on system robustness within federated learning architectures for data marketplaces. It discusses designing a robust model aggregation protocol that excludes low-quality or malicious contributions, ensuring inclusiveness and resilience against attacks.

Robustness is normally classified within Data Quality metrics but given its broader perspective and meaning within different domains, in the present work, it has been included in two domains, Data Quality and Operational Efficiency.

3.6.1.6 Data Governance and Compliance Metrics

Data governance refers to the definition of processes, methods, roles, policies, standards, and metrics for managing and ensuring the proper use, discovery, collection, processing, analysis, disposal, and storage, of data within an organisation [79]. By implementing sound governance, data can be accurate, secure, consistent, and accessible for authorised use while meeting compliance requirements and supporting organisational goals. Importantly, since governance implies a link to the strategies, security, and definition of business architecture, amongst other higher hierarchy specifications, the metrics described here ARE NOT used to cover the wide scope of governance. Instead, governance is used as a cluster that facilitates checks with compliance, safety, and security. With these considerations in mind, in this document, we have categorised the observed metrics and considerations related to Governance into three primary clusters: (1) Compliance, (2) Risks, Safety and Security, and (3) Openness and Ownership.

3.6.1.6.1 Compliance

Compliance involves metrics designed to track adherence to legal, standards, and regulatory requirements. Given the broadness of this aspect, they were further subdivided in two cases: (1) External compliance - In other words legal and regulatory requirements and (2) Internal compliance - regulated by standards, policies and other definitions that facilitate handling of data within the organisation (that can further be fed into External Requirements – see Figure 5). These types of metrics are often context-dependent due to the regulations and policies nuances, independent of standardised models across industries. Furthermore, aggravating the discrepancy in the data monetisation environment, the maturity level in the domain is too low for the existence of broad standards, metrics, and policies.

External Compliance, Data Principles and Maturity Score - In terms of External compliance these include indicators for legal compliance, Compliance Costs (within the Data Valuation Techniques Cluster), Licensing adherence, and the presence of data-sharing policies.

The evaluations conducted in [129, 79] bridge gaps in the literature by identifying and assessing key data governance evaluations and considerations (and thus metrics). These topics are related to Policy, Oversight (or Audit), Data principles and Practices (and correct application of Standards), Maturity Score (From Maturity Assessment) and Data Issue Management, which together can provide a structured framework for understanding and improving governance practices. For example, the DAMA-DMBOK and COBIT-based approach [129] ensures a comprehensive evaluation of governance maturity, offering actionable insights for organisations aiming to improve their data governance capabilities.

Based on the recognition of data principles and practices, oversight, and data issue management, and the recognition of External Compliance and Internal compliance needs, we have incorporated within External Compliance needs Oversight, Data Principles and Practice, and Issue Management metrics that help to encapsulate compliance requirements.

Oversight and Data Principles and Practices - Even though there was no direct mention of Oversight metrics, an evaluation for the organisation’s capability to monitor and validate compliance with governance standards, processes, and policies should be included. Furthermore, the definition of policies for periodic audits to ensure data integrity and proper governance implementation should be in place.

Data Principles and Practices (also referred to as Data Standards, or Standardisation) are related to the rapid discovery, collection, processing, analysis, storage, and defensible disposal of large volumes and fast streams of structured and unstructured data with security, privacy, and cost efficiency. Standards and policies from early adoptions can be used to ensure a correct implementation of the mentioned processes. Nevertheless, even though standardisation is commonly mentioned in the literature, it is not necessarily linked to generalised metrics. For example, some works [46, 129, 124] clearly establish its relevance within evaluation frameworks/metrics, but fail to specify how to measure it. Thus, their approaches have been grouped, helping to evaluate their consideration.

Regulatory Compliance (linked to Compliance Cost), as a metric, evaluates and emphasises adherence to legal requirements and regulations, helping to minimise risks related to non-compliance. For instance, Compliance Cost and Risk Score are integral components in evaluating financial data models. Adherence to a standard such as ISO 20022 ensures regulatory alignment in electronic data exchanges within the financial sector, streamlining international financial operations and mitigating compliance risks [100]. In cybersecurity contexts, compliance metrics assess the alignment of data protection practices with national security regulations. This illustrates their critical role in ensuring operational security and adherence to legal requirements [62].

Furthermore, ownership metrics identified in data integration efforts underscore the importance of clear responsibility and stewardship. This reflects the context-dependent nature of compliance metrics, which adapt to organisational complexities such as varying data governance frameworks [129].

The growing emphasis on compliance reflects increased regulatory scrutiny of organisations managing large volumes of sensitive data. As noted in studies [40, 107], metrics like Compliance Cost and Licensing are crucial for aligning data practices with regulatory standards. These works

Funded by the European Union

also highlight the challenges faced by data owners in navigating complex legal frameworks and the financial implications of maintaining compliance.

Licensing is often context-dependent, given its reliance on industry-specific regulations and national legal frameworks [4, 23, 14]. Licensing, as observed in Figure 5 also impacts Ownership metrics. Full ownership provides the freedom to use, modify, and share data without external constraints, while restrictive licenses or reliance on third-party services reduce control. Open licensing and internally managed datasets enhance ownership value, making it a critical factor in determining how organisations can leverage their data assets. Since Outright ownership can be seen as the opposite of licensing restrictions, only Licensing was considered as a metric.

Internal Compliance, Data Principles and Policy - These metrics focus on and assess the conformity of data practices with established internal standards and policies. A definition of the former was already provided.

When referring to policy, we are referring to a set of formalised principles, rules, or guidelines, created by organisations, institutions, or governments, as internal or operational guidelines [23, 69]. Often, policies are designed to ensure compliance with legal requirements while adding organisation-specific rules or strategies. For instance, [69] defines a policy to meet GDPR compliance but may also introduce stricter internal controls to safeguard sensitive information. The measure of policy as a metric is like data principles and practices, where the Percentage of Adherence, audit results, survey, and feedback can be used.

Traceability Refers to the ability to track and verify data throughout its lifecycle. It involves understanding data lineage—how data flows from its source through various transformations to its final use—and provenance, which captures the origin and history of the data, including any changes made. Maintaining audit trails ensures that all modifications are recorded, including who made them, when, and what was changed, supporting accountability and transparency. Traceability also plays a key role in meeting regulatory and governance standards by demonstrating data integrity and responsible data handling. Additionally, addressability defines how the origin of data can be identified and, if needed, contacted or referenced.

Other metrics useful to estimate traceability includes: Data Lineage Completeness measures the percentage of data elements for which lineage information is available. Provenance Documentation (PD) assesses the percentage of data elements with documented provenance. Audit Trail Coverage evaluates the percentage of data changes that are recorded in an audit trail. Compliance Rate measures the percentage of data processes that comply with established governance and regulatory standards.

Different metrics, derived from the AGDI (Agent-Goal-Decision-Information) model, can facilitate traceability. These metrics include NDG, NDGI, NDI, NGD, NGI, NID, and NIG [66, 108]. Nevertheless, these metrics are related to a model and thus could be context dependent. Independently of this, these metrics have been grouped as Traceability and left for readers to specify the applicability of general metrics or application of Decision-Information models within their business model.

Lineage focuses on the Origin, transformation, and general History of data. It can be settled as a requirement within Internal and External Compliance (i.e. it can be incorporated within Schema, Metadata, and different compliance metrics). In the present work, it has been set as a metric defining a suitable existence, or not, of provenance information.

While traceability metrics are directly applicable to data management, particularly in the context of decision support systems, lineage is a broader concept that includes the ability to link and follow relationships between data, processes, decisions, and goals within a system.

Schema, as a metric, evaluates compliance with predefined data structures, for instance, the presence or absence of desired attributes such as Clarity of definition, Comprehensiveness, Flexibility, Robustness, Precision of domains, Identifiability, Obtainability, Relevance, adaptability to semi-structured data, semantic alignment, standardisation for interoperability, and support for integration processes or others [11, 24, 63]. Due to the need to estimate quality metrics and define schema boundaries, it is essential to include quality metrics related to this cluster.Given the need for estimation of quality metrics and limits to establish Schema, the need for quality metrics linked to this cluster is a must. For example, as mentioned in [11] semi-structured data is data that have a structure which has some degree of flexibility. Semi-structured data is also referred to as schemeless.

Asseen in [18], several metrics have been proposed to assess the quality of data models at the conceptual, logical and physical levels. These different requirements are rate based and are intrinsically linked to the Schema. For example, NFT measures the number of fact tables that follow the Schema. Given the diversity (NFT, NDT, NSDT, NAFT, amongst others), readers are encouraged to review this manuscript. In the present work, all these metrics have been categorised as Schema.

3.6.1.6.2 Openness and Ownership

Metrics related to openness, particularly in the context of Open Government Data (OGD) initiatives, measure the extent of data sharing and transparency, highlighting the need for accessible data that also respects regulatory obligations. These metrics are typically context-sensitive, influenced by specific data-sharing agreements and national transparency policies.

Ownership is a key dimension in data valuation, reflecting the level of control and rights an organisation has over a dataset. It includes outright ownership, licensing restrictions, and Service Agreements [41]. Importantly, ownership also influences other valuation factors like Other costs, Privacy, and Usability (see Figure 2). Higher ownership often involves greater upfront costs but minimises long-term licensing fees and dependency risks. It ensures better integration, uninterrupted access, and increased flexibility for critical applications [23, 41].

Openness refers to the degree to which datasets are accessible, usable, and reusable by the public (depending on the intended use of data). It encompasses the provision of data in machine-readable formats, under licenses that enable free use, redistribution, and modification, and often emphasises accessibility, discoverability, and compliance with open data principles [65, 70, 124, 93, 15]. Thus, as a metric, Openness, evaluates the extent to which a dataset adheres to principles of Accessibility - linked and defined within the Operational Efficiency cluster, machine-readability and non-proprietary formats (i.e. Interoperability), and open Licensing. In its estimation,

Funded by the European Union

formats like the 5-star Model of Openness [104] can be used. Additionally, depending on the type of use defined for the data, factors such as Easy-to-use can impact the metric.

3.6.1.6.3 Risk, Privacy, and Security

These metrics evaluate potential risks associated with data management, system security, and data privacy. In term of risk Concepts, key metrics include:

Risk Cost and Risk Score - For a definition of Risk Cost, please refer to the Data Monetisation section. The Risk Score corresponds to a quantified metric measuring the level of risk associated with data, derived from factors such as sensitivity, compliance requirements, potential impacts of data breaches, and the likelihood of exploitation. It supports prioritisation in data protection and guides decision-making in risk management strategies [75, 69].

Protection Expense is related to the cost of applying protection measures (e.g., encryption, or access control), quantifying the financial impact of data breaches and compliance issues. It is context-dependent, varying according to the legal environment, the nature of the data, and the organisation’s risk profile. It is also part of the Data Monetisation cluster. Regarding security and privacy management are metrics aiming to measure data integrity, preventing unauthorised access, and ensure compliance with privacy laws. The application of Privacy metrics and frameworks is wide, as observed in different sectors including IoT [30], Federated Learning [115], marketplaces [72], and balancing payment and privacy for participants in a data collection mechanism [127].

Differential Privacy - When focusing on Privacy concepts, metrics such as Differential Privacy [127, 30, 74, 115] provide a quantitative measure of privacy by bounding the change in the output distribution of an algorithm when a single individual’s data is added or removed from a dataset (help to assess datasets modifications). Other metrics, such as Inferential Privacy [30], support the estimation of differential privacy.

Privacy Budget, contrary to what its name might suggest, is not a metric related to economic concepts. Instead, it quantifies the amount of privacy a data owner is willing to ”spend” in exchange for participation in a data marketplace or collaborative model training. It determines the level of perturbation (noise) added to the data or gradients during processing, thereby balancing privacy and utility [40, 115] and importantly, it is also related to Differential Privacy and its calculation and thus, has been placed hierarchically in a lower level (as seen in Figure 5) together with metrics such as Number of Sensitive Fields [69] and Data Centre Security Metrics.

Privacy Level is a quantitative measure of data protection, often defined through differential privacy, where the parameter (ϵ) balances data utility and privacy. Lower ϵ values indicate stronger privacy with reduced data leakage, while higher values allow more utility but weaker privacy guarantees [127, 74, 35, 125]. Closely related, and linked, privacy sensitivity reflects the value individuals or organisations place on privacy, influencing the required privacy level, with higher sensitivity necessitating stricter privacy settings. In data marketplaces, this sensitivity impacts compensation mechanisms by aligning privacy preferences with incentives [74, 125].

To check the application of privacy concepts, manuscripts such as [125] highlight the application of privacy-preserving techniques, including differential privacy, in scenarios involving sensitive governmental datasets. For instance, the use of synthetic data in New Zealand's government sector incorporates privacy metrics to minimise re-identification risks [125]. This context-specific approach reflects the reliance on privacy techniques tailored to the needs of the data being handled. Moreover, the balance between data utility and privacy, central to differential privacy applications, is discussed as a key consideration [125].

Security Composite Efficiency Indicator (SCEI) and Security Level Index (SLI) - When referring to security, metrics like the SCEI and the Security Level Index assess the effectiveness of implemented security measures and are generally applicable across industries. For example, the first serves as a holistic measure of the performance of Information Security and Cybersecurity Systems (ISCSS) [62]. It is calculated as the mean of five sub-components:

(1) The Equipping Coefficient with Cyber Defence Means which quantifies system preparedness against cyberattacks;

(2) The Technical Readiness Coefficient which evaluates the operational readiness of cybersecurity mechanisms;

(3) The Equipping Coefficient with Serviceable Cyber Defence Means, which Assesses the availability of practical and functional cyber defence tools;

(4) The Staffing Coefficient with IT System Administrators which reflects the adequacy of IT administrative resources; and

(5) The Staffing Coefficient with Service Personnel which represents the sufficiency of service personnel for cybersecurity operations.

Security as a metric or SLI is a quantitative measure that evaluates the effectiveness of security mechanisms in protecting systems and data. It encompasses indicators such as encryption strength, access control, incident response, and compliance with security standards. Closely tied to Access Security, SLI assesses the ability to regulate and restrict data access, ensuring system resilience against unauthorised use while maintaining operational integrity [20, 62]. For the specific considerations of Data Centres, [95] has defined different metrics that help track the security level: Average Comparisons per Rule, Accessibility Surface, Application Transaction Rate, Concurrent Connections, Connections Establishment Rate, Connection Tear Down Rate, Defence Depth, Data Transmission Exposure, Firewall Complexity, Interface Accessibility Surface, IP Fragmentation Handling, Illegal Traffic Handling, Rule Area, Reachability Count, Rogue Change Days, and Vulnerability Exposure. All these metrics are encapsulated as Data Centre Security Metrics that can be fed to determinate a SLI.

SLI is critical in hierarchical systems, where it stratifies data protection based on sensitivity and operational requirements. This optimises resource allocation and mitigates risks. In IoT systems, trust frameworks are integrated to evaluate data reliability and prevent malicious activities [17, 61]. Additionally, SLI supports large-scale systems by addressing challenges in data volume and velocity [50, 134].

Maturity Model is a structured framework designed to assess the current state of data governance within an organisation. It identifies existing gaps and outlines the necessary steps to achieve a more advanced governance level. By aligning governance practices with strategic goals and evaluating their effectiveness, the maturity model ensures data management evolves to meet organisational needs. This model is typically tailored to specific domains, industries, and data types, enabling it to adapt to Big Data challenges [79].

Metrics play a critical role in feeding and informing the maturity model. Accurate and well-defined metrics provide the data-driven insights to evaluate progress and determine areas requiring improvement. Also, these metrics allow organisations to track their adherence to governance principles, ensuring alignment with compliance requirements and organisational goals. By doing so, they directly contribute to the estimation and refinement of the maturity model.

The strong connection between metrics, governance of the data and the approaches to track and measure them, and the maturity model, makes the latter be strongly dependent on the Governance and Compliance cluster, as observed in Figure 8. Thus, an additional Cluster with a clear differentiation of colour, is incorporated that is fed by the metrics and KPIs affected by Governance & Compliance, fostering this connection. Therefore, this cluster emphasises the importance of regulatory adherence, ethical data management, and the consistent application of governance principles. Through this integration, the maturity model, which is part of the products of T4.3 in DATAMITE, becomes not only a tool for assessment but also a mechanism to drive continuous improvement in governance practices.

3.6.1.7 Operational Efficiency Metrics

After reviewing the literature, we decided to classify metrics and KPIs related to operations in the domain of data into three key areas: Utilisation, Communication, & Performance, Resilience, and Energy Efficiency and Optimisation. The analysis emphasises strictly documented metrics, avoiding assumptions not directly supported by the sources. Additionally, the context dependency of each metric category is highlighted when applicable.

3.6.1.7.1 Resilience

These metrics primarily emphasise the reliability and resilience of data systems, and indirectly, that of data. These metrics address aspects such as:

- System Robustness
- Availability which is linked to Retrievability [39, 65, 118, 23, 50, 76, 54, 84, 134, 93, 55]

- Error Rates also referred to as Error Ratio, Error Count, Inter-Server Error Rate (ISER) which are highly contextual and including uplink, downlink, and trouble tickets, amongst others [36, 112, 51, 40, 122, 55, 15, 37]
- Success Rate (similar to the previous one)
- Maintainability [11,118]

In general, these indicators are essential for ensuring continuous access to data resources, regardless of whether they are hosted on cloud platforms or within on-premise data centres [16, 20, 114]. Metrics directly linked to data resilience are also related to data Integrity.

System Robustness and Redundancy - As previously mentioned, we have defined a separate robustness metric linked to the system instead of data itself (i.e. System Robustness, which can also be recognised to as System Stability). Additionally, Robustness is also considered in terms of replication integrity and equitable data distribution in data markets, emphasising the need to prevent biased outcomes [2, 50, 68]. Note that redundancy of the datasets, which facilitate the estimation of Scalability, is linked to Uniqueness, so no further definition is provided in the appendix section for redundancy. Building on this, it is important to clarify that redundancy within the datasets, particularly as it relates to overlapping or duplicative information, can also be encapsulated within the Clarity metric. This approach reflects the view that reducing unnecessary repetition enhances interpretability and supports more accurate valuation. In contrast, redundancy at the service level, such as duplicated processes, fallback systems, or replicated service pathways, has been maintained as a standalone metric within the operational efficiency cluster. This separation allows for a more granular evaluation of system-level resilience and resource optimization, distinguishing between informational and infrastructural redundancy.

Availability and Retrievability - Availability can be defined as the extent to which data can be accessed and retrieved without error or restriction, ensuring that datasets are easily accessible for their intended use by the intended users [39, 65, 23, 50, 76, 84, 93, 55]. Additionally, it encompasses the readiness and openness of data to be used by the public or other stakeholders [23]. Retrievability [65] is closely linked to availability as both address the ease of accessing data.

While availability ensures that the data exists and is open for use, Retrievability focuses on the actual ability to fetch and use the data effectively without technical or administrative barriers. Given the proximity, we have clustered them as Availability. Furthermore, Availability can be linked to other metrics, facilitating its expression as a KPI (Check the appendix section for more information). For example, securing data to its intended use implies also securing authorised users access and management of the data (i.e. Accessibility), linked to communication metrics in Figure 8 [93].

Additionally, system reliability measured through System Robustness - not to be mistaken with data reliability, named Integrity in this work - ensures consistent system and access availability without frequent downtime, which is a core aspect of Availability.

Redundancy and Backup - Depending on the system type (service or data), metric and concepts such as data Redundancy which supports availability by maintaining multiple copies of data, facilitating system stability under stress, and Backup, facilitate the estimation of Availability. Based on [50], Backup is defined as a critical component of the broader system - backup and recovery capabilities - which focuses on the system's ability to preserve data securely and ensure its availability during recovery scenarios. In our work, to facilitate it conceptualisation, we have grouped both concepts, recovery and system backup, just as Backup. As expected, this metric together with Scalability (defined next) facilitate system robustness.

Scalability, Interoperability, Extensibility, and Adaptability - One key metric within this cluster is Scalability. As observed in Figure 7, it is an intermediate component between System Robustness and Interoperability, thus, Availability (of Data Systems). Scalability refers to the ability of a system, framework, or architecture to handle increased workloads or growing amounts of data by adding resources (i.e. hardware or software) or optimising existing ones, without compromising performance or efficiency [90, 105, 117]. System Concurrent Processing Capability refers to a system's ability to handle multiple processes simultaneously. Scalability is directly linked to this capability, as a scalable system must be able to manage increased concurrency [50]. Another linked concept (grouped with Scalability) is Elasticity. Elasticity is the ability of a system to dynamically adapt to changing workloads by provisioning or removing resources in real-time [76, 117]. Even though not linked in the figure, other metrics can be linked to measure capabilities. For example, the system throughput refers to the amount of data processed, or the number of tasks completed within a specific time frame. For instance, in IoT systems, improvements in both throughput and latency reflect the system's ability to handle increasing data loads and user demands without performance degradation [90, 117].

Other factors that help to measure Scalability are:

- Data Redundancy Reduction, linked to Backup [117] (Efficient deduplication and compression techniques, widely adopted in large-scale storage systems like cloud platforms, have been shown to significantly enhance overall data quality and system performance [110]);
- Interoperability, which is also linked to Compatibility, Applicability, Integration Capabilities, and Implementability, refers to the ability of different systems/datasets to communicate, to be integrated, and exchange data seamlessly. Interoperability is a key factor in the syntactic aspect of data asset valuation. Thus, it can be dependent on other syntactic aspects of data, such as different data Quality factors (Consistency and Granularity), Accessibility, and Schema (and other factors related to the governance of the data) [11, 75];
- Extensibility, which is linked to the ability to extend (add, not necessarily modify) systems, without major changes [117]. Examples include [34]; and
- Adaptability, also referred to as flexibility or versatility, represents the ability of a system (e.g. infrastructure and IoT [117, 124]), process (e.g. adaptability evaluation using Pettigrew’s Context-Content-Process (CCP) model [118]), or service (e.g. Data Model [100]) to efficiently adjust to evolving conditions, changing requirements, or dynamic environments. It ensures systems remain effective across various contexts by allowing seamless integration [44, 110 117].

Mean Time to Data Loss MTTDL and Expected Annual Fraction of Data Loss (EAFDL) - Within this group, as previously mentioned, there is MTTDL (linked to EAFDL) [54] that measures the average time expected before a storage system experiences data loss due to failures that exceed its redundancy and recovery capabilities. It provides a high-level estimate of the frequency of data loss events. These metrics, even though described separately in the taxonomy, are defined as an alternative to measure reliability (and thus integrity) for a specific scenario. As defined in [54], MTTDL and EAFDL are critical metrics used to evaluate the reliability of storage systems.

Error Rate, Success Rate, and Maintainability - Finally, these three metrics, as observed in Figure 7, are the last components linked to Resilience. Maintainability is described by its dependency on Maintenance Frequency and the ability of the data itself to be maintained, thus dependency on the type of data governance and quality components [11, 118]. Other metrics such as Maintenance Frequency underscore the need for regular system checks, helping to minimise downtimes and maintain smooth operations [118].

Error Rate is a performance metric that quantifies the proportion of failed operations in a system, dataset, network, or process over a given period. It is commonly used in service reliability, communication networks, and data processing environments to assess the effectiveness and stability of operations. The error rate can be expressed in different forms, including failure occurrences per transaction, per unit time, per content, or per processed data request. For example, Error Rate is used as a KPI for service troubleshooting, particularly for monitoring fluctuations in success rates and error counts in data centres [112]. In database systems, Error Rate was used to identify anomalies and inconsistencies [51].

Similarly, Success Rate and Task Completion Rate are highlighted as important indicators of system performance and user satisfaction, measuring the proportion of tasks or transactions successfully completed [44, 23, 118, 75].

3.6.1.7.2 Utilisation and Performance

Several key performance metrics in computing, data management, and networking can be attributed to either Technology & Infrastructure or operational efficiency, depending on their context. Technology & Infrastructure (discussed in a following section) focuses on the hardware,

Funded by the European Union

network, and system architecture that define the technical capabilities and limitations of a system. In contrast, operational efficiency refers to how effectively those resources are managed, tracked, optimised, and utilised to enhance performance, as demonstrated in a few real scenarios such as customer relationship management (CRM) data analysis and high-speed railway data management [114, 31, 19, 119]. Given these considerations, metrics like Latency, Bandwidth, Network Traffic Overhead, and Throughput are linked to both clusters. Furthermore, some metrics, like System Utilisation (encompassing CPU Use, memory, slot utilisation, storage usage, Overall Storage Efficiency and Disk Usage) can be representative of several possible objects, given their highly contextual dependency. In other words, System Utilisation is a collection of metrics that are dedicated to expressing the current use of the system, rather than measuring the full capacity of it (i.e. Technology and Infrastructure). Full references are provided in the appendix section. For more information on these metrics, readers are encouraged to check [112, 119, 32, 47, 51, 95].

Latency, Bandwidth, Network Traffic Overhead, and Throughput – These metrics focus on assessing the Responsiveness and Speed (discussed later on) of data processing systems. Latency and Latency-related metrics, including system response time, communication latency, database access latency, transaction processing speed, uplink/downlink communication latency, inter-server communication latency, database access latency, and transaction finality time are particularly crucial in environments where real-time data handling is essential [47, 51, 91, 97, 19, 90, 122, 56, 102, 6, 5]. This includes applications in streaming services, IoT systems, and cloud computing, where low latency and high throughput are vital for efficient performance. End-to-end latency, inter-server latency, and uplink/downlink communication latency are commonly highlighted as key measures for optimising real-time data management, particularly in distributed systems where minimising delays is crucial [19].

Techniques such as data filtering, traffic reduction, and information-oriented traffic management have been shown to improve latency, reduce packet loss, and enhance data throughput, optimising real-time data processing in IoT and Big Data systems [56, 102]. In addition, strategies such as in-memory processing and load balancing are frequently employed to lower latency and boost throughput, particularly in resource-constrained scenarios such as edge computing environments [47, 90]. In blockchain applications, reducing network latency and increasing transaction processing speed are key focus areas for improving system performance in high-demand environments [91].

Responsiveness, Speed, Response Time, Elapsed Time, and Runtime – Responsiveness, closely linked to Response Time, Runtime (also named Processing Time), and Elapsed Time, serve as key indicators in evaluating the efficiency of computing environments and computing processes. Importantly, while Speed tends to be confused with Responsiveness, the former refers to the raw processing capability of a system — such as how quickly a database query is executed. Speed is typically measured in terms of latency (e.g., milliseconds per query).

Responsiveness is a broader concept that emphasizes the system's ability to react promptly to user inputs or external requests. While it includes Speed (since fast processing helps responsiveness), it also considers other factors like communication delays, queuing, system load, and how quickly the system responds overall from the user's perspective. Given this extension or hierarchy, Speed [39, 11, 44, 50, 13] could be included as one metric within Responsiveness.

Response time refers to the total duration from the moment a request is made to when a response is received. It is often evaluated using metrics such as average response time and the 95th percentile response time, which provides insight into worst-case performance scenarios [119]. Systems with optimised response times demonstrate better performance, especially in applications requiring real-time interactions, such as financial transactions, machine learning inference pipelines, and high-frequency trading platforms.

Similarly, Elapsed Time accounts for the full duration of a process or transaction. It encapsulates all contributing factors such as network delay, resource contention, and processing bottlenecks [91]. This metric is particularly relevant in cloud computing, where system workloads fluctuate dynamically, affecting transaction speeds and response predictability. Research indicates that reducing response time and optimising elapsed time through efficient workload distribution and caching mechanisms significantly enhances Responsiveness [53, 91, 119, 102].

Runtime refers to the total duration during which a program, process, or system executes from start to completion. It includes processing time, memory access, input/output operations, and system overhead. In performance evaluation, runtime efficiency directly impacts responsiveness, as shorter runtimes enable faster task execution and lower response times [102, 110].

Service responsiveness is a fundamental factor in customer satisfaction, especially in logistics and e-commerce platforms, where rapid service execution is essential. Studies highlight that responsiveness, when linked to response time, plays a key role in shaping user expectations. Delays beyond one second impacting perceived service efficiency, and latencies exceeding ten seconds leading to transaction abandonment [137]. Thus, reducing the service response time leads to benefits in system efficiency, user retention, and user satisfaction.

In data-intensive environments, network traffic overhead and bandwidth availability further influence system responsiveness. High-throughput systems must balance network congestion, latency variability, and database access speed to ensure seamless data delivery [102]. Advanced methodologies, such as dynamic load balancing, adaptive caching, and predictive prefetching, have been proposed to mitigate response time degradation and improve overall responsiveness [6]. Additionally, distributed architectures utilising multi-server deployment strategies have shown significant improvements in reducing server response time and transaction finality time, ensuring consistent performance across various workloads [53].

In conclusion, responsiveness is a multifaceted metric that integrates response time, latency, and elapsed time to evaluate system efficiency in diverse computing environments. Whether in real-time data streaming, cloud-based service management, or high-performance computing, optimising these parameters is essential for maintaining competitive and reliable system performance.

Additionally, Figure 7 includes a conglomerate of other metrics called Network Metrics. These are specific for data centres and include: Diameter Stretch, Path Stretch, Maximum Relative Size, and Network Utilisation [95]. Other metrics are also described in the same manuscript, but given their direct linkage to Energy Usage, they have been set in a following section.

Encryption Time (ET), Decryption Time (DT), and Data Centre Security Metrics – Concerning Privacy and Security, metrics assessing the effectiveness of mechanisms designed to safeguard data are key. From a performance perspective, these indicators include, for example, ET and DT [31], which measure the efficiency of encryption processes, ensuring that data protection measures do not excessively delay system performance. Other metrics that measure the performance of security, such as the number of weak logins and Detection Performance have been encapsulated within Data Centre Security Metrics. Importantly, do not confuse metrics related to the complexity and level of security with the focus of this cluster which is the operational efficiency of the processes involved in it.

Even though not highly covered in the manuscript found, metrics for system security are widely covered in general in the literature and could be linked within this cluster. Nevertheless, when data valorisation and monetisation is enforced in the analyses, not enough manuscripts were found. Independently, readers are encouraged to review [95], which broadly cover system security and metrics for it.

3.6.1.7.3 Energy Efficiency Metrics

These metrics and KPIs focus on tracking energy consumption while maintaining optimal system performance. Given the broadness of these metrics, an additional figure connected with the main taxonomical one was provided. (i.e. Figure 9 connected to Figure 7 and Figure 8). A general description of the metrics is provided in the appendix section, while the dependencies are defined based on expert judgment and information extracted from [67, 85, 95, 101, 121, 135]. Some of the metrics were eliminated or merged with other more relevant, given the low information about them (e.g. H-POM, SI-POM, OSWE, PESavings), the disconnection with operational unit / system performance on operations (e.g. Technology Carbon Efficiency (TCE)), or the general concept that is not necessarily associated to data centre (e.g. Material Recycling Ratio - MRR).

In addition to the primary metrics discussed, Figure 9 incorporates various metrics related to utilities that support data centres and data management equipment. These diverse metrics include, for example, ACE, DCLD, and Airflow Efficiency. Depending on environmental conditions, these metrics may also involve seasonal adaptations.

The more commonly known metric dedicated to data centres is Power Usage Effectiveness (PUE), which is the ratio of the total amount of energy used by a data centre to the energy used by the computing equipment alone. The most well-known data centre efficiency metrics at the moment are defined in ISO/IEC 30134-2 / EN 50600-4-2. Beyond the standard PUE metric, several derivatives have been developed to provide more granular insights into specific aspects of data centre efficiency [21, 32, 48, 72, 85, 95, 99, 101, 121, 122, 135]:

- Partial Power Usage Effectiveness (pPUE);
- Data Centre Efficiency (DCE);
- PUE Scalability (PUEscalability);
- Server Power Usage Efficiency (SPUE);
- PUE Categories (PUE1, PUE2, PUE3, PUE4).
- Carbon Usage Effectiveness (CUE) which measures carbon emissions associated with the energy used by a data centre;
- Energy Data Centre (DC) which represents the total energy consumption of a data centre, encompassing both IT hardware energy usage and supporting infrastructure (e.g., cooling, power distribution, and lighting).;
- Renewable Energy Factor (REF) which measures the proportion of a data centre’s total energy consumption that is sourced from renewable energy;
- On-site Energy Fraction (OEF) which measures the proportion of the data centre’s energy demand met by on-site renewable sources over a given period;
- On-site Energy Matching (OEM) which assesses how well the on-site renewable energy generation aligns with the data centre’s energy demand;
- Data Centre Infrastructure Efficiency (DCiE), Energy Reuse Effectiveness (ERE), and Energy Reuse Factor (ERF), which evaluate the efficiency of energy use, helping to measure the proportion of waste energy, that can be reused. .

These metrics vary considerably depending on infrastructure type, such as cloud computing setups versus traditional data centres. Variations arise due to differences in energy consumption patterns, cooling systems, and integration with renewable energy sources [95, 99].

Additional metrics included within this cluster (as seen in Figure 9) includes:

- Adaptability Power Curve (APC) which shows how efficiently power adapts to changing workloads, ensuring scalability, reliability, and minimal energy waste;
- Corporate Average Data Centre Efficiency (CADE);
- Compute Power Efficiency (CAPE) which measures the overall energy efficiency of an organisation's data centres by averaging their efficiency scores, typically using Power Usage Effectiveness (PUE) and IT performance metrics;
- Various data-centre-specific energy metrics: DCcE, DCeP, DCLD, DCPE, DC-FVER, DWPE that are related to the facility performance.

Even though several of the metrics previously mentioned (like PUE) tend to be defined under an IT perspective of equipment, they are more general and can be seen from a facility perspective.

On the other hand, there exist IT/rack-equipment-specific metrics, such as:

- H-POM which is the ratio of AC hardware load at the plug to the DC hardware compute load, which is the amount of power used only by the computing components;
- ITEE which assesses the energy efficiency of IT equipment by comparing its capacity with the energy it consumes;
- ITEU which measures how much of the IT equipment's capacity is being utilised;
- SWaP which assesses how effectively a system utilises physical space and power consumption relative to its performance output;
- DCPD which is the power consumption per rack within a data centre;
- PDE a variation of the PUE metric, called power density efficiency (PDE)

Complex metrics such as Adaptability Power Curve (APC) have also been used with relative success that checks the energy flexibility in data centres participating in demand response programs [123]. Similarly, Adaptation of Data Centre to Available Renewable Energy (APCren) [86] measures how well the data centre adapts to the availability of renewable energy sources, influencing energy cost savings. A higher APC and APCren value indicate a greater capacity of the data centre to shift its energy consumption in response to energy and renewable energy availability. These metrics have been grouped as APCren and considered within a greener perspective.

Finally, given the variety of operational units in HVAC systems, it’s essential to specify their availability, capacity, and efficiency. While some specific metrics are included in the figure, due to the relative functionality of operational units, some metrics found in references are categorised as ”Other HVAC.” (e.g. Cooling Effectiveness Rate (CER) which can also be confused with Connection Establishment Rate (CER)). For more information on general metrics related to utilities, please refer to [67, 95, 116, 121].

3.6.1.8 Technology and Infrastructure

As previously mentioned, several key performance metrics in computing, data management, and networking can be attributed to either Technology & Infrastructure or operational efficiency, depending on their context. In this section, we discuss the hardware, the network, and the system architecture that determine the system's capabilities and limitations. After performing the analysis in Technology and Infrastructure related metrics and KPIs, we identified two key areas: Ingestion Capabilities & Capacity and Task Placement.

3.6.1.8.1 Ingestion Capabilities & capacity

Ingestion Capabilities is a key component (or metric) that quantifies the ability of a data collection system to efficiently and reliably acquire, process, and store data from distributed sources, particularly in high-velocity environments such as the IoT, edge computing, and industrial sensor networks.

Communication Metrics - Ingestion performance is influenced by several ”Communication” components such as Throughput, Latency, Network Efficiency and Energy Consumption. These components determine a system’s ability to handle continuous and concurrent data streams [126].

Based on the previous considerations, and as observed in Figure 8, Ingestion capabilities are represented by what we refer to as Communication Metrics — which are not included in the Appendix section, as they consist of a weighted collection of sub-metrics that support data acquisition, accessibility, and storage—and by component velocities. Several of these metrics have been mentioned before. Data throughput, for example, defines the rate at which data is successfully ingested into the system, typically measured in records per second or megabytes per second. Higher throughput indicates a more efficient ingestion pipeline, as demonstrated by systems such as IoTDB, which achieves an insertion rate of up to ten million values per second [126].

Velocity describes the rate at which data is generated, ingested, and processed over time. It reflects the continuous flow and accumulation of data within a system. A system can exhibit high Speed but low Velocity if it processes small amounts of data very quickly. Conversely, a system can have high Velocity if it handles large, continuous streams of incoming data in real time, even if individual processing actions are not exceptionally fast.

Traffic Energy, Traffic Ratio, and Hop Distance - Traffic Energy quantifies the energy consumption associated with the transmission and processing of ingested data. This aspect is critical in resource-constrained environments such as edge computing and IoT deployments, where optimising energy efficiency directly impacts system sustainability [36, 126].

Traffic Ratio (linked to Internal Traffic and External Traffic) further refines the evaluation of ingestion efficiency by measuring the proportion of useful data successfully transmitted relative to total network traffic, providing insight into the effectiveness of the system in minimising redundant or unnecessary transmissions [36].

Hop Distance is a key network performance metric in distributed and cloud computing, measuring the number of intermediary nodes a data packet traverses before reaching its destination. It directly affects latency, energy consumption, and network efficiency. This metric consists of uplink/downlink hop distance (UDHD) and inter-server hop distance (ISHD). UDHD quantifies the hops between a data centre gateway and computing servers, where fewer hops reduce transmission delays and congestion. ISHD measures the hops between computing servers, with higher values leading to increased latency and energy use. Optimising hop distance through hierarchical routing or server-centric architectures enhances performance and reduces the cloud infrastructure environmental impact [122].

System capacity – These metrics are fundamental in evaluating hardware components' characteristics and their impact on system utilisation and scalability. These metrics encompass various aspects, including CPU performance, cache size, memory capacity, available slots, storage space, disk throughput, Input/Output Operations Per Second (IOPS - number of read and write operations a storage device or storage array, can handle per second), and power capacity, amongst others [51,126]. The interplay amongst these factors is complex, as each contributes to optimal system performance. Achieving optimal values for these metrics involves strategies that consider usage patterns, logistical considerations, and other contextual factors. In the realm of data monetisation, effectively managing these metrics is crucial, as they directly impact the system's ability to handle data processing demands and support scalable monetisation strategies. Given the broadness of their specification, they have only been grouped as System Capacity in Figure 8. Few other considerations are made as follows.

Scalability, already mentioned in the Operational Efficiency perspective, is a critical challenge in IoT and blockchain environments due to limited CPU, memory, and bandwidth, which impede efficient data processing and system performance. [17] highlights the necessity of trust-based data quality frameworks to ensure reliable processing in resource-constrained environments. Similarly, blockchain scalability is hindered by high computational costs associated with consensus mechanisms and transaction validation, necessitating tiered architectures and off-chain storage solutions.

Storage bandwidth and Input/Output Operations Per Second (IOPS) – These metrics are key system capacity measurements, as insufficient IOPS can degrade system responsiveness and increase data retrieval latency. [20] proposes hierarchical storage frameworks and automated data migration strategies to optimise storage efficiency and reduce performance bottlenecks.

Integrating edge computing, SSD storage, and decentralised processing can enhance scalability, mitigating latency and computational overhead in large-scale IoT-blockchain systems.

A review of [20, 21, 102] highlights key advancements in distributed storage, SSD utilisation, and hierarchical storage frameworks. [102] emphasises the role of Hadoop MapReduce in handling large datasets, requiring scalable and efficient storage architectures. [20] explores tiered storage architectures, where HDFS (Hadoop Distributed File System) manages large datasets, and SSDs enhance high-speed access to frequently used data. [21] underscores the importance of low-latency storage solutions, including SSDs, to sustain high-throughput data processing.

3.6.1.8.2 Task Placement

Task Placement Ensuring efficient data processing is a critical factor in optimising system performance. According to [6, 86], task scheduling and workload allocation play essential roles in ensuring that tasks are executed on the most appropriate system resources. Several key performance indicators (KPIs) can be used to evaluate scheduling and workload allocation, including some already covered such as PUE, CER, and different metric to check profiles (e.g. Data Centre Energy Profile Change (DCA)). The use of metrics that track changes on racked metrics together with allocation constraints and IT power budgets can help to define strategies for task placements. Similarly, other important metrics include the tracking of energy consumption shifts and Adaptation of Data Centre to Available Renewable Energy (APCren) to evaluate how well the system integrates with renewable energy sources. Additionally, [21] evaluates software performance by analysing its ability to manage large datasets without overloading system resources, particularly within data centre and cloud environments. Metrics such as Energy Reuse Effectiveness (ERE) for energy conservation and CO2 Savings.

3.6.1.9 Innovation and Growth – Oriented

Finally, in this section, we present metrics linked to new methods and techniques focused in data use and management. These metrics are essential for assessing the ongoing transformation and adaptation of enterprises as they integrate innovative solutions, refine their processes, and respond to emerging technological and market trends. Furthermore, this cluster includes key performance indicators (KPIs) that facilitate the measurement and tracking of system or enterprise growth, specifically by assessing the progression of different subsystems and their impact on overall business development. To provide a structured approach to tracking such evolution, maturity models play a crucial role in evaluating an enterprise's growth trajectory (and therefore their direction connection as observed in Figure 8). These models offer a staged framework that allows organisations to measure their progress across predefined levels — ranging from initial or ad-hoc implementation to optimised and data-driven maturity. By leveraging such models, enterprises can systematically monitor their progress, identify gaps, and strategically enhance their operational and analytical capabilities. Considering previous definitions, innovation and growth-oriented KPIs are categorised into three key domains: Inclusiveness, Values, and Fair Data Practices – Assessing ethical, transparent, and equitable data use; Novel Metrics and KPIs – Evaluating the effectiveness of new performance indicators in driving business innovation; and Diffusion, Distribution, and Expansion – Measuring scalability, data integration, and market growth.

3.6.1.9.1 Inclusiveness, Values, and Fair Data Practices

This cluster evaluates the Reliability, Transparency (linked to Market Penetration Metric), and equitable access to data in shared ecosystems. Trust metrics, including Trustworthiness, Trust Scores, Assurance, Transparency, amongst others, are used in IoT networks to assess data reliability, particularly in heterogeneous sensor environments, ensuring data quality.

Trustworthiness is a key metric in decentralised systems [15, 29, 37, 41, 118, 124]. For example, the DeDa platform integrates blockchain mechanisms for data integrity and supports data monetisation [91], while martFL applies verifiable aggregation techniques to improve accuracy in federated learning systems [68]. The metric Trustworthiness (referred to here also as Assurance or Trust Score) is used to evaluate the reliability and credibility of entities within a digital ecosystem. This metric is often derived from factors such as identity verification, transaction history, reputation, and credibility as perceived by transacting parties.

In the context of a decentralised IoT data marketplace, a Trust Score (TS) is calculated based on two primary components: Reputation which is determined by the verification status of an entity's identity, often validated using decentralised identifiers (DIDs); and Credibility which reflects the confidence an entity has accumulated from past transactions, assessed through ratings and feedback provided by transacting parties [29]. Therefore, Trust Score or Assurance Score serves as a critical metric in digital systems, particularly in data transactions, decentralised platforms, and blockchain-based marketplaces.

Trust in data ecosystems is crucial for effective data monetisation. An information system evaluation model underscores the need for objective measurement and assurance in data transactions [118]. A quality assessment framework is also introduced in [102] to enhance transparency in big data processing [102]. The impact of data monetisation on societal benefits is another key consideration. Research highlights how data-driven innovation enhances social welfare by improving decision-making and service delivery [124]. Furthermore, [93] highlights the importance of open government data in promoting transparency and accessibility. This is done by ensuring that datasets are available in usable formats for broader public engagement and evaluation, facilitating equitable opportunities and access for stakeholders.

Finally, inclusiveness and FAIR metrics are used to address equitable data access and transparency. The CkanFAIR tool evaluates dataset compliance with the FAIR (Findable, Accessible, Interoperable, Reusable) principles to enhance data reusability and openness [73]. The marketplace for data framework explicitly incorporates Shapley fairness principles to ensure a fair allocation of value amongst data contributors by considering their marginal contributions in cooperative settings [2]. Additionally, DeDa promotes inclusiveness by enabling diverse data sources to participate in data sharing and trading using decentralised mechanisms. Some of these metrics were already defined within the Data Quality Metric (Advance Metrics) and thus, as observed in Figure 8, a direct linkage (only to those types of metrics related to fairness) is provided.

Confidence and Inclusiveness - Another metric that can help track system reliability is Confidence which is defined as a measure of the reliability of a rule [120]. It expresses how likely the consequent is to be true given that the antecedent is true. It answers the question: ”If someone buys bread, how likely are they to also purchase butter?” Confidence helps you evaluate whether a rule is strong or weak. A high confidence means that once the condition is met (buying bread), the expected outcome (buying butter) is very likely to occur.

In the case of federated learning, the Inclusiveness metric [68] helps measures the percentage of benign data providers (DPs) whose local models are selected for aggregation in a federated learning framework. It contrasts with robustness, which quantifies the exclusion of malicious DPs. The document highlights a trade-off between inclusiveness and robustness, where higher inclusiveness allows more benign participants but risks including low-quality models, while higher robustness improves security but may exclude legitimate contributors.

Transparency can be seen as a complex metric that impacts data and data processes. Securing that data is measurable, verifiable, and interpretable, supporting data quality, governance, and model fairness [63, 82, 129]. In terms of data transparency, it is linked to Reputation. This is due to its close link to service-level agreements (SLAs) and business processes, where transparency ensures that the data is measured and reported effectively under previously defined processes. Transparency can be considered to ensure that data sources, transformations, and decision-making criteria are openly documented and accessible. This helps prevent bias (and therefore its connection to Values and Fair data practices) by allowing stakeholders to audit and validate data usage.

Statistical parity [82] is used as a fairness measure to evaluate whether a decision-making model is biased against a particular group. It is particularly useful in the context of fairness-aware models, where the aim is to ensure equitable treatment of individuals regardless of sensitive attributes.

3.6.1.9.2 Diffusion, Distribution, and Expansion

These metrics assess the strategic flow and utilisation of data within and across organisational boundaries. The concept of information diffusion evaluates the effectiveness of data insights dissemination throughout an enterprise, influenced by factors such as organisational structure and data accessibility policies.

Information Diffusion, also referred to as data distribution efficiency, is a fundamental metric in distributed systems and federated learning frameworks, as it quantifies the effectiveness of data propagation across nodes and stakeholders [26, 126]. In the context of digital information asset evaluation, information diffusion is closely linked to the data accessibility and sharing, directly influencing its utility within a given system.

As outlined in [124], the information value is determined by several characteristics, including Information Diffusion, its Structure, and data Accuracy. In large-scale computational environments, such as those modelled in [26], the efficiency of data distribution significantly impacts performance, particularly in terms of load balancing, network latency, and task scheduling.

Discoverability, linked to platform performance, assesses openness and demonstrates the importance of making datasets easily accessible to stakeholders [93, 83]. Discoverability is the degree to which a dataset, information system, or platform enables users to locate, access, and understand its content efficiently. This is done through search mechanisms, metadata, and structured organisation. It measures how easily users can find relevant datasets or resources, often assessed by factors such as indexing, categorisation, metadata quality, and search functionality.

Lifecycle, also called Shelf Life, brings temporal considerations. In the context of decision-based valuation, it represents the duration for which a decision node remains active and capable of generating value [111, 129, 15, 37]. It is useful for projecting income and determining an amortisation schedule. Since information is perishable, its usefulness can change over time, making the lifecycle estimation a critical factor in valuation methodologies.

3.6.1.9.3 Novel

This cluster encompasses different metrics that could be linked, indirectly, to other previously covered clusters. Nevertheless, given their low coverage and connection to specific components within the data market environment, we have decided to cluster them here. For example, social welfare metrics are employed to assess the collective benefits of data-sharing activities, particularly within federated learning and data marketplace contexts, where maximising data utility can have broader social impacts [115]. Social welfare metrics are context-dependent, shaped by specific collaborative goals and data governance frameworks in practice.

Growth Rate, as a metric, could be interpreted as the rate at which data volume, data records, or other measurable data entities increase over a given period [15]. This metric helps organisations track the expansion of their datasets and assess whether this growth aligns with their data value strategies. Importantly, depending on their rate of growth (linear, exponential, cyclical), processes for management can be improved.

As expected, other metrics could be clustered here; it is up to users of the present taxonomic approach to specify the allocation of new metrics. Nevertheless, these definitions can impact the relative importance of balanced scorecard components.

3.6.2 RQ2 - *What are the main trends and relations observed between the observed KPIs and Metrics for data monetisation/valuation?*

Figure 1 up to Figure 9 illustrate the taxonomic approach proposed for structuring Metrics and Key Performance Indicators (KPIs). This approach integrates well-established strategic frameworks, such as the Balanced Scorecard (BSC), with data monetisation principles to provide a comprehensive methodology for assessing strategic performance.

To visually represent the relationships and dependencies amongst these elements, two key components are used:

- Arrows denote dependencies between different metrics and KPIs.
- BSC dependencies are grouped under the concept we introduce as Strategic Data Value. This term encapsulates the most relevant metrics within the taxonomy. It effectively represents an aggregate KPI that conveys the status of an organisation's strategic implementation.

The taxonomy itself is structured across three hierarchical levels, ensuring a logical flow from high-level strategic insights to granular technical data:

- Higher level – This level aligns with the fundamental structure of the Balanced Scorecard (BSC) and consists of KPIs that provide a broad overview of the status of each main BSC component. These KPIs determine how well an organisation is progressing in achieving its strategic goals.
- Middle level – This level is closely linked to data stewards and is responsible for aggregating specific metrics into well-defined KPIs within complex clusters. It acts as an intermediary layer that refines raw technical data into meaningful performance indicators.
- Technical level – The foundation of the taxonomy, this level encompasses the processes involved in data collection, processing, and metric extraction. It ensures that all necessary data points are captured and structured in a way that feeds into middle-level KPIs, enabling accurate performance assessment.

By structuring the taxonomy in this way, we create a systematic and scalable method for evaluating an organisation’s strategic data value. This hierarchical approach not only ensures coherence across different levels of decision-making but also enables enterprises to monitor, analyse, and optimise their strategic objectives with increased precision. The integration of the BSC framework with data monetisation principles allows for a more dynamic and data-driven evaluation, making this approach highly adaptable to various business and technological environments.

3.6.2.1.1 How the Taxonomy Was Created:

The taxonomy presented in Figure 1 up to Figure 9 was developed through a structured, expert-driven process that combined top-down alignment with the BSC components with a bottom-up empirical evaluation. Its construction integrated insights from experts and a systematic clustering of academic and technical literature on data valuation, performance metrics, and data governance (i.e. systematic literature review).

The process began by identifying and analysing a large set of metrics and KPIs referenced across state-of-the-art publications in the data monetisation environment. These sources were evaluated and grouped based on their conceptual similarities and logical dependencies. Metrics with strong thematic cohesion, such as those related to risk, cost, utility, and ownership, were clustered into categories, forming the basis for the technical level of the taxonomy. These clusters were refined through multiple analyses against practical use and BSC components.

From there, the middle layer of the taxonomy was defined. This layer functions as a semantic bridge between the granular, metric-rich technical level and the strategic overview found in the upper level. Clusters were evaluated not only by their intrinsic coherence but also by their functional role in common data stewardship activities. Key constructs such as Data Quality, Governance & Compliance, and Ingestion Capabilities emerged naturally as aggregators of technical clusters that feed into broader decision-making processes

3.6.2.1.2 How to Use the Taxonomy:

When considering the relationship between KPIs and metrics, it is imperative to note that while they are interconnected, the estimation of a KPI does not necessarily require the calculation of all related metrics. For example, a Financial KPI can be effectively represented by incorporating key financial metrics such as Operating Expenditure (OPEX), Capital Expenditure (CAPEX), Return on Investment (ROI), and Net Present Value (NPV). However, not all these metrics are mandatory for KPI estimation. Due to context dependency and the need to define specific thresholds, including limits, benchmarks, and minimum or maximum values (which are often dictated by strategic considerations), a tailored selection of metrics should be established to ensure an accurate representation of a given KPI that effectively reflects overall strategic performance.

Determining the optimal set of metrics and KPIs is beyond the scope of the taxonomy. Instead, appropriate decision-making methodologies should be applied to identify the most relevant indicators in a given context.

In general, the objective is to track the overall strategic performance using the measured Strategic Data Value. This serves as a representative measure for evaluating strategies or datasets. If the focus is on datasets rather than databases, data centres, or other data-related components, Strategic Data Value can be leveraged to assess dataset value within a strategic framework. Achieving this requires clearly defining dependencies between data valuation techniques and middle-level KPIs, ensuring a robust representation of dataset value to the overall strategy.

Furthermore, by appropriately assigning contributions (i.e., weights) to each of the main BSC components, the Strategic Data Value can be systematically estimated. This weighting approach ensures that each component's impact on the overall strategy is properly accounted for. This facilitates a more accurate and meaningful assessment of strategic success.

3.6.2.1.3 How to Start Using the Taxonomy as a Framework

Given the inherent complexity of the data environment, defining the most suitable metrics can be challenging. As observed during the development of the taxonomy, irrespective of the complexity of individual metrics, the clusters that were most independent — and therefore essential as a starting point — were those associated with Governance & Compliance. Establishing a robust foundation for governance facilitates a structured and efficient approach to metric definition and implementation.

A well-defined governance framework should be built on fundamental principles such as schema, lineage, and Data Principles and Practices. By clearly articulating these foundational elements, organisations can accelerate the structured development and integration of additional metrics. Moreover, fundamental data quality metrics serve as the cornerstone upon which other metrics are derived. By integrating governance principles with data quality considerations, a comprehensive framework can be established. This framework can include well-defined structures, policies, and best practices that support the systematic development of metrics.

The impact of Governance & Compliance extends beyond metric definitions — it directly influences data integrity, security, and regulatory adherence. Data governance reduces risks related to data mismanagement, privacy breaches, and regulatory penalties by ensuring that data is accurate, traceable, and compliant. Furthermore, governance structures facilitate seamless data sharing while maintaining consistency and accountability by enabling interoperability across systems.

By leveraging Data Principles and Practices, organisations can create an adaptable framework that allows for continuous evaluation and refinement of metrics. This iterative process enables organisations to test the relative importance of different metrics and refine their relevance over time. Furthermore, as a system matures — a process that can be monitored through a Maturity Model — decision-making methodologies such as Analytic Network Process (ANP) and Analytic Hierarchy Process (AHP) can be employed to systematically identify and prioritise the most suitable metrics. These methodologies facilitate a structured approach to metric selection, either per cluster or through an aggregated assessment, ensuring alignment with the organisation's strategic goals.

Ultimately, the incorporation of robust governance and compliance considerations not only enhances metric selection but also strengthens the overall data management strategy. By establishing governance as a foundational pillar, organisations can ensure that their data-driven decision-making processes are reliable, scalable, and aligned with both regulatory requirements and business objectives.

4 ANP and AHP applied to metrics and KPIs within data Environment

This section provides a short description of the connection between the strategies, the balanced scorecard, the different metrics, and a decision-making approach. These four different topics are brought together as a web-based tool within DATAMITE, which is presented at the end of this section.

Previously, in the introduction section, Strategy, Metrics and KPIs connection was described without going into a deep level of complexity. Within this section, a higher insight is performed describing an approach to interconnect them. The initial interconnection was made through the

Balanced Scorecard (BSC) Framework without going in detail on how this framework can be connected to each strategy. This section explores how strategies, the balanced scorecard, metrics/KPIs, and decision-making approaches interconnect, culminating in a web-based tool developed within the DATAMITE project. Building on the earlier overview, this section dives deeper into these connections, starting with a recap of key strategies (Section 4.1), criteria for achieving business goals (4.2), and linking metrics and KPIs to decision-making (4.3). It then introduces the ANP and AHP methods as strategy-independent decision-making tools (4.4), outlines their implementation in DATAMITE (4.5), and concludes with initial validation results, key insights, and future development plans (4.6).

4.1 Strategies and Connection of DATAMITE T4.1 with T4.2.

Deliverable D4.1 offers a structured approach for organisations aiming to unlock the latent value of their data. By distinguishing between internal and external monetisation strategies and providing practical tools like the 4x4 matrix and an in-depth questionnaire, it helps businesses navigate the choices needed to create (or refine) successful data-driven business models. Whether a company’s goal is improving in-house operations, selling insights to external clients, or forging data-sharing alliances, D4.1 sets the foundational framework that can be adapted to various industries and organisational sizes. The overarching emphasis is on aligning data-focused initiatives with core business objectives, ensuring compliance with regulations, and fostering an environment where data fuels both innovation and revenue growth.

The strategies derive from a thorough literature review, case studies, pilot interviews, and the project’s overarching conceptual framework. D4.1 groups these strategies using Gassmann’s Business Model Framework (focusing on value creation, value proposition, revenue mechanisms, and target customers), supplemented by Osterwalder’s Value Proposition Canvas to clarify how data-based offerings align with user losses, gains, and needs.

D4.1 formally identifies twelve monetisation pathways. They are broadly organised by direct vs. indirect monetisation and further grouped under internal, external, or “non-monetary” categories. These strategies are:

Data-as-a-Service(DaaS): *Core Idea* - Sell (often raw) datasets or data streams to external customers. *Key Requirements* - Large data volume, anonymisation capabilities, a robust delivery platform. *Revenue Model* - Typically one-time sales (per dataset) or recurring subscriptions.

Information-as-a-Service(DaaS): *Core Idea* - Provide structured insights, analyses, or reports instead of raw data. *Key Requirements* - Analytical expertise, data visualisation, industry-specific know-how. *Revenue Model* - Subscription-based, pay-per-report, or consultative upcharges.

Answers-as-a-Service(DaaS):
Core Idea - Move beyond static reports to deliver data-driven recommendations or solutions (e.g., consulting, interactive dashboards). *Key Requirements* - Advanced analytics, domain expertise, “decision support” tools. *Revenue Model* - Often subscription plus premium “expert service” fees.

Reduce Costs (Internal Monetisation): *Core Idea* - Use data internally to streamline operations and slash expenses — no external sales required. *Key Requirements* - Strong data culture, cost-mapping, KPI tracking to measure savings. *Revenue Model* - “Hidden” monetisation through increased margins rather than direct sales.

Data Network: *Core Idea* - Connect multiple stakeholders (e.g., suppliers, partners, industry peers) to exchange data for mutual benefit. *Key Requirements* - Trusted relationships, secure data-sharing protocols. *Revenue Model* - Membership fees, brokerage or facilitation fees, or value-added services.

Wrapping: *Core Idea* - Embed new data-driven features into existing products/services (“wrap” them with insights) to justify higher prices or retain customers. *Key Requirements* - A clear link between the product's functionality and the data-driven addition. *Revenue Model* - Typically integrated into the product's price, boosting margins.

Servitisation: *Core Idea* - Transform a traditionally one-off product sale into an ongoing service contract via continuous data collection and analytics (e.g., predictive maintenance). *Key Requirements* - Hardware capable of data capture, advanced analytics, recurring subscription or "pay per outcome". *Revenue Model* - Leasing, pay-per-use, or subscription for real-time monitoring and proactive service.

New Products/Services: *Core Idea* - Create brand-new offerings by leveraging internal data (e.g., launching a new digital product suite). *Key Requirements* - Innovation mindset, product development resources, market analysis. *Revenue Model* - Standard sales, subscription, or license fees for the new product/service.

Data Bartering: *Core Idea* – Trade data or insights for valuable assets (e.g., software tools, marketing benefits) rather than direct cash. *Key Requirements* – Clear guidelines on data ownership, robust legal frameworks, mutually beneficial partnerships. *Revenue Model* – Indirect — value exchanged in kind instead of via payment.

Share-for-Free: *Core Idea* – Offer data openly (e.g., open data initiatives) to gain market traction or to foster developer ecosystems. *Key Requirements* – High-level anonymisation, well-defined scope and terms of use. *Revenue Model* - Often used to build brand awareness; potential for future upsells.

Data Platform Providing: *Core Idea* - Operate a platform that directly hosts and distributes datasets from multiple sources to paying customers. *Key Requirements* - Advanced data storage, integration, security measures, user-friendly environment. *Revenue Model* - Commission fees, subscriptions, or listing fees for data providers.

Data Platform Refining: *Core Idea* - Specialise in enhancing or transforming data from multiple external contributors (e.g., cleaning, enriching) before it goes on sale. *Key Requirements* - Data curation and metadata management, thorough knowledge of data quality standards. *Revenue Model* - Fees based on how much refining is performed, or a premium charge to data buyers for higher-quality datasets.

While several strategies may coexist within one organisation — either combined or sequenced over time — each requires unique metrics to monitor its effectiveness, growth, and value creation within the organisation. These metrics, combined with the strategic goals derive specific KPIs to track, in this way the unique outcomes/criteria (e.g., cost reduction vs. new revenue), so the KPIs must capture those specific objectives. Furthermore, some organisations could start with a given strategy (e.g. DaaS), but as they mature, more advanced offerings (e.g., AaaS, or Servitisation) would require different performance measures. To be more precise, some KPIs and metrics that can be seen as relevant for DaaS includes subscription rate, number of data transactions, data quality scores, anonymisation/compliance checks, churn rate for subscriptions. As an organisation evolves to an AaaS goal, they can include additional KPIs and metrics such as repeat customers, insight accuracy, customer satisfaction scores, predictive model accuracy, amongst others.

Then, as expected, there is a direct connection between T4.1 and T4.2, through the links that exist between strategies and KPIs. Regretfully, the maturity on metrics and KPIs within the data market environment is relatively low, it is thus hard to classify and connect suitable metrics and KPIs to strategies. Independently of this, the new taxonomy proposed here is a first step in that direction. Additionally, as discussed later, the web-based tool created during this work (and presented in section 4.4) provides initial metrics and KPIs that could be useful for data monetisation. These metrics can then be further expanded by user interests by adding new ones (e.g. repeat customers that is not part of the taxonomy derived). Finally, KPIs (as grouped by BSC components) can then be measured as a weighted aggregation of normalised metrics (as defined in Section 2.1).

4.2 Criteria to Achieve Business Goals

In an organisation’s architectural framework — whether focused on business processes, data infrastructure, or operational blueprints — clearly defined criteria play an important role in unifying disparate functions and aligning each toward a set of overarching objectives or strategies. By articulating specific concerns such as cost, productivity, quality, employee satisfaction, safety, learning and growth, and customer satisfaction, leadership ensures that resource allocation and data analytics converge on the factors that matter most.

Beyond simplifying the landscape (including those of potential metrics), criteria also enable the organisation to remain adaptable. For instance, linking learning and growth directly to performance measurement emphasises that architecture must evolve alongside both the market environment and emerging internal opportunities. Importantly, codifying these concerns within the architecture’s design fosters resilience and sustainability over the long term, as the systems and processes are continually refined through feedback loops intended to capture skill gaps, compliance demands, and inefficiencies.

The role of these criteria in translating high-level strategies into quantifiable actions is equally significant. By specifying core concerns — ranging from cost per unit to customer satisfaction index — each department gains a clear sense of how day-to-day decisions and work streams contribute to larger strategic ambitions. Because cost, quality, safety, and other priorities are viewed as universally relevant across the enterprise, they provide a common language that fosters cross-functional collaboration. This shared set of measures ensures that each functional area, whether it concerns product development or sales, interprets its targets in a manner that aligns with broader goals.

Even though it could be defined that criteria are not relevant by themselves to the goals of DATAMITE, their relevance within enterprises and the approach of [138], makes their consideration a must. In fact, based on this work, seven criteria are easily recognised to help organisations achieve their strategies. Each criterion is framed differently in DATAMITE given the specific domain it supports. The criteria are as follows:

Cost: Monetary value of all inputs or resources that a company uses to produce goods or services. The grading is the relative importance, not that it is increased or reduced.

Productivity: Productivity in business refers to the efficiency with which inputs (such as labour, capital, and raw materials) are converted into outputs (goods and services). High productivity means achieving more output from the same or fewer inputs, which is crucial for a company's profitability, competitiveness, and long-term sustainability.

Quality: Increase quality, ensuring fewer defects, higher conformance to standards, and overall product/service reliability. Data quality refers to the condition of data based on various attributes (e.g. accuracy, completeness, consistency, etc.) that determine its reliability and utility for specific purposes.

Employee Satisfaction: Refers to the extent to which employees are happy, content, and fulfilled with their jobs and work environment. It encompasses a variety of factors that contribute to an employee's overall well-being and their feelings about their role within the organisation.

Security: Involves implementing measures and practices to protect data from unauthorised access, misuse, corruption, or theft. Ensuring the security of datasets is crucial for maintaining the confidentiality, integrity, and availability of data, especially when dealing with sensitive or personal information.

Learning and Growth: Refer to the continuous process of improving the capabilities, skills, and knowledge of an organisation's workforce to drive innovation, efficiency, and competitive advantage. The components can include technological advancement, training and development, organisational culture, knowledge management, etc.

Funded by the European Union

Customer Satisfaction: Refers to the measure of how well a company's products or services meet or exceed customer expectations. It is a key indicator (subdivided in indicators such as product quality, service quality, etc.) of customer loyalty, repeat business, and overall business success.

4.3 Decision-Making as a solution to define Metrics and KPIs

According to Rodrigues et al. [138], structured decision-making methods are especially well-suited for selecting metrics and Key Performance Indicators (KPIs) because they impose systematic, transparent procedures on what could otherwise be an unmanageable list of possibilities. In their work on an analytic network process (ANP) model for the tooling and die industry, they illustrate how decision-making approaches offer three principal benefits.

First, they allow organisations to translate diverse stakeholder viewpoints into quantifiable comparisons. The ANP framework incorporates feedback loops and interdependencies, enabling managers to articulate not just how one KPI compares to another, but also how each might influence or be influenced by a broader network of company objectives and constraints.

Second, decision-making techniques supply robust mechanisms to handle large volumes of qualitative judgments. In many industrial settings, knowledge resides with multiple experts — production coordinators, quality engineers, financial controllers — who each bring specialised insight. Structured models, such as ANP, convert these subjective inputs into normalised weights or priorities, ensuring that each opinion is accounted for without reducing the process to mere guesswork or gut instinct.

Finally, they help detect and resolve inconsistencies that inevitably arise in pairwise comparisons, particularly when the KPI set is extensive. Through iterative sessions and validation checks, experts can refine their inputs until the model achieves an acceptable level of consistency. This procedure boosts confidence in the final prioritisation, as the selected KPIs are not only grounded in expert knowledge but also vetted for logical coherence. By methodically linking criteria such as cost, productivity, and quality to broader organisational goals, these decision-making approaches unify strategic aims with day-to-day performance measurement in a way that is both collaborative and data-driven.

4.4 Analytical Network Process (ANP) and the Analytical Hierarchical Process (AHP)

Decision-making in complex environments frequently requires the consideration of multiple, often conflicting criteria, which is referenced as Multiple Criteria Decision-making (MCDM). Traditional decision models, which emphasise linear optimisation or single-objective evaluation, frequently fall short in capturing the nuanced preferences and interdependencies amongst decision elements. To address the challenges of MCDM, Thomas L. Saaty introduced the Analytic Hierarchy Process (AHP) and later extended it into the Analytic Network Process (ANP).

The AHP is a decision-support tool that decomposes a problem into a hierarchy of sub-problems, each of which can be analysed independently. The process begins by defining the overall goal of the decision problem, which forms the top level of the hierarchy. Below this, decision criteria and sub-criteria are structured in successive levels, culminating in a set of alternatives at the lowest level. This hierarchical model facilitates a systematic comparison of various decision elements in pairs, allowing for subjective judgments to be quantified. The comparisons are performed using a fundamental scale of absolute numbers (from 1 - equal importance to 9 - extreme importance), reflecting preferences between pairs of elements.

Pairwise comparison matrices are constructed for each hierarchy level, and the relative weights of the elements are derived using eigenvector methods. These weights represent the priorities of the elements with respect to their parent criteria. An important feature of the AHP is the inclusion of a consistency check, whereby a consistency ratio (CR) is computed to assess the logical coherence of all the judgments. Once consistency is established, the weights are synthesised across the hierarchy to determine the overall ranking of the decision alternatives.

One of the principal limitations of AHP is its assumption of independence amongst the elements at each level. In many real-world scenarios, such as technology assessment or strategic planning, the interrelationships and feedback amongst decision elements play a crucial role and cannot be ignored.

The key advantage of ANP lies in its ability to compare elements not only within the same cluster but also across different clusters, providing a comprehensive view of the factors influencing a decision. This, however, introduces the challenge of increased complexity and potential inconsistencies, as the process involves a larger number of pairwise comparisons. Despite this, ANP effectively captures the inherent complexity of decision-making situations, making it particularly useful in reflecting the relative importance of Key Performance Indicators (KPIs).

The methodology of ANP begins with the development of a network model that identifies all relevant clusters and elements, along with the dependencies amongst them. Like in AHP, pairwise comparisons are then conducted to determine the influence of elements with respect to each other. These comparisons lead to the construction of a supermatrix, a partitioned matrix that captures the influence of all elements on each other across clusters. The supermatrix is then normalised and raised to limiting powers to derive the so-called limit supermatrix, which represents the long-term stable priorities of the elements.

To summarise, both the ANP and AHP offer powerful tools for structured decision-making. but given the relative complex environment in connecting several strategies with metrics and KPIs that need to be linked to the objectives of the different enterprises, the ANP was selected as the main approach in DATAMITE.

In general, the use of the ANP and AHP process to define metrics and KPIs is not broadly covered in literature. One of the pioneer works in this area is the work of Rodriguez et al. [138] which directly links the BSC components within an ANP structure. This work was used as the backbone for defining the final ANP structure within DATAMITE, which is specifically covered in following sections.

4.4.1 ANP Formulation

Let $C1, C2, \ldots, Cm$ be clusters in the decision network. In our case, these clusters correspond to the different components of the BSC, plus one dedicated to the criteria used to foster the achievement of the goals, and another linked to the strategies, so m = 6. Each cluster C_i contains elements $e_1(i), e_2(i), \ldots, e_k(i)$. These elements can influence or be influenced by other elements within the same cluster or different clusters. In our case, these elements correspond to the respective metrics, strategies (from T4.1) and specific criteria.

Based on the structure covered in the following section, the pairwise comparison of elements e_i and e_j within a cluster can take place and is noted $a_{i,j}$. Based on the axiom of reciprocity, each pairwise comparison follows the following rule:

$$a_{i,j} = \frac{1}{a_{j,i}}$$

where *i,j* represent the *i* and *j* elements of the pairwise comparison process. One important characteristic included in ANP processes is that the pairwise comparison can be aggregated, allowing to ponder relative feedback from several stakeholders within the same domain. In our case, this aggregation is driven with the idea that stakeholders (k) with similar base strategies from T4.1, can be aggregated to have an integrative and initial recommendation. In other words, provide a recommendation of what metrics should be considered by users of a specific strategy. This can be achieved by calculating the aggregation G as:

$$G = \sqrt[n]{\prod_{1}^{n} a_k}$$

Once all the pairwise comparisons are performed, a matrix is constructed by respecting the axiom of reciprocity. This matrix has the form

$$A \;=\; [\,[1, a12, \ldots, a1n], [1/a12, 1, \ldots, a2n], \ldots, [1/a1n, 1/a2n, \ldots, 1]\,].$$

With this matrix, the priority vector w can be computed using the principal eigenvector of A.

$$A \cdot w \;=\; \lambda_{max} \cdot w$$

This eigenvector, λ_{max}, can be used to estimate the consistency of the pairwise comparisons performed within the cluster. The consistency is calculated as:

$$CI \;=\; \frac{\lambda_{max} - n}{n - 1}$$

which depending on its value, can lead to a rejection of the information provided.

Once the different vectors are estimated based on the different clusters, a supermatrix can be constructed. The supermatrix W is a partitioned matrix that holds all the priority vectors representing the influence of one element on another. It is structured as:

$$W = [[W_{11}, W_{12}, ..., W_{1n}], [W_{21}, W_{22}, ..., W_{2n}], ..., [W_{n1}, W_{n2}, ..., W_{nn}]]$$

Each $W_{i,j}$ is a submatrix representing the influence of elements in cluster C_i with respect to cluster C_j. One important characteristic of this supermatrix is that at this stage, interlinkage between clusters has not been incorporated. Thus, at this stage further pairwise comparisons can be performed to incorporate further dependencies (i.e. priority vectors). If only one dependency is foreseen between elements of two clusters (many to one or one to one), the resulting priority vector on the individual cluster will have only one component, translating in a normalised value of 1. Contrarily, if there are no dependencies, the priority vector has only zeros values.

To ensure the entire matrix is stochastic (values are probabilities and the sum of the values of a row is 1), each block column of the supermatrix is weighted by the importance of the corresponding cluster (also obtained by pairwise comparison).

The final steps involve constructing a modified supermatrix, named the limit supermatrix that corresponds to raising the weighted supermatrix to large powers until it converges to a steady state. In other words, this involves:

$$W_{limit} = \lim_{k \to \infty} W_{modified}^{k}$$

The limit supermatrix provides the final global priorities of the elements — fully accounting for all direct and indirect influences within the network. These values show the *relative importance* of each element when considering the entire system dynamics. In the case of DATAMITE, these relative importances, hereafter referred as Relative Weights, can be used to accumulate metrics values to calculate global KPIs. In this form, individual metrics, connected true ANP estimations calculate domain specific interest KPIs (e.g. Quality, Trustworthiness, etc.) that can then be further aggregated by the Relative Weights of the BSC domains. In this way a global metrics can be calculated to encapsulate the Relative Value of a dataset.

Importantly, if identical Relative Weights are used over two datasets, the final Relative Values can be used directly to select the most useful of both. Furthermore, this value can be used within benchmarking process to further perform comparison and dataset selection.

4.4.2 ANP Structure for Data Monetisation

As previously mentioned, the first stage in constructing and evaluating the ANP process is to define a suitable structure that defines the interactions. Given the link between ANP and the BSC components, the structure developed by Rodriguez et al. [138] can be used almost directly. This structure is presented in the following diagram.

Funded by
the European Union

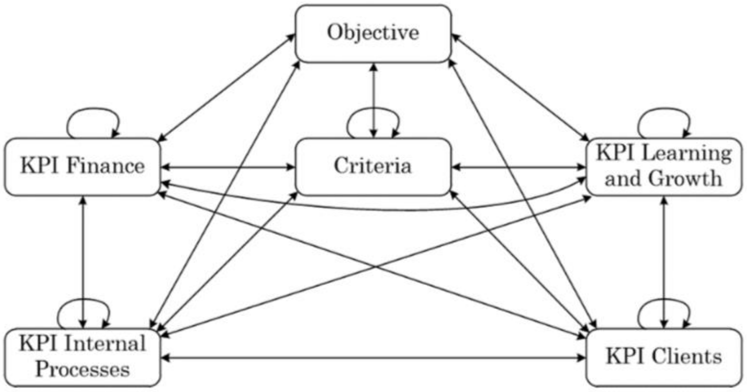


Figure 12 ANP structure used in DATAMITE; directly extracted from [138].

As observed in Figure 12, each BSC component are interlinked within themselves and with other BSC components. The elements interconnected correspond to the metrics from the taxonomy that share clusters. In other words, they belong to the same BSC component. Importantly, not all the metrics should be involved on each cluster, only the relevance to the specific user of this framework.

Additionally, each BSC cluster is connected to the others, through interconnection between metrics, and linked to criteria and objectives. The objective in DATAMITE corresponds to the global KPI that represent the strategic interest of the user. Thus, by agglomeration of different relevant users metrics, similar to the taxonomy diagram, users can estimate a global metric for their strategies.

4.5 ANP tool for Metrics and KPIs for Data Monetisation

Despite the recognised importance of robust metrics, many companies find themselves overwhelmed by the sheer abundance of available KPIs (as relatively summarised in the taxonomy). Traditional scorecards and heuristic approaches — while practical in small-scale settings — struggle to capture the complex, networked relationships that drive modern data

Funded by the European Union

monetisation models. As Rodrigues et al. have shown [138] decision-making methods such as the ANP, are well equipped to handle large sets of qualitative and quantitative inputs.

Building on this notion, T4.2 has introduced a web-based tool underpinned by the ANP method, specifically tailored to define metrics and KPIs for data monetisation strategies. The tool’s architecture comprises a backend capable of administering user’s loggings and complex ANP calculations and a frontend designed to guide decision-makers through defining, prioritizing, and refining their performance metrics. Distinguishing it from generic spreadsheet-based solutions, the tool harnesses both the computational rigor of ANP and a user-friendly interface that lowers the barrier to adoption in organisational settings that range from small and medium-sized enterprises to multinational firms.

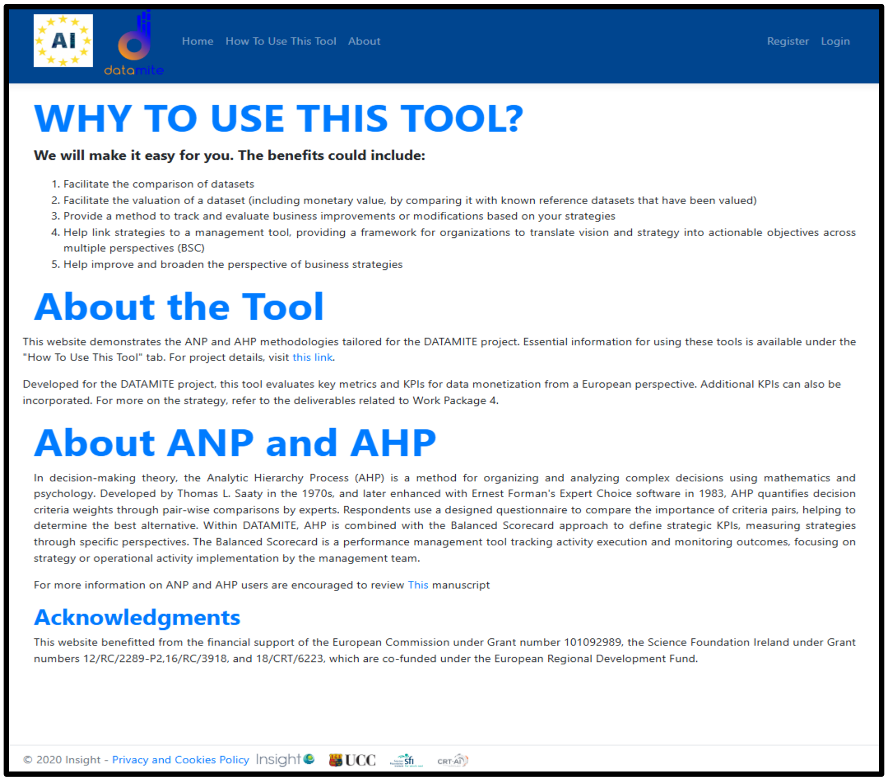
Central to the tool’s development is the need to capture the overlapping influences amongst criteria such as cost efficiency, innovation capacity, regulatory compliance, and customer value realisation. In data monetisation contexts, these dimensions are not isolated: innovations that drive new subscription revenue streams, for example, may require adjustments to compliance frameworks when data assets involve personally identifiable information or confidential corporate records. In this sense, an ANP-based approach is uniquely positioned to capture the reciprocal relationships amongst criteria that would otherwise be artificially ranked in a linear hierarchy. The interface invites domain experts to collaboratively assess how each criterion relates to the others, effectively mapping a network of cause-and-effect relationships that more accurately reflect the organisational reality of data-driven initiatives.

Moreover, the web service architecture confers practical benefits for distributed teams. By consolidating the ANP methodology into a centralised platform, organisations can incorporate multiple viewpoints — business analysts, legal advisors, IT specialists, and executive leadership — into a single decision-making process. This collaborative approach reduces blind spots and fosters alignment, ensuring that the eventual prioritisation of KPIs resonates with both strategic objectives (increasing revenue, improving market position, mitigating risk) and day-to-day operational realities (workforce availability, data infrastructure readiness, development timelines). The tool also embeds iteration and inconsistency checks mechanisms that allow for recalibration when contradictory pairwise judgments arise, thus improving the reliability of results.

In sum, this web-based ANP tool delivers a structured, repeatable approach to a challenge that often defies simple guidelines: the identification of high-value metrics and KPIs for data monetisation. While anchored in rigorous decision-analytic theory, it is designed to be practically deployed, factoring in the constraints and complexities that real-world organisations face. By making specialised methods such as the Analytic Network Process more accessible and fostering richer stakeholder engagement, the tool aims to elevate both the quality and the transparency of metric selection — an outcome that ultimately promotes more informed, data-driven strategic planning in the digital economy.

The tool is available online by using the following link https://datamite.insight-centre.org. The tool is currently running on a server in the Insight Centre but will be made available as an open-source project including the source code and a Docker container soon. The objective is to make available for a local deployment within an organisation. The tool is adaptable and extensible by allowing the creation of new strategies, criteria, metrics, and KPIs by various stakeholders connecting to it.

The following figure shows the homepage frontend of the tool's first version. At the top of the page, there are tabs that allow users to: (1) access further information on ***how to use this tool*** (which includes basic concepts, including definitions of metrics, BSC, and the ANP methodology); (2) get further information ***about*** the creators and general nomenclature; and (3) ***register*** or ***log in***. Currently, the service follows the security and privacy protocols, as established at the Insight Centre (see this link https://www.insight-centre.org/privacy-statement/). But these are momentary countermeasures, for testing and validation, since the idea is, again, for this tool to be deployed locally.



WHY TO USE THIS TOOL?

We will make it easy for you. The benefits could include:

1. Facilitate the comparison of datasets
2. Facilitate the valuation of a dataset (including monetary value, by comparing it with known reference datasets that have been valued)
3. Provide a method to track and evaluate business improvements or modifications based on your strategies
4. Help link strategies to a management tool, providing a framework for organizations to translate vision and strategy into actionable objectives across multiple perspectives (BSC)
5. Help improve and broaden the perspective of business strategies

About the Tool

This website demonstrates the ANP and AHP methodologies tailored for the DATAMITE project. Essential information for using these tools is available under the "How To Use This Tool" tab. For project details, visit this link.

Developed for the DATAMITE project, this tool evaluates key metrics and KPIs for data monetization from a European perspective. Additional KPIs can also be incorporated. For more on the strategy, refer to the deliverables related to Work Package 4.

About ANP and AHP

In decision-making theory, the Analytic Hierarchy Process (AHP) is a method for organizing and analyzing complex decisions using mathematics and psychology. Developed by Thomas L. Saaty in the 1970s, and later enhanced with Ernest Forman's Expert Choice software in 1983, AHP quantifies decision criteria weights through pair-wise comparisons by experts. Respondents use a designed questionnaire to compare the importance of criteria pairs, helping to determine the best alternative. Within DATAMITE, AHP is combined with the Balanced Scorecard approach to define strategic KPIs, measuring strategies through specific perspectives. The Balanced Scorecard is a performance management tool tracking activity execution and monitoring outcomes, focusing on strategy or operational activity implementation by the management team.

For more information on ANP and AHP users are encouraged to review This manuscript

Acknowledgments

This website benefitted from the financial support of the European Commission under Grant number 101092989, the Science Foundation Ireland under Grant numbers 12/RC/2289-P2,16/RC/3918, and 18/CRT/6223, which are co-funded under the European Regional Development Fund.

© 2020 Insight - Privacy and Cookies Policy

Figure 13 ANP tool Homepage

Registering and logging in provide access to new tabs and information. The main new tab opened when accessing your individual service, is the tab called “My Evaluations”. Within it, users can start or recall ongoing ANP evaluations. Each ANP evaluation requires to set a specific name to it and define the user’s application domain. In the current version, the application domain includes Service Users, Data Provider, and Service Stakeholder. In future versions, as detailed later, this could be provided by users to capture specific feedback from the difference stakeholders within a company. This will facilitate later aggregation of information.

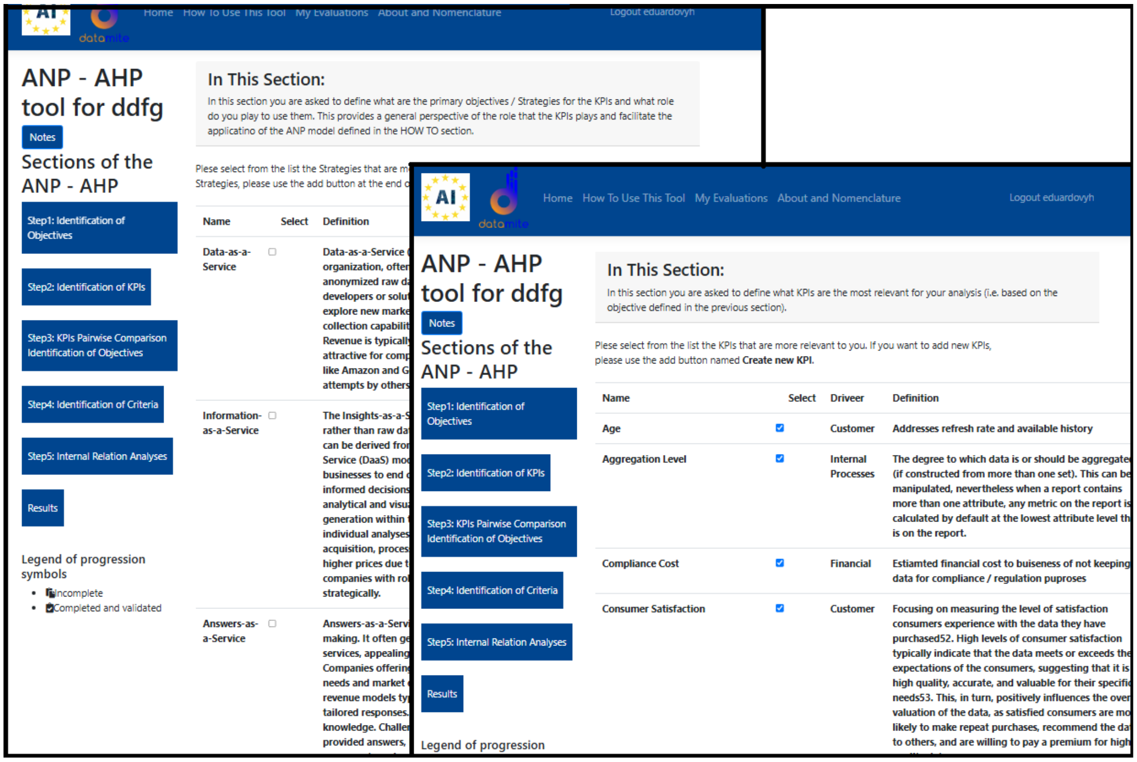
One user can have several ANP evaluation processes at the same time, given possible strategies changes, modifications of metrics and KPIs selection, or the need to update previous ongoing evaluations.

The ANP process is divided into steps that the user can access as he completes them via the left navigation bar as shown in Figure 14. These steps are:

(1) Identification of Objectives – In this step, a strategy from those described in Section 4.1 is selected. Only one can be selected at a time. In case another strategy is needed, a separate evaluation needs to be created. The user is then asked to provide the pairwise comparison of the 4 BSC families of metrics/KPIs. As shown in Figure 15, a slider is used to determine the user's preference between two components. The range is between -8 to 8 to represent a whole range of the Saaty scale (bidirectional). After each BSC component pairwise comparison is performed, a consistency calculation is performed. If a huge discrepancy on the information provided by the users (values over 1.4 - calculated automatically by the tool), a warning message is displayed, and the user is asked to revise the pairwise comparisons.

(2) Identification of KPIs – In this step, a user selects metrics/KPIs that are the most relevant for the selected strategy. Unfortunately, in this first version of the tool, no recommendation for the most suitable metrics and KPIs is provided (further internal and external evaluations are required). Any number of metrics/KPIs can be selected. However, since the number of pairwise comparisons grows exponentially with the number of elements selected, it is advised to select a reasonable number of them.

Funded by
the European Union


AI
datamite
Home
How To Use This Tool
My Evaluations
About and Nomenclature
Logout eduardovyh
Notes
Step1: Identification of Objectives
Step2: Identification of KPIs
Step3: KPIs Pairwise Comparison Identification of Objectives
Step4: Identification of Criteria
Step5: Internal Relation Analyses
Results

ANP - AHP tool for ddfg

Sections of the ANP - AHP

In This Section:

In this section you are asked to define what are the primary objectives / Strategies for the KPIs and what role do you play to use them. This provides a general perspective of the role that the KPIs plays and facilitate the applicatino of the ANP model defined in the HOW TO section.

Plese select from the list the Strategies that are m
Strategies, please use the add button at the end o

Name	Select	Definition
Data-as-a-Service	☐	Data-as-a-Service (organization, often anonymized raw da developers or solut explore new marke collection capabilit Revenue is typically attractive for comp like Amazon and G attempts by others
Information-as-a-Service	☐	The Insights-as-a-S rather than raw dat can be derived fron Service (DaaS) moc businesses to end c informed decisions analytical and visua generation within t individual analyses acquisition, proces higher prices due t companies with rol strategically.
Answers-as-a-Service	☐	Answers-as-a-Servi making. It often ge services, appealing Companies offering needs and market c revenue models typ tailored responses. knowledge. Challer provided answers,

Legend of progression symbols

- Incomplete
- Completed and validated

ANP - AHP tool for ddfg

Sections of the ANP - AHP

In This Section:

In this section you are asked to define what KPIs are the most relevant for your analysis (i.e. based on the objective defined in the previous section).

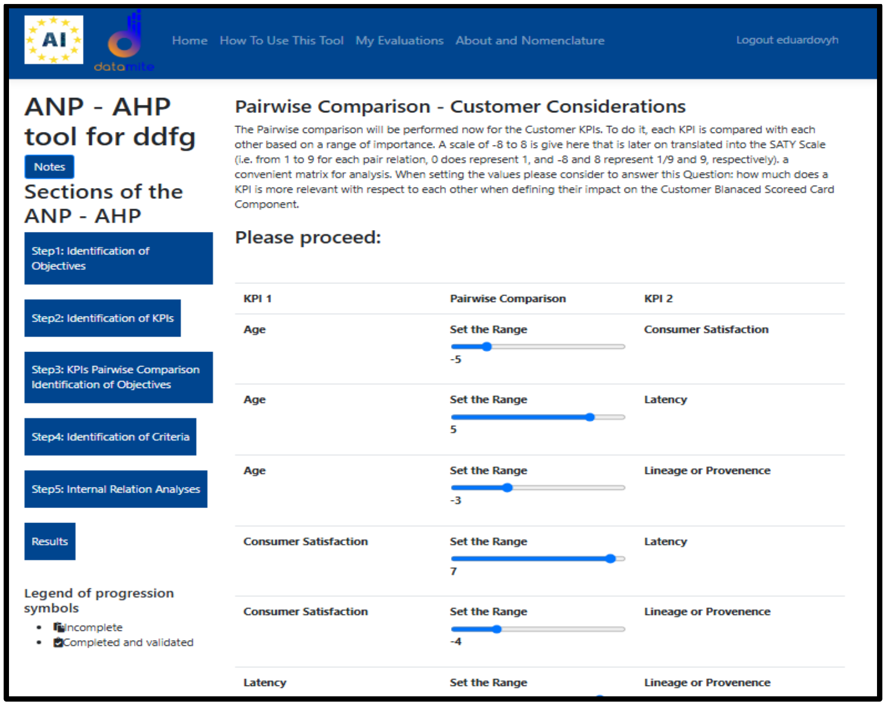
Plese select from the list the KPIs that are more relevant to you. If you want to add new KPIs, please use the add button named **Create new KPI**.

Name	Select	Driveer	Definition
Age	☑	Customer	Addresses refresh rate and available history
Aggregation Level	☑	Internal Processes	The degree to which data is or should be aggregate (if constructed from more than one set). This can be manipulated, nevertheless when a report contains more than one attribute, any metric on the report is calculated by default at the lowest attribute level th is on the report.
Compliance Cost	☑	Financial	Estiamted financial cost to buiseness of not keeping data for compliance / regulation puproses
Consumer Satisfaction	☑	Customer	Focusing on measuring the level of satisfaction consumers experience with the data they have purchased52. High levels of consumer satisfaction typically indicate that the data meets or exceeds the expectations of the consumers, suggesting that it is high quality, accurate, and valuable for their specifi needs53. This, in turn, positively influences the over valuation of the data, as satisfied consumers are mo likely to make repeat purchases, recommend the da to others, and are willing to pay a premium for high

Legend of progression

Figure 14 ANP tool Strategy and Metrics Definition – Step 1 (left) and 2 (right) are presented here.

(3) KPIs Pairwise Comparison – In this step, pairwise comparisons are provided by the user for metrics/KPIs within the same BSC category (e.g. Financial or Customer). Figure 15 illustrates this process. As for the pairwise comparisons of the BSC families, if discrepancies are detected in the pairwise comparisons of metrics/KPIs, the user is prompted to revise them. If no metric/KPI is selected in a whole BSC category, then the whole branch is eliminated. If only one metric/KPI is selected within a BSC category, then no pairwise comparison is required and the importance of the metric/KPI is set to the importance of the BSC category itself.


AI
datamite
Home
How To Use This Tool
My Evaluations
About and Nomenclature
Logout eduardovyh
Notes
Step1: Identification of Objectives
Step2: Identification of KPIs
Step3: KPIs Pairwise Comparison Identification of Objectives
Step4: Identification of Criteria
Step5: Internal Relation Analyses
Results

ANP - AHP tool for ddfg

Sections of the ANP - AHP

Legend of progression symbols

- Incomplete
- Completed and validated

Pairwise Comparison - Customer Considerations

The Pairwise comparison will be performed now for the Customer KPIs. To do it, each KPI is compared with each other based on a range of importance. A scale of -8 to 8 is give here that is later on translated into the SATY Scale (i.e. from 1 to 9 for each pair relation, 0 does represent 1, and -8 and 8 represent 1/9 and 9, respectively). a convenient matrix for analysis. When setting the values please consider to answer this Question: how much does a KPI is more relevant with respect to each other when defining their impact on the Customer Blanaced Scoreed Card Component.

Please proceed:

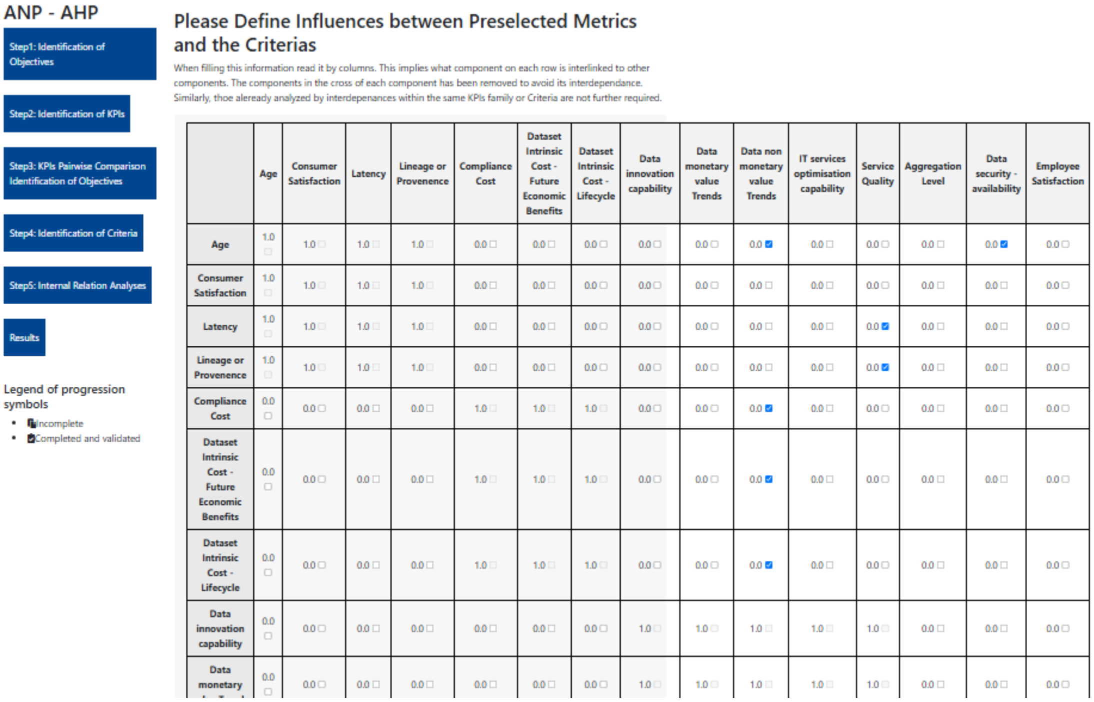
KPI 1	Pairwise Comparison	KPI 2
Age	Set the Range -5	Consumer Satisfaction
Age	Set the Range 5	Latency
Age	Set the Range -3	Lineage or Provenence
Consumer Satisfaction	Set the Range 7	Latency
Consumer Satisfaction	Set the Range -4	Lineage or Provenence
Latency	Set the Range	Lineage or Provenence

Figure 15 ANP tool - Pairwise Comparisons

(4) Identification of Criteria – In this step, the user selected the criteria that are the most relevant to their strategy. Once again, if multiple criteria are selected, pairwise comparisons need to be provided.

(5) Internal Relation Analyses – In this step, as shown in Figure 16, the user is asked to specify if there are interdependencies between metrics/KPIs of different BSC clusters. A matrix of checkboxes is displayed for the user to select which metrics/KPIs may influence each other. The interdependencies of metrics/KPIs pertaining to the same BSC cluster has already been determined in the previous steps and these checkboxes are thus automatically set. If interdependencies were selected, then new pairwise comparisons are performed and need to be provided by the user.

Funded by the European Union



ANP - AHP

Step1: Identification of Objectives
Step2: Identification of KPIs
Step3: KPIs Pairwise Comparison Identification of Objectives
Step4: Identification of Criteria
Step5: Internal Relation Analyses
Results

Legend of progression symbols

- Incomplete
- Completed and validated

Please Define Influences between Preselected Metrics and the Criterias

When filling this information read it by columns. This implies what component on each row is interlinked to other components. The components in the cross of each component has been removed to avoid its interdependance. Similarly, thoe aleready analyzed by interdepenances within the same KPIs family or Criteria are not further required.

	Age	Consumer Satisfaction	Latency	Lineage or Provenence	Compliance Cost	Dataset Intrinsic Cost - Future Economic Benefits	Dataset Intrinsic Cost - Lifecycle	Data innovation capability	Data monetary value Trends	Data non monetary value Trends	IT services optimisation capability	Service Quality	Aggregation Level	Data security - availability	Employee Satisfaction
Age	1.0 □	1.0 □	1.0 □	1.0 □	0.0 □	0.0 □	0.0 □	0.0 □	0.0 □	0.0 ☑	0.0 □	0.0 □	0.0 □	0.0 ☑	0.0 □
Consumer Satisfaction	1.0 □	1.0 □	1.0 □	1.0 □	0.0 □	0.0 □	0.0 □	0.0 □	0.0 □	0.0 □	0.0 □	0.0 □	0.0 □	0.0 □	0.0 □
Latency	1.0 □	1.0 □	1.0 □	1.0 □	0.0 □	0.0 □	0.0 □	0.0 □	0.0 □	0.0 □	0.0 □	0.0 ☑	0.0 □	0.0 □	0.0 □
Lineage or Provenence	1.0 □	1.0 □	1.0 □	1.0 □	0.0 □	0.0 □	0.0 □	0.0 □	0.0 □	0.0 □	0.0 □	0.0 ☑	0.0 □	0.0 □	0.0 □
Compliance Cost	0.0 □	0.0 □	0.0 □	0.0 □	1.0 □	1.0 □	1.0 □	0.0 □	0.0 □	0.0 ☑	0.0 □	0.0 □	0.0 □	0.0 □	0.0 □
Dataset Intrinsic Cost - Future Economic Benefits	0.0 □	0.0 □	0.0 □	0.0 □	1.0 □	1.0 □	1.0 □	0.0 □	0.0 □	0.0 ☑	0.0 □	0.0 □	0.0 □	0.0 □	0.0 □
Dataset Intrinsic Cost - Lifecycle	0.0 □	0.0 □	0.0 □	0.0 □	1.0 □	1.0 □	1.0 □	0.0 □	0.0 □	0.0 ☑	0.0 □	0.0 □	0.0 □	0.0 □	0.0 □
Data innovation capability	0.0 □	0.0 □	0.0 □	0.0 □	0.0 □	0.0 □	0.0 □	1.0 □	1.0 □	1.0 □	1.0 □	1.0 □	0.0 □	0.0 □	0.0 □
Data monetary	0.0 □	0.0 □	0.0 □	0.0 □	0.0 □	0.0 □	0.0 □	1.0 □	1.0 □	1.0 □	1.0 □	1.0 □	0.0 □	0.0 □	0.0 □

Figure 16 ANP tool – Interdependencies selection

(6) Results – After all the information is gathered, the supermatrix, and the limiting supermatrix can be computed. This supermatrix provides enough information to encapsulate the relative importance factors of each BSC component, each metric, and each criterion. More importantly, a direct connection of metrics to cluster (and sub cluster given the taxonomy) can be defined, which can be translated into agglomerated KPIs to represent the individual contribution of the different enterprise areas to achieve strategic goals.

The tool provides as a result a spider web diagram (see Figure 17) and a table (see Figure 18) that contain a visualisation of the relative importance of the selected metrics (diagram) and the impacts and metrics weight useful for KPIs estimation on each strategy and BSC component. These results can be automatically exported as a PDF report.

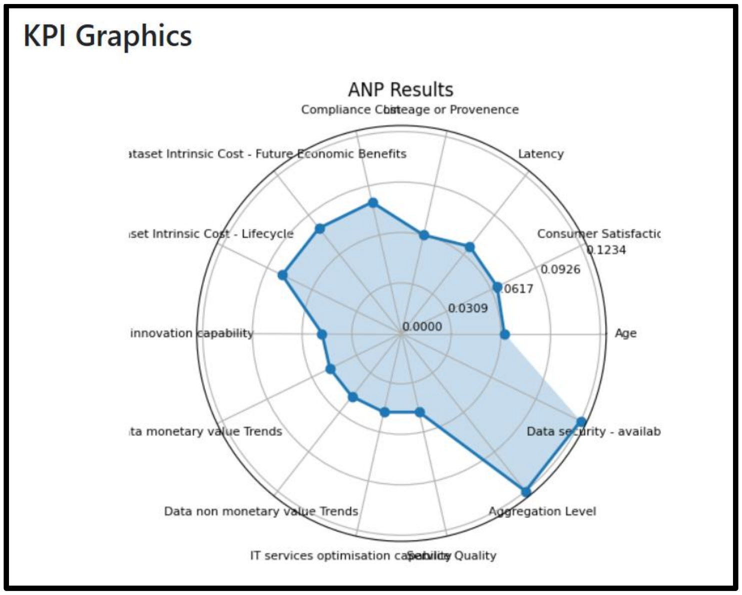
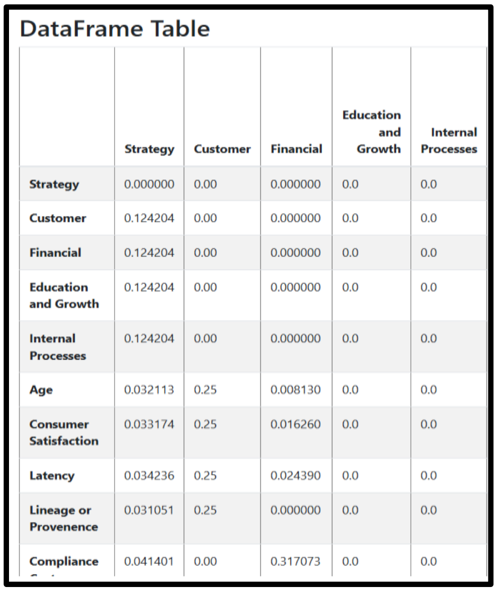


Figure 17 ANP tool – Example of a spider web diagram.


DataFrame Table

	Strategy	Customer	Financial	Education and Growth	Internal Processes
Strategy	0.000000	0.00	0.000000	0.0	0.0
Customer	0.124204	0.00	0.000000	0.0	0.0
Financial	0.124204	0.00	0.000000	0.0	0.0
Education and Growth	0.124204	0.00	0.000000	0.0	0.0
Internal Processes	0.124204	0.00	0.000000	0.0	0.0
Age	0.032113	0.25	0.008130	0.0	0.0
Consumer Satisfaction	0.033174	0.25	0.016260	0.0	0.0
Latency	0.034236	0.25	0.024390	0.0	0.0
Lineage or Provenence	0.031051	0.25	0.000000	0.0	0.0
Compliance	0.041401	0.00	0.317073	0.0	0.0

Figure 18 ANP tool – Example of a result table

Funded by
the European Union

Users can take further advantage of these results to connect them with DATAMITE service main Quality estimations to directly estimate relative KPIs useful for valuation of dataset, based on different approaches structured in T4.3 or use these relative KPIs as direct comparison values.

4.6 Validation and Future Work

4.6.1 Validation

Validation of the work performed in T4.2 was driven by two approaches:

- From one side to validate the metrics and KPIs with the technical components, i.e. validation of the systematic literature review.
- And from another part, validation of the tool that links BSC components, as key KPIs classes and sub-classes, though ANP, to the different metrics.

It is important to establish at this point that the tool constructed within DATAMITE is an additional component that could help users to extend the implementation of a BSC-based framework to estimate KPIs for their own strategies. This tool also can facilitate interaction with other components within WP4, since it can provide techniques to estimate contextualised measures of the dataset value. Furthermore, this value can later, through proper techniques such as benchmarking, be conceptualised as a ‘price’ proxy for the dataset.

For the validation of the first component, the resulting metrics, KPIs and taxonomy were shared with the technical partners (within WP2 and WP3) of DATAMITE. This information can be captured to improve outcomes from the quality metrics components. In fact, possible modifications could be included in the final product to incorporate final ANP process weights to generate global quality metrics of processed information.

For the validation of the second component, a workshop was performed within the consortium General Assembly - Aachen in which use cases and other consortium participants could evaluate the web tool directly. Since the tool focuses on the idea of containing all the information for users to interact with it, minimal help was provided during the validation process. The focus on this validation was not the technical part which has been widely ingrained already in literature. Instead, it is in the comprehension of the material shown, how easy and clear is the interaction with the service, and how relevant and understandable the main outcomes are.

Additionally, external stakeholder's evaluations were considered. Regretfully, a predominant insight from the first evaluation was a marked need for considerable improvement of the existing web platform. Despite sound technical performance and accurate output calculations, many users expressed difficulty in understanding the navigation structure and the interactions required to load datasets or switch between different dashboards. The interface design, while robust on the back end, lacked clarity in terms of visual guidance, button placement, and informational hierarchy in the frontend. In effect, participants from the first evaluation often resorted to trial-and-error to locate key functionalities such as generating specialised KPI, reports, or defining strategy components, leading to frustration and a slowed adoption rate.

Furthermore, participants highlighted that understanding the usability of an ANP process for metrics, although highly beneficial to experts, was overwhelming for new users. This resulted in confusion about which metrics to prioritise and how these metrics could be effectively combined to extract meaningful business insights. Moreover, the labelling of certain fields did not always match domain-specific terminology, compounding the perception of a steep learning curve.

As a result, external validation was delayed until modification of the frontend was performed and new components were included to facilitate interaction with the tool and interaction, in general, with outputs of WP4.

4.6.2 Future Work

An extended phase of T4.2 has been approved, further amplifying the scope of work originally envisioned for this task. This extended phase not only continues the development of the core tool but also introduces a new deliverable designed to unify insights from T4.1 up to T4.5. By consolidating the insights and methodologies from these tasks, T4.2 can expand its reach to accommodate additional components — most notably, Large Language Models (LLMs). The integration of LLMs holds considerable promise for enhancing user interactions and interpretability. For instance, natural-language queries could guide users through complex processes of KPI selection or data valuation procedures, thus contributing to a more intuitive experience. This expanded functionality would also create new opportunities for developing complementary tools, extending the scope of T4.2 beyond the current design.

In light of these expansions, user-friendliness remains a pivotal concern. As noted in the validation results, the tool's interface and feature set require thorough redesign to effectively support both novice and expert stakeholders. A new user-friendly interface, in particular for pairwise comparisons is currently being developed. Incorporating an LLM-based assistant could drastically simplify the user's journey, transforming multi-step tasks — such as uploading datasets or interpreting final KPI reports — into guided, conversation-like interactions. Early testing has indicated that such enhancements could be carried out in close collaboration with partners (FIR and 1001Lakes), whose involvement will help shape a more engaging and instructive evaluation process for new deployments of the tool.

In parallel, the Value Modelling approach is under active development. This methodology will provide a refined framework for connecting T4.2 metrics with the broader valuation and decision-making processes envisioned in T4.3. A dedicated session in Bilbao brought these concepts for refinements or alternative viewpoints. These discussions aimed to define a cohesive strategy that binds T4.2's data-centric metrics with T4.3's valuation techniques, ensuring both technical and methodological compatibility.

5 Conclusions

The work carried out in this document illustrates how T4.2 has achieved its primary goals of (1) establishing an objective approach to defining dataset value, (2) identifying and structuring existing approaches for the evaluation of datasets, (3) define criteria for the fair value evaluation of datasets and (4) offering mechanisms for computing and interpreting these indicators.

As described initially and through the document, all these objectives were achieved. In particular, the taxonomy introduced in Section 2, combined with the balanced scorecard approach, gives practitioners a straightforward way to classify and contextualise KPIs across domains such as finance, customer impact, internal processes, and learning and growth. This structure not only ensures that relevant metrics are aligned with high-level business strategies but also enables stakeholders to measure progress in a manner that is consistent, transparent, and actionable.

More specifically, and as previously mentioned, Objective (1) was achieved by establishing an aggregated framework and a taxonomy to objectively determine dataset value by linking key metrics – derived from strategies. Objective (2) involved conducting a systematic literature review on metrics and KPIs in data monetisation, which enabled us to identify and structure existing evaluation approaches and develop a taxonomy to correlate similar metrics, paving the way for enhanced valuation models in T4.3. Objective (3) focused on defining valuation criteria and establishing KPIs in categories and was achieved using the Balanced Scorecard framework, thus translating strategic objectives into actionable metrics for holistic evaluation. Finally, Objective (4) provided the mechanisms to compute these KPIs through a set of formulas and computational guidelines, offering a systematic, data-driven method for fair valuation (which are described in the annex as a table and has been provided to the technical components of DATAMITE).

A core achievement of T4.2 is the ANP-based web tool, which provides a practical means to navigate the KPI landscape, enabling organisations to prioritise metrics in line with their strategic aims. However, the validation process (Section 4.6.1) underscored the need for substantial refinements to enhance user-friendliness. The iterative feedback loop confirms that while the underlying analytic framework is sound, significant improvements are required to ease adoption by both technical and non-technical stakeholders. Future enhancements, including the potential integration of large language models (LLMs), are already under development, offering a promising route to more intuitive interfaces and conversational guidance.

6 References

[1] Davide Adami, Rosario G Garroppo, Stefano Giordano, and Stefano Lucetti. On synchronization techniques: performance and impact on time metrics monitoring. International Journal of Communication Systems, 16(4):273–290, 2003.

[2] Anish Agarwal, Munther Dahleh, and Tuhin Sarkar. A marketplace for data: An algorithmic solution. In Proceedings of the 2019 ACM Conference on Economics and Computation, pages 701–726, 2019.

[3] Salim Ahmad, Sanjeev Kumar, Munish Kumar, Rajiv Kumar, Minakshi Memoria, Amarjeet Rawat, Ashulekha Gupta, et al. The importance of quantifying financial returns on information system (is) investment for organizations: An analysis. In 2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON), volume 1, pages 197–200. IEEE, 2022.

[4] Haydar Akyurek, Cora Scholl, Regina Stodden, Tobias Siebenlist, and Agnes Mainka. Maturity and usability of open data in north rhinewestphalia. In Proceedings of the 19th Annual International Conference 138 on Digital Government Research: Governance in the Data Age, pages 1–10, 2018.

[5] Faiga Alawad and Frank Alexander Kraemer. Value of information in wireless sensor network applications and the iot: A review. IEEE Sensors Journal, 22(10):9228–9245, 2022.

[6] Muhammad Salek Ali, Massimo Vecchio, and Fabio Antonelli. A blockchain-based framework for iot data monetization services. The Computer Journal, 64(2):195–210, 2021.

[7] Paul Alpar and Sven Winkelstr¨ater. Assessment of data quality in accounting data with association rules. Expert Systems with Applications, 41(5):2259–2268, 2014.

[8] Santiago Andres Azcoitia, Marius Paraschiv, and Nikolaos Laoutaris. Computing the relative value of spatio-temporal data in data marketplaces. In Proceedings of the 30th International Conference on Advances in Geographic Information Systems, pages 1–11, 2022.

[9] Otmane Azeroual, Anastasija Nikiforova, and Kewei Sha. Overlooked aspects of data governance: Workflow framework for enterprise data deduplication. In 2023 International Conference on Intelligent Computing, Communication, Networking and Services (ICCNS), pages 65–73. IEEE, 2023.

[10] Donald P Ballou and Giri Kumar Tayi. Enhancing data quality in data warehouse environments. Communications of the ACM, 42(1):73–78, 1999.

[11] Carlo Batini, Cinzia Cappiello, Chiara Francalanci, and Andrea Maurino. Methodologies for data quality assessment and improvement. ACM computing surveys (CSUR), 41(3):1–52, 2009.

[12] Malika Bendechache, Judie Attard, Malick Ebiele, and Rob Brennan. A systematic survey of data value: Models, metrics, applications and research challenges. IEEE Access, 2023.

[13] Laure Berti-Equille. Data veracity estimation with ensembling truth discovery methods. In 2015 IEEE International Conference on Big Data (Big Data), pages 2628–2636. IEEE, 2015.

[14] Khadidja Bouchelouche, Abdessamed R´eda Ghomari, and Leila Zemmouchi-Ghomari. Enhanced analysis of open government data: Proposed metrics for improving data quality assessment. In 2022 5th International Symposium on Informatics and its Applications (ISIA), pages 1–6. IEEE, 2022.

[15] Rob Brennan, Judie Attard, and Markus Helfert. Management of data value chains, a value monitoring capability maturity model. 2018.

[16] Rob Brennan, Judie Attard, Plamen Petkov, Tadhg Nagle, and Markus Helfert. Exploring data value assessment: a survey method and investigation of the perceived relative importance of data value dimensions. 2019.

[17] John Byabazaire, Gregory O'Hare, and Declan Delaney. Using trust as a measure to derive data quality in data shared iot deployments. In 2020 29th International Conference on Computer Communications and Networks (ICCCN), pages 1–9. IEEE, 2020.

[18] Coral Calero, Mario Piattini, Carolina Pascual, and Manuel A Serrano. Towards data warehouse quality metrics. In DMDW, page 2, 2001.

[19] Riya Chakraborty, Lohit VijayaRenu, Zhenzhao Wang, and Praveen Killamsetti. Sparrow tracer: Scalable real time metrics from event log pipelines at twitter. In 2022 International Conference on Computational Science and Computational Intelligence (CSCI), pages 623–627. IEEE, 2022.

[20] Zhibo Cheng, Yanhua Wu, Taifeng Li, and Yanming Chen. Research on a hierarchical storage framework and algorithm of high-speed railway maintenance big-data based on multi metric features. In 2022 IEEE 5th International Conference on Big Data and Artificial Intelligence (BDAI), pages 93–98. IEEE, 2022.

[21] MV Chilukuri, Masliza Mohd Dahlan, and Chan Chuey Hwye. Benchmarking energy efficiency in tropical data centres–metrics and mesurements. In 2018 International Conference and Utility Exhibition on Green Energy for Sustainable Development (ICUE), pages 1–10. IEEE, 2018.

[22] Rene Bødker Christensen, Shashi Raj Pandey, and Petar Popovski. Semi-private computation of data similarity with applications to data valuation and pricing. IEEE Transactions on Information Forensics and Security, 18:1978–1988, 2023.

[23] Pin-Yu Chu and Hsien-Lee Tseng. A theoretical framework for evaluating government open data platform. In Proceedings of the International Conference on Electronic Governance and Open Society: Challenges in Eurasia, pages 135–142, 2016.

[24] J´ulia Colleoni Couto and Duncan Dubugras Ruiz. An overview about data integration in data lakes. In 2022 17th Iberian Conference on Information Systems and Technologies (CISTI), pages 1–7. IEEE, 2022.

[25] Istvan David and Dominik Bork. Infonomics of autonomous digital twins. In International Conference on Advanced Information Systems Engineering, pages 563–578. Springer, 2024.

[26] Leandro Batista de Almeida, Eduardo Cunha de Almeida, John Murphy, E Robson, and Anthony Ventresque. Bigdatanetsim: A simulator for data and process placement in large big data platforms. In 2018 IEEE/ACM 22nd International Symposium on Distributed Simulation and Real Time Applications (DS-RT), pages 1–10. IEEE, 2018.

[27] Shaleen Deep and Paraschos Koutris. Qirana: A framework for scalable query pricing. In Proceedings of the 2017 ACM International Conference on Management of Data, pages 699–713, 2017.

Funded by
the European Union

[28] Mario Jose Divan. Strategy for the data monetization in tune with the data stream processing. In 2017 6th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO), pages 81–91. IEEE, 2017.

[29] Akanksha Dixit, Arjun Singh, Yogachandran Rahulamathavan, and Muttukrishnan Rajarajan. Fast data: a fair, secure, and trusted decentralized iiot data marketplace enabled by blockchain. IEEE Internet of Things Journal, 10(4):2934–2944, 2021.

[30] Roy Dong, Lillian J Ratliff, Alvaro A C´ardenas, Henrik Ohlsson, and S Shankar Sastry. Quantifying the utility–privacy tradeoff in the internet of things. ACM Transactions on Cyber-Physical Systems, 2(2):1–28, 2018.

[31] Raghunadha Reddi Dornala, Sudhir Ponnapalli, Kalakoti Thriveni Sai, and Sreenu Bhukya. An enhanced data quality management system in cloud computing. In 2024 5th International Conference on Mobile Computing and Sustainable Informatics (ICMCSI), pages 788–796. IEEE, 2024.

[32] Corentin Dupont, Mehdi Sheikhalishahi, Federico M Facca, and Fabien Hermenier. An energy aware application controller for optimizing renewable energy consumption in data centres. In 2015 IEEE/ACM 8th International Conference on Utility and Cloud Computing (UCC),

pages 195–204. IEEE, 2015.

[33] European Central Bank. Banking supervision statistics framework. https://www.bankingsupervision.europa.eu/framework/ statistics/html/index.en.html, n.d. Accessed: 2025-01-17.

[34] Exasol. Extensibility in databases: The key to future-proofing your data strategy, 2025. Accessed: 29-Jan-2025.

[35] Raul Castro Fernandez, Pranav Subramaniam, and Michael J Franklin. Data market platforms: Trading data assets to solve data problems. arXiv preprint arXiv:2002.01047, 2020.

Funded by
the European Union

[36] Claudio Fiandrino, Dzmitry Kliazovich, Pascal Bouvry, and Albert Y Zomaya. Performance and energy efficiency metrics for communication systems of cloud computing data centers. IEEE Transactions on Cloud Computing, 5(4):738–750, 2015.

[37] Mike Fleckenstein, Ali Obaidi, and Nektaria Tryfona. A review of data valuation approaches and building and scoring a data valuation model. 2023.

[38] Julien Freudiger, Shantanu Rane, Alejandro E Brito, and Ersin Uzun. Privacy preserving data quality assessment for high-fidelity data sharing. In Proceedings of the 2014 ACM Workshop on Information Sharing & Collaborative Security, pages 21–29, 2014.

[39] LF Garifova. Infonomics and the value of information in the digital economy. Procedia economics and finance, 23:738–743, 2015.

[40] Lodovico Giaretta, Thomas Marchioro, Evangelos Markatos, and ˇSar¯unas Girdzijauskas. Towards a decentralized infrastructure for data marketplaces: narrowing the gap between academia and industry. In Proceedings of the 1st International Workshop on Data Economy, pages 49–56, 2022.

[41] Sreenivas Gollapudi and Debmalya Panigrahi. Fair allocation in online markets. In Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, pages 1179–1188, 2014.

[42] Sreenivas Gollapudi and Debmalya Panigrahi. Fair allocation in online markets. In Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, pages 1179–1188, 2014.

[43] Keith Grueneberg, Seraphin Calo, P Dewan, Dinesh Verma, and Tristan O'Gorman. A policy-based approach for measuring data quality. In 2019 IEEE International Conference on Big Data (Big Data), pages 4025–4031. IEEE, 2019.

[44] Naniek Utami Handayani, Wina Dara Kusuma, Zainal Fanani Rosyada, Yusuf Widharto, and Ajeng Hanifah. Usability evaluation of" inventory information system" design of disaster management in Yogyakarta province-indonesia. In Proceedings of the 2020 International Confer-

ence on Engineering and Information Technology for Sustainable Industry, pages 1–6, 2020.

[45] Junyi He, Qian Ma, Meng Zhang, and Jianwei Huang. Optimal fresh data sampling and trading. In 2021 19th International Symposium on Modeling and Optimization in Mobile, Ad hoc, and Wireless Networks (WiOpt), pages 1–8. IEEE, 2021.

[46] Bernd Heinrich, Diana Hristova, Mathias Klier, Alexander Schiller, and Michael Szubartowicz. Requirements for data quality metrics. Journal of Data and Information Quality (JDIQ), 9(2):1–32, 2018.

[47] Martin Hilgendorf, Vincenzo Gulisano, Marina Papatriantafilou, Jan Engstr¨om, and Binay Mishra. Forte: an extensible framework for robustness and efficiency in data transfer pipelines. In Proceedings of the 17th ACM International Conference on Distributed and Event-based

Systems, pages 139–150, 2023.

[48] Thi Thao Nguyen Ho and Barbara Pernici. A data-value-driven adaptation framework for energy efficiency for data intensive applications in clouds. In 2015 IEEE conference on technologies for sustainability

(SusTech), pages 47–52. IEEE, 2015.

[49] Xueyang Hu, Mingxuan Yuan, Jianguo Yao, Yu Deng, Lei Chen, Qiang Yang, Haibing Guan, and Jia Zeng. Differential privacy in telco big data platform. Proceedings of the VLDB Endowment, 8(12):1692–1703, 2015.

[50] Lu Huang. Research on evaluation system of digital rural management information system under internet technology. In Proceedings of the 2024 6th Asia Pacific Information Technology Conference, pages 1–9, 2024.

[51] Shiyue Huang, Ziwei Wang, Xinyi Zhang, Yaofeng Tu, Zhongliang Li, and Bin Cui. Dbpa: A benchmark for transactional database performance anomalies. Proceedings of the ACM on Management of Data, 1(1):1–26, 2023.

[52] Zhong Huang. Design and development of university information system based on mdm-a case study of the service satisfaction evaluation system. In Proceedings of the 2022 3rd International Conference on Internet and E-Business, pages 153–160, 2022.

[53] Jinho Hwang, Ahsen Uppal, Timothy Wood, and Howie Huang. Mortar: Filling the gaps in data center memory. In Proceedings of the 10th ACM SIGPLAN/SIGOPS international conference on Virtual execution environments, pages 53–64, 2014.

[54] Ilias Iliadis and Vinodh Venkatesan. Expected annual fraction of data loss as a metric for data storage reliability. In 2014 IEEE 22nd International Symposium on Modelling, Analysis & Simulation of Computer and Telecommunication Systems, pages 375–384. IEEE, 2014.

[55] Anne Immonen, Pekka Paakkonen, and Eila Ovaska. Evaluating the quality of social media data in big data architecture. Ieee Access, 3:2028–2043, 2015.

[56] Abdallah Jarwan, Ayman Sabbah, and Mohamed Ibnkahla. Information-oriented traffic management for energy-efficient and lossresilient iot systems. IEEE Internet of Things Journal, 9(10):7388–7403, 2021.

[57] Kenta Kanamori, Kota Tsubouchi, Junichi Sato, and Tatsuru Higurashi. Location yardstick: Calculation of the location data value depending on the users' context. In 2020 IEEE International Conference on Big Data (Big Data), pages 1545–1554. IEEE, 2020.

[58] Robert S Kaplan, David P Norton, et al. The balanced scorecard: measures that drive performance, volume 70. Harvard Business Review Boston, MA, USA, 2005.

[59] Staffs Keele et al. Guidelines for performing systematic literature reviews in software engineering, 2007.

[60] Vijay Khatri and Carol V Brown. Designing data governance. Communications of the ACM, 53(1):148–152, 2010.

[61] Igor Khokhlov and Leon Reznik. What is the value of data value in practical security applications. In 2020 IEEE Systems Security Symposium (SSS), pages 1–8. IEEE, 2020.

[62] Ihor Kozubtsov, Oleksii Silko, Lesia Kozubtsova, Mahmood Jawad Abu-AlShaeer, Laith S Ismail, and Mustafa Mohammmed Jassim. A method for calculating efficiency indicators of information security systems. In 2024 35th Conference of Open Innovations Association

(FRUCT), pages 388–398. IEEE, 2024.

[63] Pavel Krasikov and Christine Legner. A method to screen, assess, and prepare open data for use. ACM Journal of Data and Information Quality, 15(4):1–25, 2023.

[64] Sesillia Fajar Kristyanti, Tien Fabrianti Kusumasari, and Ekky Novriza Alam. Operational dashboard development as a data quality monitoring tools using data deduplication profiling result. In 2020 6th International Conference on Science and Technology (ICST), volume 1, pages

1–6. IEEE, 2020.

[65] Sylvain Kubler, Jeremy Robert, Yves Le Traon, Jurgen Umbrich, and Sebastian Neumaier. Open data portal quality comparison using ahp. In Proceedings of the 17th international digital government research conference on digital government research, pages 397–407, 2016.

[66] Manoj Kumar et al. Validation of data warehouse requirements-model traceability metrics using a formal framework. In 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom), pages 216–221. IEEE, 2015.

[67] Babak Lajevardi, Karl R Haapala, and Joseph F Junker. An energy efficiency metric for data center assessment. In IIE Annual Conference. Proceedings, page 1715. Institute of Industrial and Systems Engineers (IISE), 2014.

[68] Qi Li, Zhuotao Liu, Qi Li, and Ke Xu. martfl: Enabling utility-driven data marketplace with a robust and verifiable federated learning architecture. In Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security, pages 1496–1510, 2023.

[69] Tianyi Li, Gregorio Convertino, Ranjeet Kumar Tayi, and Shima Kazerooni. What data should i protect? recommender and planning support for data security analysts. In Proceedings of the 24th International Conference on Intelligent User Interfaces, pages 286–297, 2019.

[70] Xiao-Tong Li, Jun Zhai, Gui-Fu Zheng, and Chang-Feng Yuan. Quality assessment for open government data in china. In Proceedings of the 2018 10th International Conference on Information Management and Engineering, pages 110–114, 2018.

[71] Xijun Li, Jianguo Yao, Xue Liu, and Haibing Guan. A first look at information entropy-based data pricing. In 2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS), pages 2053–2060. IEEE, 2017.

[72] Z Li, M Wang, Z Xu, and G Tang. A comprehensive review of data centre integrated with renewable energy. 2022.

[73] Matteo Lia and Davide Damiano Colella. Ckanfair: a digital tool for assessing the fair principles. In 2023 IEEE International Conference on Big Data (BigData), pages 3980–3984. IEEE, 2023.

[74] Jinfei Liu, Jian Lou, Junxu Liu, Li Xiong, Jian Pei, and Jimeng Sun. Dealer: an end-to-end model marketplace with differential privacy. Proceedings of the VLDB Endowment, 14(6), 2021.

[75] Kecheng Liu, Hua Guo, Tao Wang, and Haotian Su. A semiotic framework for data asset valuation. In International Conference on Logistics, Informatics and Service Sciences, pages 878–887. Springer, 2023.

[76] Yongwen Liu, Moez Esseghir, and Leila Merghem Boulahia. Cloud service selection based on rough set theory. In 2014 International Conference and Workshop on the Network of the Future (NOF), pages 1–6. IEEE, 2014.

[77] Hsin-Ke Lu, Peng-Chun Lin, Chia-Hui Lo, and Mei-Yao Wu. A review of information system evaluation methods. In dalam International Conference on Software and Computer Applications (ICSCA 2012), Singapore, 2012.

[78] Puneet Mahajan. Textual data quality at scale for high dimensionality data. In 2022 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI), volume 1, pages 1–4. IEEE, 2022.

[79] Piyush Malik. Governing big data: principles and practices. IBM Journal of Research and Development, 57(3/4):1–1, 2013.

[80] Gomez-Omella Meritxell, Basilio Sierra, and Susana Ferreiro. On the evaluation, management and improvement of data quality in streaming time series. IEEE Access, 10:81458–81475, 2022.

[81] Robert Mieth, Juan M Morales, and H Vincent Poor. Data valuation from data-driven optimization. IEEE Transactions on Control of Network Systems, 2024.

[82] Francesco Paolo Nerini, Paolo Bajardi, and Andre Panisson. Value is in the eye of the beholder: A framework for an equitable graph data evaluation. In The 2024 ACM Conference on Fairness, Accountability, and Transparency, pages 467–479, 2024.

[83] Edobor Osagie, Mohammad Waqar, Samuel Adebayo, Arkadiusz Stasiewicz, Lukasz Porwol, and Adegboyega Ojo. Usability evaluation of an open data platform. In Proceedings of the 18th annual international conference on digital government research, pages 495–504, 2017.

[84] Taha Ouachani, Rachid Jahidi, and Bouchra Lebzar. Information system performance and evaluation: A theoretical review. The International Journal of Business Management and Technology, 7(1), 2022.

[85] PCMag. 5 green computing metrics, 2025. Accessed: 2025-02-04.

[86] Dirk Pesch, Alan McGibney, Piotr Sobonski, Susan Rea, Thomas Scherer, L Chen, Ton Engbersen, Deepak Mehta, Barry O'Sullivan, Enric Pages, et al. The genic architecture for integrated data centre energy management. In 2015 IEEE/ACM 8th International Conference on Utility and Cloud Computing (UCC), pages 540–546. IEEE, 2015.

[87] Kai Petersen, Robert Feldt, Shahid Mujtaba, and Michael Mattsson. Systematic mapping studies in software engineering. In 12th International Conference on Evaluation and Assessment in Software Engineering (EASE) 12, pages 1–10, 2008.

[88] Kai Petersen, Sairam Vakkalanka, and Ludwik Kuzniarz. Guidelines for conducting systematic mapping studies in software engineering: An update. Information and software technology, 64:1–18, 2015.

[89] Mark Petticrew and Helen Roberts. Systematic reviews in the social sciences: A practical guide. John Wiley & Sons, 2008.

[90] VJS Pranavasri, Leo Francis, Ushasri Mogadali, Gaurav Pal, SVSLN Surya Suhas Vaddhiparthy, Anuradha Vattem, Karthik Vaidhyanathan, and Deepak Gangadharan. Scalable and interoperable distributed architecture for iot in smart cities. In 2023 IEEE 9th World Forum on Internet of Things (WF-IoT), pages 01–06. IEEE, 2023.

[91] Minfeng Qi, Zhiyu Xu, Ziyuan Wang, Shiping Chen, and Yang Xiang. Deda: A defi-enabled data sharing and trading system. In Proceedings of the Fourth ACM International Symposium on Blockchain and Secure Critical Infrastructure, pages 47–57, 2022.

[92] Alfonso Quarati and Monica De Martino. Open government data usage: a brief overview. In Proceedings of the 23rd international database applications & engineering symposium, pages 1–8, 2019.

[93] Vigan Raca, Goran Velinov, Betim Cico, and Margita Kon-Popovska. Measuring the government openness using an assessment tool: Case study of six western balkan countries. In 2021 10th Mediterranean Conference on Embedded Computing (MECO), pages 1–5. IEEE, 2021.

[94] Divya Rao and Wee Keong Ng. Information pricing: A utility based pricing mechanism. In 2016 IEEE 14th Intl Conf on Dependable, Autonomic and Secure Computing, 14th Intl Conf on Pervasive Intelligence and Computing, 2nd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech), pages 754–760. IEEE, 2016.

[95] V Dinesh Reddy, Brian Setz, G Subrahmanya VRK Rao, GR Gangadharan, and Marco Aiello. Metrics for sustainable data centers. IEEE Transactions on Sustainable Computing, 2(3):290–303, 2017.

[96] Xiaoqi Ren, Palma London, Juba Ziani, and Adam Wierman. Joint data purchasing and data placement in a geo-distributed data market. In Proceedings of the 2016 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Science, pages 383–384, 2016.

[97] Xiaoqi Ren, Palma London, Juba Ziani, and Adam Wierman. Datum: Managing data purchasing and data placement in a geo-distributed data market. IEEE/ACM Transactions on Networking, 26(2):893–905, 2018.

[98] Shazia Sadiq, Tamraparni Dasu, Xin Luna Dong, Juliana Freire, Ihab FIlyas, Sebastian Link, Miller J Miller, Felix Naumann, Xiaofang Zhou, and Divesh Srivastava. Data quality: The role of empiricism. ACM SIGMOD Record, 46(4):35–43, 2018.

[99] Patrick Katuta Samuel, Peter Musau Moses, Segera Davies, and Cyrus Wekesa. Analysis of energy utilization metrics as a measure of energy efficiency in data centres: Case study of wananchi group (kenya) limited data centre. In 2022 IEEE PES/IAS PowerAfrica, pages 1–5. IEEE, 2022.

[100] Nirmal B Satyendra, Neeraja K Swami, and Priyanka V Bhailume. Evaluation of banking standards to ascertain their suitability for building data models for big data based data lake for banking domain.In 2020 IEEE International Conference on Technology, Engineering, Management for Societal impact using Marketing, Entrepreneurship and Talent (TEMSMET), pages 1–7. IEEE, 2020.

[101] Sdia. Data center metrics, 2025. Accessed: 2025-02-04.

[102] Mohamed Adel Serhani, Hadeel T El Kassabi, Ikbal Taleb, and Al-ramzana Nujum. An hybrid approach to quality evaluation across big data value chain. In 2016 IEEE International Congress on Big Data (BigData Congress), pages 418–425. IEEE, 2016.

[103] Flavia Serra, Veronika Peralta, Adriana Marotta, and Patrick Marcel. Use of context in data quality management: a systematic literature review. ACM Journal of Data and Information Quality, 16(3):1–41, 2024.

[104] Nigel Shadbolt, Kieron O’Hara, Tim Berners-Lee, Nicholas Gibbins, Hugh Glaser, Wendy Hall, et al. Linked open government data: Lessons from data. gov. uk. IEEE Intelligent Systems, 27(3):16–24, 2012.

[105] Raunak Shah, Koyel Mukherjee, Atharv Tyagi, Sai Keerthana Karnam, Dhruv Joshi, Shivam Pravin Bhosale, and Subrata Mitra. R2d2: Reducing redundancy and duplication in data lakes. Proceedings of the ACM on Management of Data, 1(4):1–25, 2023.

[106] Ting Shao, Peiting Yang, and Hongbo Jiang. Evaluation of users’ participation in value co-creation of open government data platform. In Proceedings of the 2023 7th International Conference on E-Commerce, E-Business and E-Government, pages 51–57, 2023.

[107] Christian Sillaber, Andrea Mussmann, and Ruth Breu. Experience: Data and information quality challenges in governance, risk, and compliance management. Journal of Data and Information Quality (JDIQ), 11(2):1–14, 2019.

[108] Tanu Singh and Manoj Kumar. Empirical validation of requirements traceability metrics for requirements model of data warehouse using svm. In 2020 IEEE 17th India Council International Conference (INDICON), pages 1–5. IEEE, 2020.

[109] CN Sowmyarani, LG Namya, GK Nidhi, and P Ramakanth Kumar. Score, arrange and cluster: A novel clustering-based technique for privacy preserving data publishing. IEEE Access, 2024.

[110] Christoph Stach, Julia Bracker, Rebecca Eichler, Corinna Giebler, and Bernhard Mitschang. Demand-driven data provisioning in data lakes: Barents—a tailorable data preparation zone. In The 23rd International Conference on Information Integration and Web Intelligence, pages 187–198, 2021.

[111] Jacques B Stander. The modern asset: big data and information valuation. PhD thesis, Stellenbosch: Stellenbosch University, 2015.

[112] Ya Su, Youjian Zhao, Wentao Xia, Rong Liu, Jiahao Bu, Jing Zhu, Yuanpu Cao, Haibin Li, Chenhao Niu, Yiyin Zhang, et al. Coflux: robustly correlating kpis by fluctuations for service

troubleshooting. In Proceedings of the International Symposium on Quality of Service, pages 1–10, 2019.

[113] D Sudharson, P Divya, K Saranya, Aman Kumar Dubey, Shreya Vijay, and G Mayuri. A novel ai framework for assuring data sustainability in health care dataset. In 2023 Third International Conference on Ubiquitous Computing and Intelligent Information Systems (ICUIS),

pages 161–166. IEEE, 2023.

[114] Youngjung Suh. Exploring the impact of data quality on business performance in crm systems for home appliance business. IEEE Access, 2023.

[115] Peng Sun, Guocheng Liao, Xu Chen, and Jianwei Huang. A socially optimal data marketplace with differentially private federated learning. IEEE/ACM Transactions on Networking, 2024.

[116] Sunbird. Top 30 data center sustainability metrics, 2025. Accessed: 2025-02-04.

[117] Amir Taherkordi, Frank Eliassen, and Geir Horn. From iot big data to iot big services. In Proceedings of the symposium on applied computing, pages 485–491, 2017.

[118] Nora Taibouni and Rachid Chalal. A toolbox for information system evaluation. In Proceedings of the 2nd International Conference on Big Data Technologies, pages 283–290, 2019.

[119] Luis Tom´as and Johan Tordsson. Cloud service differentiation in overbooked data centers. In 2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing, pages 541–546. IEEE, 2014.

[120] Anh Duy Tran, Somjit Arch-int, and Ngamnij Arch-int. Measures of dependency in metric decision systems and databases. In 2017 International Conference on Soft Computing, Intelligent System and Information Technology (ICSIIT), pages 107–113. IEEE, 2017.

[121] Mohammed Joda Usman, Abdul Samad Ismail, Gaddafi Abdul Salaam, Hassan Chizari, Omprakash Kaiwartya, Abdulsalam Yau Gital, Muhammed Abdullahi, Ahmed Aliyu, and Salihu

Idi Dishing. Energy-efficient nature-inspired techniques in cloud computing datacenters. Telecommunication Systems, 71:275–302, 2019.

[122] Saumitra Vatsal and Shalini Agarwal. Energy efficiency metrics for safeguarding the performance of data centre communication systems by green cloud solutions. In 2019 International Conference on Cuttingedge Technologies in Engineering (ICon-CuTE), pages 136–140. IEEE, 2019.

[123] Andreea Valeria Vesa, Tudor Cioara, Ionut Anghel, Marcel Antal, Claudia Pop, Bogdan Iancu, Ioan Salomie, and Vasile Teodor Dadarlat. Energy flexibility prediction for data center engagement in demand response programs. Sustainability, 12(4):1417, 2020.

[124] Gianluigi Viscusi and Carlo Batini. Digital information asset evaluation: Characteristics and dimensions. In Smart Organizations and Smart Artifacts: Fostering Interaction Between People, Technologies and Processes, pages 77–86. Springer, 2014.

[125] Alex X Wang, Colin R Simpson, and Binh P Nguyen. Enhancing data governance through data-centric ai: Case study in new Zealand government sector. In 2024 16th International Conference on Computer and Automation Engineering (ICCAE), pages 64–69. IEEE, 2024.

[126] Chen Wang, Jialin Qiao, Xiangdong Huang, Shaoxu Song, Haonan Hou, Tian Jiang, Lei Rui, Jianmin Wang, and Jiaguang Sun. Apache iotdb: A time series database for iot applications. Proceedings of the ACM on Management of Data, 1(2):1–27, 2023.

[127] Weina Wang, Lei Ying, and Junshan Zhang. The value of privacy: Strategic data subjects, incentive mechanisms and fundamental limits. In Proceedings of the 2016 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Science, pages 249–260, 2016.

[128] Viola Wenz, Arno Kesper, and Gabriele Taentzer. Clustering heterogeneous data values for data quality analysis. ACM Journal of Data and Information Quality, 15(3):1–33, 2023.

[129] Nugroho Wibisono, Mohammad Amin Soetomo, Heru Purmono Ipung, Marastika Wicaksono Aji Bawono, and Eka Budiarto. Data integration readiness analysis in merged

telecommunication company. In 2021 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS, pages 317–322. IEEE, 2021.

[130] Haocheng Xia, Jinfei Liu, Jian Lou, Zhan Qin, Kui Ren, Yang Cao, and Li Xiong. Equitable data valuation meets the right to be forgotten in model markets. Proceedings of the VLDB Endowment, 16(11):3349–3362, 2023.

[131] Tao Xiaoming, Wang Yu, Peng Jieyang, Zhao Yuelin, Wang Yue, Wang Youzheng, Hu Chengsheng, and Lu Zhipeng. Data component: An innovative framework for information value metrics in the digital economy. China Communications, 21(5):17–35, 2024.

[132] Anran Xu, Zhenzhe Zheng, Qinya Li, Fan Wu, and Guihai Chen. Vap: Online data valuation and pricing for machine learning models in mobile health. IEEE Transactions on Mobile Computing, 2023.

[133] Anran Xu, Zhenzhe Zheng, Fan Wu, and Guihai Chen. Online data valuation and pricing for machine learning tasks in mobile health. In IEEE INFOCOM 2022-IEEE Conference on Computer Communications, pages 850–859. IEEE, 2022.

[134] Mehdi Yalaoui and Saida Boukhedouma. A survey on data quality: principles, taxonomies and comparison of approaches. In 2021 International Conference on Information Systems and Advanced Technologies (ICISAT), pages 1–9. IEEE, 2021.

[135] Ynvolve. Mastering data center efficiency: The power metrics, 2025. Accessed: 2025-02-04.

[136] Ziba Yusufi, Simon J Preis, Daniel Kraus, Udo Kruschwitz, and Bernd Ludwig. Data value assessment in semiconductor production. 2022.

[137] Jing Zhang and Xiao-Ping Liu. Evaluation of integrated logistics information system based on perception. In Proceedings of the 2018 1st International Conference on Internet and e-Business, pages 319–323, 2018.

[138] Diego Rodrigues, Radu Godina, Pedro Espadinha de Cruz. Key Performance Indicators Selection through an Analytical Network Process Model for Tooling and Die Industry. Sustainability, 2021, 13, 13777.

● Appendix A: KPIs, Metrics, and its Definitions

KPIs Calculations

Generally speaking, any metric can become a KPI as long as its delimited between maximum and minimum ranges. These ranges are normally linked to targets or benchmarks that facilitate the definition of the KPI (e.g. KPIi = $M_{current}$ $-M_{minimum}$/($M_{benchmark}$ $-M_{minimum}$)), where M is a metric i. Furthermore, KPIs Can be calculated as a ponder between different components that allows its full estimation. For example, accessibility can be defined based on Availability, Usability, Interoperability, Inclusivity, and Security and Permissions (see table). Thus a weighted sum of each factor (i.e KPI = Σ wi $*$ KPI_i +Σ wj $*Metrics_j$.), would allow the estimation of the defined KPI. Furthermore, selective factoring such as: Accessibility = max (0; 1-(Delivery time – Request time)/(Deadline time - Request time)) can be employed. Finally, KPIs can be calculated based on soft approaches such as questionnaires. For example, for platform accessibility (or a component of accessibility such as inclusivity, if desired to be considered), Users can rate based on search clarity, navigation simplicity, and the ability to retrieve data efficiently (or other dependent factors). When estimating KPIs and metrics, score approaches can be used such as the following example for Accuracy : - Use predefined accuracy values based on the quality of data: - 0.2: Poor unusable data - 0.4: Substandard data quality - 0.6: Usable and functional data quality - 0.8: Excellent data quality - 1: Perfect data quality.

The nature of the ranges and scores can be technical or non, nevertheless, they have to be objectively regulated within a range of 0 and 1 in order to be agglomerated into a general KPI or being used as its own.

Metrics

Metrics are normally measured directly from data or estimated based on a combination of collected information and proper transformation by specific approaches, equations, and algorithms. Generally, to transform a metric into a range, is necessary to know (or define): Minimum, Maximum, Differential, or Target values (target can assume different names based on the domain, e.g. Request).

Thus, each metric can have in addition to the pondered value, an assigned limiting value such as those defined. (e.g. Target Accuracy). By defining these last, a translation into KPIs can be achieved. Metrics (and some KPIs) can be defined in function of an efficiency term. Efficiency can be defined as the ratio of output (performance or value) to input (cost, time, or resources). It essentially measures how effectively a process, system, or resource performs relative to what is invested in it. We can define a general efficiency KPI that integrates multiple aspects of cost, accessibility, availability, timeliness, and scalability as a withed added value of other efficiencies like operational efficiency, energy efficiency, efficiency in handling big data, etc.

Node

A Node is a data. Is designed to convert data into valuable assets. It consists of various parts, such as information and its quality, which collectively serve a specific function. Like physical assets, Nodes have associated costs and the potential to generate value. They provide structure and accountability to data, bridging the gap between raw information and actionable insights for an organisation. The nodes are composed of 8 Ranges The decision, the required information, the value range, the required ROI, lifecycle, maximum cost, accuracy, and frequency. Check [117] For more information and definition of a Node.

Appendix B: Metrics and KPIs Estimation

Name	Secondary Name	Ref.	BSC	Definition	Equation (if applicable)
Access Cost	Bandwidth Cost	97, 105	Data Valuation Technique	Data valuation technique which measures access cost based on historical logs and potential cloud pricing, influenced by bandwidth in geo-distributed systems.	$Access\ Cost = A \cdot r$ - Where A is the number of accesses, and r is the read cost per unit. Similarly, $Bandwidth\ Cost = D \cdot t$ - Where: D: Data transferred (e.g., in GB or TB) and t: Cost per unit of data transferred (e.g., cost per GB).
Access Frequency	Number of Requests, (directly linked to Views), Arrival Rate, Access Interval, number of accesses, usage over time	12, 15, 20, 37, 48, 92, 111, 112	Customer Needs and Satisfaction	Overall demand for platform content, capturing the number of times datasets, services, or features are accessed, irrespective of whether by single or multiple users. Also, total count of operations, transactions, or queries made to a system within a given period.	Can be calculated based on metrics such as the number of requests, views, arrival rate, access intervals, and usage over time. The total count of requests or views in a specified period is: $Access\ Frequency = \sum_{i=1}^{n} Request_i$ The arrival rate, representing request per unit time, is given by: $\lambda = {Total\ Request\ in\ Period}/_{\Delta t}$ Access Interval, indication the average time between consecutive accesses can be calculated based on previous ones. $Access\ Interval = \frac{\Delta t}{Access\ Frequency}$
Accessibility	Search, Functionality, Navigation and Link	11, 14, 17, 23, 93, 97, 105, 124, 134,	Data Valuation Technique & Governance and Compliance & Tech. and Infra.	Accessibility, in the context of data and information systems, refers to the ease with which users can access and utilise data. It encompasses several key aspects: **Availability**: Data should be readily available to authorised users when needed, without unnecessary delays or barriers. **Usability**: The data should be presented in a format that is easy to understand and navigate, allowing users to efficiently find and use the information they need. **Interoperability**: Data should be compatible with various systems and applications, enabling seamless integration and exchange of information across different platforms. **Inclusivity**: Accessibility also considers the needs of diverse user groups, including those with disabilities. This means ensuring that data and information systems are designed to be usable by individuals with varying abilities, such as providing alternative formats or assistive technologies. **Security and Permissions**: While ensuring accessibility, it is also important to manage access controls and permissions to protect sensitive data from unauthorised access. **Open to use / licensing**: There is no limit on users to use the data **Format**: Follow pre-structured definitions for the data content structure, for example, does the dataset match Berners-Lee's "linked data principles"?	KPI - Estimation based on key components as a KPI. For example, following [65], accessibility can be estimated by its: **Open to Use**: within this scale: 0 = Not open to use (need to apply to the responsible department); 5 = Use the data with limitations; 10 = The data is open to use without limitations; **Automatic Reading capability**: Using this scale: 0 = Automatic reading language or API format is not provided; 5 = proportion of datasets with APIs between 0% and 50%; 10 = proportion of datasets with APIs between 50% and 75%; 15 = proportion of datasets with APIs is more than 75%; **Search functionality**: 0 = With the "search" button, but couldn't find the subject matter; 5 = There are matching datasets, but still some irrelevant datasets; 10 = key-in the name of the subject matter and we could search the datasets successfully by clicking the "search" button; 15 = Besides what mentioned above, the function of word association enables users to search efficiently **Format**: 0 = None of the datasets is downloadable; 5 = Datasets that are in PDF, JPEG, or other specific format; 10 = Users could use the application to read structured data such as EXCEL files; 15 = Users could get access to the datasets that are non-proprietary such as CSV or XML files; 20 = Users could use the format standard; for example, URIs are used to express data in order to let users understand the location of data in the data network, also people/machines could directly access, save, apply every single data in the datasets; 25 = Besides 4-star, users could link the data to other people's data as an extension of related content.

Funded by the European Union

				Overall, accessibility is crucial for maximizing the value of data, as it ensures that users can effectively leverage the information for decision-making, analysis, and other purposes. Measures how accessible the platform is for users in terms of ease of navigation, search, and data retrieval.	Other components are described in [65] so review as needed.
Accuracy		111, 39, 11, 10, 70, 83, 46, 60, 127, 107, 82, 40, 68, 2, 35, 137, 118, 23, 124, 77, 113, 9, 125, 55, 13, 102, 17, 80, 43	Data Quality	Accuracy in the context of scientific analysis is a comprehensive measure of how closely data or results align with the true, intended, or expected values. It encompasses multiple facets that ensure the reliability, applicability, and validity of the information within the specific scientific domain. The accuracy of data or results is not only about correctness but also about their overall suitability for achieving the desired objectives in a given context. The complexity of the relationship between accuracy and the value gained depends on the type of information and the organisation. Moreover, unless specified, the basic calculation of value is obtained by meeting the required accuracy. As defined by [12], Accuracy depends on Range, Consistency, Typicality and Moderation Metrics. Nevertheless, different key elements can also be used to explain accuracy: **Precision**: Refers to the granularity or level of detail in the data. In science, precision ensures that measurements are not only correct but also detailed enough to provide meaningful insights. For example, reporting a temperature as 37.456°C instead of rounding to 37°C can be crucial in laboratory experiments. **Timeliness**: Information must reflect the most current state. In scientific research, outdated data can lead to incorrect conclusions, making the timeliness of updates or recalibrations essential. **Relevance**: Accuracy is context dependent. Data must be both correct and directly applicable to the research question or hypothesis. For instance, accurate but irrelevant data adds no value to the scientific inquiry. **Completeness**: Scientific accuracy requires that no critical data points are missing. Incomplete datasets can lead to skewed or biased results, undermining the validity of findings. **Traceability**: The provenance of data ensures it can be verified and validated. For science, traceability means being able to track data back to its source, such as instrumentation or observation records, to confirm its authenticity. **Tolerance and Range**: Recognises that minor inaccuracies or deviations may be acceptable within specified thresholds. Scientific accuracy involves defining these thresholds to minimise the impact on overall validity. Furthermore, in terms of range it specifies limits for acceptance too. **Context Dependence**: Accuracy must meet the specific standards and expectations of the scientific discipline. For example, a minor deviation in physics experiments might be	To calculate accuracy, the following methods can be used, incorporating explicit details from the reviewed texts: **Exact Accuracy Calculation**: Define the tolerance and resolution of the data acquisition process: Example: For a thermometer with a tolerance of ±0.1°C and a required accuracy of ±1°C: $Accuracy = 1 - \frac{Tolerance}{Requirement} = 1 - \frac{0.1}{1} = 0.9$ This yields an accuracy of 0.9 or 90%. **Estimated Accuracy Based on Data Quality**: Assign predefined values to estimate accuracy qualitatively: 0.2: Poor, unusable data; 0.4: Substandard quality; 0.6: Usable, functional data; 0.8: Excellent quality; 1.0: Perfect data Weighted KPI for Accuracy Evaluate accuracy by considering its key components (e.g., range, consistency, typicality, moderation) and assign weights to each component. Calculate each component's contribution using: $Accuracy = 1 - (Ceiling - Required) \cdot Value\ Gain\ Rate$. Aggregate the weighted values for an overall accuracy score, as detailed in [80]. **Correctness-Based Accuracy**: Calculation using the formula from [46] to calculate accuracy based on correctness: $DQ = 1/(d(w, wm) + 1)$, where: w: Stored data value wm: Real-world value d: Distance measure, such as the number of differing positions in a data string. **Metrics from Confidence Intervals**: As described in [80], accuracy can also be assessed using statistical confidence intervals: Range: Proportion of data points within predefined valid intervals. Consistency: Proportion of values within an 80% confidence interval of a Gaussian distribution. Typicality: Proportion of data points within a 95% confidence interval. Moderation: Proportion of values within a 99% confidence interval. **Accuracy through Deviation Metrics**: In [70], accuracy is linked to deviation analysis: Calculate the deviation of data values from predefined benchmarks. Use statistical measures such as mean absolute error (MAE) or root mean square error (RMSE) to quantify deviation-based accuracy. **Application-Specific Thresholds**: Can be described as context-dependent, with thresholds defined by application needs. Specify acceptable tolerance levels based on: (1) The criticality of the data's role in decision-making. (2) Domain specific accuracy standards.

				acceptable, whereas in clinical trials, even small inaccuracies can have serious consequences. **Consistency**: Data should be reliable and uniform across various trials or applications. Inconsistencies in repeated experiments can signal issues with data accuracy. **Typicality**: Measures whether the data reflects typical or expected conditions. Atypical data may require additional validation to ensure its accuracy. **Moderation**: Ensures data is not skewed or biased toward extremes. Balanced data contributes to overall scientific accuracy by avoiding distortions.	
Adaptability	Versatility, Flexibility	117, 110, 44, 118, 124, 100	Operational Efficiency	Relates to how well a system can adjust to changes in its environment, such as varying network conditions, changing user needs, or different operating contexts. Adaptable systems make scalability more efficient because they can dynamically adjust their resources and behaviour to accommodate growth.	KPI or Metric - Depending on the nature of the system under consideration, adaptability can be seen as a KPI (and thus method to estimate it) that can be combined from different key factors including: **Time to Adapt (TA)** - Time to make a system functional under new requirements. **Success Rate (SR)** - Percentage of successful transitions to new conditions without failure. **Cost of Adaptation (CA)** - Cost (in time, effort, or resources) incurred to accommodate changes. **Number of Support (NUS)** - The variety of functionalities, legacy applications, languages, formats, or other subjective considerations that the system supports. **Performance Consistency (PC)** - Ability to maintain performance. **Effort for Transition (ET)** - The ease of switching between use cases. **Range of Customisation (RC)** - The breadth of adjustable parameters or features. **Ease of Configuration (EC)** - How quickly and effortlessly users can implement changes. **Impact of Changes (IC)** - The degree to which adjustments disrupt system operations.
Adaptability Power Curve (APC)	Adaptation of Data Centre to Available Renewable Energy (APCren)	86,123	Operational Efficiency	Adaptability of a data centre's power consumption by quantifying the deviation of actual monitored power from the baseline power profile over a given time interval. Similarly, APCren can be seen as a modified version that looks for the renewable energy sources influences.	$APC = \frac{\sum_{start}^{end} \lvert P_{monitored} - P_{baseline} \rvert}{\sum_{start}^{end} P_{baseline}}$ Where P is the power consumed on the respective interval.
AEUF		95, 101	Operational Efficiency	Assess how effectively a data centre employs its airside economiser system for "free" cooling.	AEUF = Time Air Economiser in Use/Total Time
Age	Data Creation, Time Index, Age of Information	41, 45, 5	Data Quality	Refers to how recent or old the data is. The age of data can impact its relevance and accuracy, especially in fast-moving fields like technology or finance.	Direct measurement or formally, Age at time t is defined as: $\Delta(t) = t - H_t$ - where H_t denotes the timestamp of the most recent data update prior to t.
Airflow Efficiency (AFE)		116	Operational Efficiency	How efficiently air moves from the supply to the return.	$AFE = \frac{Fan\ Power}{Airflow}$

Funded by the European Union

Annual Fraction of Data Loss (AFDL)		54	Data Quality	Measures the fraction of stored data lost annually, assessing reliability in distributed systems.	Direct measurement.
Availability	Retrievability	39, 65, 23, 50, 76, 84, 93, 55	Operational Efficiency	Degree to which data is accessible and retrievable when needed, by authorised users or systems. It ensures that the data can be accessed in a timely manner without interruption, and is crucial for maintaining smooth operations, decision-making, and service continuity. As a KPI it could have the following Key Aspects: **Accessibilit**y: Ensuring that the data is available to authorised users when they need it, whether for operational or analytical purposes. (i.e. a Boolean per dataset) **Reliability** (Measured through System Robustness): The data should be consistently available without frequent downtime or failures. **Fault tolerance** (Measured through System Robustness): Systems storing the data should be resilient to hardware or software failures, ensuring minimal disruption in data access. **Backup**: Implementing regular backups and quick recovery solutions ensures that data is still available after incidents like data corruption or system failures. **Redundancy**: Data availability can be enhanced through redundant systems (e.g., having multiple copies of the data across servers), so if one system fails, another is still operational.	Data Availability as a metric is typically expressed as a percentage and can be calculated, simplistically, using the following formula: $Availability = \frac{Uptime}{Total\ Time} \cdot 100$ Where: Uptime: The total time during which the data was accessible and usable by users. Total Time: The entire period under consideration (e.g., a month, year). Retrievability, on the other hand, is measured based on intrinsic operations related to data access, such as the success rate of HTTP GET operations, which determine whether the data can be retrieved without errors or restrictions [65]. As a more general metric / KPI, it can be measured by its components previously mentioned.
Availability, Capacity, and Efficiency (ACE)	-	95	Operational Efficiency	In the context of Heating, Ventilation, and Air Conditioning (HVAC) systems, performance is often evaluated based on metrics such as Availability, Capacity, and Efficiency. General expressions can be used to track ACE.	General expressions already mentioned for system Availability and Capacity can be used. Regarding Efficiency, alternative metrics such as Seasonal Energy Efficiency Ratio (SEER), Heating Seasonal Performance Factor (HSPF), and Annual Fuel Utilisation Efficiency (AFUE) among other expressions can be used depending on the HVAC functionality.
Backup	System Backup, Recovery Capabilities, Data Redundancy Reduction	50, 117	Operational Efficiency	System's ability to securely preserve data and ensure its availability for recovery during disruptions, emphasizing robust mechanisms and minimal data loss. The opposite perspective of backup is Data Redundancy Reduction, that can be a metric when the objective is a scalable service (big IoT services) without considerable repetition, as mentioned in [117].	Following [50], Backup is evaluated as part of the reliability indicators (B4) using the analytic hierarchy process (AHP). Experts assign scores (on a scale of 0 to 10) based on system performance, which are normalised and weighted according to pre-defined importance levels.
Budget	Costs Planned, IT Plan, Plan	42	KPIs and Metrics for Data Monetisation	A budget is a financial plan that outlines expected income and expenditures over a specific period. It represents an estimate of how much money an individual, business, government, or organisation expects to earn (revenue) and how much it plans to spend (expenses) to achieve certain financial or operational goals.	Direct estimation or indirectly by using Total Incomes and Total Expenses.
Cache Size		53	Technology and Infrastructure	Amount of data (measured in units such as kilobytes, megabytes, gigabytes, etc.) that a cache can store at any given time. The cache is a high-speed data storage layer that stores a subset of data, typically temporary, to serve future requests more quickly than retrieving the data from its primary storage or origin location.	Direct measurement.

CAPEX	Capital Cost, Hardware Costs, Software/Application Costs, Service Cost, Infrastructure Unit Costs	111, 3, 131, 50, 12, 15	KPIs and Metrics for Data Monetisation	One-time, upfront investments made to acquire, upgrade, or maintain physical assets or infrastructure, independent on the number of quotes needed to cover this expense. Examples: (1) Purchasing new machinery or equipment; (2) Constructing buildings or facilities; (3) Upgrading technology infrastructure (e.g., data centres, servers); (4) Acquiring servers or other long-term assets; (5) Investing in property or land. For a data and information system, it encompasses the following components: (1) Data Capturing Hardware: Computers and devices used to collect and input data; (2) Acquisition Systems: Sensors, data acquisition (DAQ) systems, and similar tools for measuring and capturing values; (3) Information Collection Mediums: Devices like PDAs and tablets used for collecting data. (4) Data Storage Hardware: Systems for storing large volumes of data. (5) Data Processing, viewing, and Analysis Hardware: Computers and servers for processing and analysing data. If existing hardware is repurposed, the cost should be based on its depreciation during use. If fully depreciated, the cost is effectively zero. Although the cost of software can be significant, other software such as Microsoft Office are typically pre-installed on the computers bought by organisations and used by employees. Therefore, they normally run at zero cost for the node and can excluded from its costs.	Direct estimation. The formula for Capital Expenditures (CAPEX) calculates the total investments made by a business in acquiring, upgrading, or maintaining physical and intangible assets. A basic representation of Capex is: $CAPEX = \Delta PPE + Depreciation\ Expense$ ∆ PPE Represents the increase or decrease in the value of property, plant, and equipment over a period: ∆ PPE = Net PPE at the End of the Period - Net PPE at the Beginning of the Period. Depreciation Expense accounts for the wear and tear or usage of physical assets over time and is typically listed on the income statement.
Carbon Emission Factor (CEF)		21	Innovation and Growth	Coefficient used to calculate the amount of CO_2 emissions produced per unit of energy consumed, typically expressed in kilograms of CO_2 per kilowatt-hour ($kgCO_2/kWh$). It varies depending on the energy source (e.g., coal, natural gas, renewable energy).	Direct estimation.
Carbon Usage Effectiveness (CUE)		86, 21, 72	Innovation and Growth	A carbon metric used to measure the environmental impact by assessing the amount of CO_2 emissions per unit of IT energy consumed.	$CUE = PUE + CEF$
Churn		49	Customer Needs and Satisfaction	Refers to the loss of customers. Prediction uses demographics, usage, transactions, interactions, sentiment and external factors to anticipate customer loss.	Defined churn per context (e.g. telecom: no recharge within 15 days of zero balance). Next, data should be collected from multiple sources, including billing records, call detail records (CDRs), customer demographics, and operational and business supporting systems (OSS/BSS). These datasets should capture payment patterns, service usage, behavioural data, and metrics related to service quality, such as call drops or network issues. Feature engineering is then applied to extract relevant attributes that capture customer satisfaction (behaviour), engagement, and service usage. Customers are labelled churners = 1 or non-churners = 0 per the definition. Using a sliding-window framework, data from month N−1 trains a model to predict churn in month N+1, often via decision trees or random forests. Finally, predicted churn is aggregated and weighted by segment importance or revenue.
Clarity	Understandability, Easy to Understand, Lack of Confusion, Unambiguity, Concise,	39, 11, 70, 83, 107, 128, 63, 10,	Customer and Market Oriented	Refers to how easily the data can be understood, interpreted, and used by its intended audience. It involves presenting data in a way that is straightforward, unambiguous, and free from unnecessary complexity. Clear data allows users to quickly grasp its meaning, draw	Pondered in function of the importance of key components, for instance, Clarity = w_1·Consistency + w_2·Readability + w_3·Contextual Information + w_4·Visualisation + w_5·Completeness, where w_i is the weight given to a component.

Funded by
the European Union

	readability, Interpretability	4, 44, 134, 17, 25		insights, and make decisions without confusion or misinterpretation. The Key Aspects of Data Clarity from a KPI perspective could be: **Simplicity and Structure**: Data should be presented in a well-organised manner, often through tables, graphs, or summaries that make complex data easier to digest. Example: Using clear, concise headers in a dataset and avoiding overly complex structures. **Consistency**: Consistent terminology, formats, and units of measurement improve clarity. Data should follow the same structure and presentation style across the dataset. **Readability**: Data should be easily readable, with clear labels, appropriate formatting, and a clean layout. Good readability ensures that the user can navigate the data without confusion. **Context, Metadata, and/or Explanation**: Providing context or explanations (e.g., metadata, labels, or definitions) that clarify what the data represents. **Redundancy/Containment Fraction**: Data should be free from unnecessary repetition or irrelevant information that can clutter the understanding (within the same dataset - since redundancy is important for availability). Example: Avoiding the use of the same information in multiple columns or rows that do not add value to analysis. **Visualisation**: Using the right type of visualisation (charts, graphs, maps, etc.) can enhance clarity by making patterns, trends, or insights easier to recognise. **Completeness**: checks if the dataset has all the required fields and values without missing information (measured as a percentage). Not all these factors are included in the main taxonomy figure, for simplicity they have been reduced there. Nevertheless, that factors could be considered as a base since several of the main components here can be related to readable, conciseness, and understandability.	
CO_2 Savings		86, 116, 95	Operational Efficiency	The metric CO_2 Savings in the document refers to the change in data centre CO_2 emissions from a given baseline. For example, this metric evaluates the reduction in CO_2 emissions achieved by implementing the GENiC system [86] and its integrated management strategies for optimizing energy usage, renewable energy integration, and heat recovery in data centre	$CO_2\ Savings = \frac{Possible\ Emissions}{Actual\ Emissions}$ There is no direct report in [86], it could be calculated based on the type of energy consumption and type of energy used. The equation might look something like this: $CO_2 Savings = \frac{CO_{2,baseline} - CO_{2,optimized}}{CO_{2,baseline}}$ Baseline: The emissions are calculated using the original energy consumption of the data centre. Optimised: The emissions calculated after implementing modifications. .
Competitive Advantage		16	Customer and Market Oriented	Degree to which data confers a unique strategic edge; impact gauged by potential consequences if competitor's access or you lose the data.	Assessed by asking "What if competitors access this data?" or "What if you lose it?"; scored as no impact (irrelevant), moderate (process insight), or significant (lost positioning).

Completeness	Appropriate Amount of data	39, 11, 10, 65, 70, 46, 60, 98, 103, 136, 41, 63, 118, 38, 124, 113, 100, 114, 61, 102, 17, 80, 43, 25, 48, 15, 37	Data Quality	Extent to which all required data is available and present within a dataset. It measures how much of the expected information is provided, indicating whether critical data fields are missing or left blank. Key aspects of data completeness include: **Presence of mandatory fields**: All required fields have values. **Proportion of missing data**: A low percentage of missing or null values indicates better completeness. **Consistency with expectations**: The dataset should align with the expected structure and content, covering all data points. **Granularity**: The data should provide enough detail to meet the desired level of analysis.	Can be estimated as a weighted combination of the key aspects that contain them. If only one factor is deemed to be important, data completeness can be calculated by using a simple formula that measures the proportion of filled (non-missing) values relative to the total number of expected values. (i.e. $Completness = \frac{Expected\ Values - Missing\ Values}{Expected\ Values} \times 100$) **Expected Values**: The total number of data points or fields that are supposed to have values. Furthermore, as defined by [46], for a relation R, let TR be the number of tuples in R that have at least one "NULL" value and let NR be the total number of tuples in R. Then, the completeness of R is defined as follows. $$completeness = 1 - \frac{TR}{NR}$$ At the data level, a piece of data value is incomplete if and only if it is "NULL"; otherwise, it is complete (i.e., the metric value is one). Here, all data values that represent missing or unknown values in a specific application scenario (e.g., blank spaces or "9/9/9999" as a date value) are represented by the data value "NULL." A tuple in a relation is defined as complete if and only if all data values are complete (i.e., none of its data values is "NULL"). Finally, from [114], given the direct relation between Completeness and Appropriate Amount of data, the latter can be calculated as $Min\ (;\ )$.
Compute Power Efficiency (CPE)		121	Operational Efficiency	Metric that assesses the efficiency of IT equipment in a data centre by measuring how much of the total power consumption is effectively used for computation.	$$CPE = \frac{Useful\ Computational\ Work}{Total\ Equipment\ Power}$$ $$CPE = \frac{Equipment\ Utilization\ Energy}{PUE}$$
Conciseness	Concise Representation, Simplicity	39, 11, 83, 107, 100	Data Quality	Principle of conveying essential information briefly, clearly and efficiently by minimizing redundancy and noise. Key characteristics of Concise Representation could include: **Brevity/Compression**: The information is presented using the fewest words, symbols, or data points necessary. It also Reflects how much data can be compacted without losing critical information. **Clarity**: The content is clear and easy to follow, even when reduced to its essential elements. **Efficiency/Signal-to-noise Ratio**: It avoids redundancy, long-winded explanations, or superfluous data. **Redundancy**: Measures the repetition of identical or similar data within a system or dataset. **Relevance**: Only the most relevant information is included, eliminating any extraneous data that does not add value to the understanding of the content. **Semantic Consistency**: Measures alignment and consistency in the terminology and format used across the dataset.	Measured by Redundancy Ratio (RR), Task Efficiency Ratio (TER), Information Density (ID), or Compression Ratio (CR): $RR = \frac{D_r}{D_t},\ \ TER = \frac{As}{Es},\ \ ID = \frac{I}{R},\ \ CR = \frac{U_s}{C_s}$ Where lower RR, TER ≈ 1, higher ID and CR indicate greater conciseness. Where D_r is the number of redundant entries, and D_t is the total number of entries. As is the actual number of steps taken, and Es is the optimal number of steps. I is the amount of meaningful information, and R is the resources used (e.g., space or time). U_s is the uncompressed size, and C_s is the compressed size.

Funded by the European Union

				Duplicate Elimination: A measure for the process of identifying and removing identical records.	
Confidence		120	Operational Efficiency & Data Quality	The confidence metric measures how strongly a dependency holds in the dataset.	Based on equation (16) in [120], the confidence metric can be expressed in terms of Γ (the family of metric neighbourhoods): Refer to manuscript for further information. The confidence metric, quantifying how reliably the neighbourhood defined by X predicts the neighbourhood defined by D, is represented as: Conf $_{xd}$ (X → D) The idea is to calculate the neighbourhood of tuple t_i with respect to the condition attributes X and a decision attribute D. Then the proximity of these neighbourhoods is used to calculate the metric.
Consistency	Linked to Heterogeneity, Veracity	75, 11, 10, 70, 63, 46, 98, 103, 136, 128, 38, 129, 113, 134, 9, 102, 17, 80, 43, 25, 48, 13, 7	Data Quality	Refers to how well data values conform to predefined rules, such as association rules in the context of relational databases. The metric for consistency, as defined by Alpar and Winkelstrater [7], evaluates the consistency of a tuple t based on whether it fulfils or violates certain rules r from a set R of association rules. Consistency means that data is uniform and reliable across all datasets, systems and applications, remaining accurate and contradiction-free. Key KPI aspects include integrity, uniformity, synchronisation, rule validation, error prevention and governance.	As a metric, consistency is more narrowly defined and can be quantified mathematically. According to [80], consistency measures the proportion of values within an 80% confidence interval for numerical variables, assuming a Gaussian distribution. It is calculated as: $$Consistency = \frac{1}{N}\sum\frac{\lvert Y_j^C \rvert}{T}$$ Where N is the number of variables, T is the total number of observations, Y_j^C represents the subset of values for variable j that fall within the 80% confidence interval. Furthermore, as recommended in [7, 46], Consistency is calculated by classifying if data fulfils, violate, or does not apply. Given different weights to each one. For the KPI, fulfils and violates can be grouped by each of the key components and then define benchmarks or limits to transform it to a KPI (i.e. group by synchronisation, rules validation, uniformity, etc.).
Containment Fraction		105	Operational Efficiency & Data Quality	Measures the extent to which one dataset is contained within another. This metric helps identify redundancy and optimise storage usage. A limit in containment fraction must exist since availability depend on duplication, for security, of identical information.	The "containment fraction" is a metric defined in the document for analysing redundancy in datasets. Specifically, the containment fraction of a dataset A in another dataset B is expressed as: CM (A, B) = |A∩B| |A|: The number of rows (or schema elements) in A; |A ∩ B|: The number of rows (or schema elements) that are common between A and B. The calculation depends on whether the containment is schema-based or data-based: For schema-level containment, |A| refers to the number of elements in the flattened schema of A, and |A ∩ B| is the number of schema elements shared between A and B. For table-level containment, |A| refers to the number of rows in A, and |A ∩ B| is the number of rows common between the two datasets. A containment fraction of 1 indicates that A is fully contained within B.
Cooling Capacity Factor (CCF)		116	Operational Efficiency	Assess the utilisation efficiency of the cooling infrastructure.	$$CCF = \frac{Test\ Cooling\ Capacity}{Critical\ Load}$$

Cooling Effectiveness Rate (CER)	Energy Effectiveness of Cooling	86	Operational Efficiency	Focuses on the effectiveness of cooling systems in the data centre. There is a direct link to other metrics related to cooling effectiveness, such as "Energy Effectiveness of Cooling Mode in a Season". Assesses how effectively the cooling system operates in different seasons, which can influence operational costs and carbon emissions.	$CER = \frac{E_{cool}}{E_{total}}$ Where E_{cool} is the energy used for cooling and E_{total} is the facility's total energy consumption.
Corporate Average Data Centre Efficiency (CADE)		95, 101	Operational Efficiency	CADE is designed to provide a comprehensive assessment of a corporation's data centre energy performance by considering both infrastructure efficiency and IT asset utilisation. CADE allows the calculation and measurement of a data centre's energy consumption so that it can be compared to the other data centres.	$CADE = DCiE \cdot IT\ Asset\ Efficiency\ (ITAE)$ Where ITAE is the average CPU Utilisation (in %).
Cost of Degradation (CoD)		109	Data Quality	This metric quantifies the loss of data quality due to privacy-preserving transformations by measuring the reduction in utility between the original and adjusted data.	$CoD(h, j) = \frac{1}{(h - j)^{\propto}}$ Where h is the total hierarchy height, j is the child level (generalised from j − 1), 0 < α ≤ 1 is a factor used to adjust the sensitivity. Higher-level generalisations (j near 0) yield larger CoD. Lower α smooths sensitivity. Summing CoD(h, j) over all attributes gives the dataset's total cost. For example, transitioning from "State Govt Job" to "Govt Job" incurs a lower CoD than transitioning from "worked" to "with-pay" because the latter involves a more generalised category. The closer the transition is to the root node, the higher the degradation in data quality.
Currency	Linked to Timeliness	134, 114, 11, 46, 107, 136, 118, 124	Data Quality	Measures the change in data value given change in time or processes in the data. Currency measures how "fresh" data is by comparing its age (time since last update) against how often it should be updated, while timeliness measures how quickly data arrives or is available when it's needed—i.e. the delay between when a value is created and when it's delivered or used.	Data currency measures how current the data is by comparing the time elapsed since its last update to the acceptable update interval: $Currency = 1 - \frac{Data\ Value}{Data\ Value\ since\ Last\ Update}$ Currency can also be assessed by comparing when data is needed versus when it is actually delivered: $currency\ change = 1 - \frac{Delivery\ Time - Expected\ Delivery\ Time}{Acceptable\ Time\ Window}$ Another method for calculating timeliness involves summing decay rates across attributes.
Data Aquisition Cost (DAC)	Cost of Procurement	68	KPIs and Metrics for Data Monetisation	Data Acquisition Cost (DAC) is the fraction of local models or data a Data Acquirer must obtain from Data Providers, each training epoch in federated learning, to build an accurate global model; all other expenses (e.g., purchase price, OPEX) are tracked separately.	DAC = Number of Local Models Procured by DA / Total Available Local Models * 100
Data Centre Compute Efficiency (DCcE)		95	Operational Efficiency	Assess the efficiency of computing resources within a data centre. It provides insights into how effectively the data centre's IT equipment is utilised to perform computational tasks relative to the energy consumed.	Refer to citations for estimation. No clear approach defined.

Funded by the European Union

Data Centre Lighting Density (DCLD)	Lighting Power Density (LPD)	95, 101	Operational Efficiency	Measures the electrical power used for lighting per unit area within a data centre, typically expressed in watts per square foot (W/ft²) or watts per square meter (W/m²). This metric helps assess the energy efficiency of the lighting system in the facility.	$DCLD\ =\frac{Area\ Data\ Centre}{Total\ Lighting\ Power}$
Data Centre Adaptation (DCA)	Data Centre Energy Profile Change	86	Operational Efficiency	The Data Centre Energy Profile Change is referred to as DCA (Data Centre Adaptation) in the context of energy metrics outlined in the GENiC framework from the manuscript. It captures the change in the data centre's energy profile from a predefined baseline. This metric is part of a broader set of evaluation measures used to assess energy efficiency, sustainability, and the ability to adapt to renewable energy integration [86].	The Data Centre Energy Profile Change (DCA) from the GENiC framework [86] is: $DCA=\frac{E_{current}-E_{baseline}}{E_{baseline}}$ Where $E_{current}$ is the total energy consumption of the data centre during the monitoring period (could be hourly, daily, or over a defined time range), and $E_{baseline}$ is the historical baseline energy consumption under similar operational conditions.
Data Ingestion Capabilities	Data Collection and Management Capabilities, Data Analysis, Data Mining, Data Sources	126, 50, 28	Technology and Infrastructure	Data Ingestion Capabilities describe a system's ability to acquire, collect and store data from diverse sources at scale and speed. Key aspects include high throughput, low latency (including delayed-data handling), scalability under growing device or data volumes, and integrity maintenance. Common technologies are time-series databases, message queues (e.g. Kafka) and compressed formats like TsFile.	$DIC=\frac{T\cdot SR}{L\cdot RU}$ Where $T=\frac{Total\ Data\ Points\ Ingested}{Time\ Taken}$; $L=\frac{Total\ Time\ Taken}{Total\ Data\ Points\ Ingested}$; $SR=\frac{Sucessfully\ Ingested\ Data}{Time\ Taken}$; $RU=\frac{Resource\ Used}{Total\ System\ Capacity}$
Data Price	Data Value, Price Function, Payoff, Reimbusement, Financial Value, Price of Information	91, 2, 74, 115, 130, 94, 131, 71, 45, 132, 133, 28	KPIs and metrics for Data Monetisation	Data price is the financial value assigned to data, determined through models or market mechanisms. It encompasses concepts like monetary cost, price functions, payoff (reward for contributions), reimbursement (compensation for costs), and volume-based metrics (e.g., price per MB). Methodologies include entropy-based valuation, utility models, and dynamic pricing, tailored to specific contexts such as privacy, real-time trading, or structured markets.	Data pricing depends on purpose (e.g. ML vs. real-time), data traits (freshness, privacy, quality) and market dynamics (demand, competition). This document summarises pricing methods tailored to dataset features, use cases and market requirements. **Contribution-Based Approaches** — Shapley Value [130] allocates price via average marginal contributions; predictive contribution [2] sets prices based on data's impact on ML accuracy. These methods require upfront CAPEX for framework development and ongoing OPEX. **Utility-Based Approaches** — CDFs [94] gauge utility before/after obfuscation to show retained value. Age of Information (AoI) [45] adjusts pricing as freshness declines, using uniform, dual or dynamic models to balance profit and user sensitivity. These models incur OPEX for continuous sampling and real-time updates. **Information-Theoretic Approaches** — Entropy-based pricing [71] uses information density to value data; Bayesian entropy [132, 133] sets prices via posterior entropy reduction in probabilistic models. These methods need substantial CAPEX for system setup and ongoing OPEX. **Dynamic and Contextual Pricing** — Real-time models [91] adjust prices for demand, usage and privacy; DeDa [91] adds blockchain elasticity; VAP [132, 133] uses contextual multi-armed bandits to optimise pricing. **Stream-Centric Pricing** — Assigns value to continuous data streams based on temporal relevance and query load, emphasizing low latency and real-time adaptability [28]. **Privacy-Integrated Approaches** — differential privacy budgets [115] balance noise and utility; obfuscation trade-offs [94] quantify privacy vs. utility for pricing.

					Market-Driven Mechanisms — Subscription/query pricing [71] charges per duration or query; auction-based models [133] sell to the highest bidder, balancing supply and demand. Furthermore, please refer to 162.pdf for a detailed explanation of the full process of price evaluation, since it is too complex to describe here. Based on 163.pdf, the initial data price is calculated as: $$P = \frac{(1 + \frac{m}{100})c}{s}$$ where $m\%$ is the desired margin, c the data cost, and s expected sales. It then adjusts by transactions and time. The final price combines both effects.
Data Principles and Practices	Data Standard Driven, Standardisation	46, 129, 124	Governance and Compliance	This metric evaluates how closely an organisation complies with defined data standards and governance policies to ensure data consistency, reliability, and alignment with its objectives. It reflects a commitment to maintaining high-quality data by enforcing predefined standards and measures compliance through clear, quantifiable thresholds in data-quality assessments [129][124].	As a KPI, it can be estimated as main components. As metric, it is estimated by tracking adherence to data-quality standards through explicit metrics such as compliance percentages and COBIT classifications like "Largely Achieved" or "Fully Achieved" and by evaluating systematic consistency measures (e.g., governance frameworks and QA processes). Regulatory compliance is measured on a binary pass/fail scale [129], supplemented by qualitative survey or expert-interview scores to validate principle adherence [129].
Data Similarity (this is different from purely "similarity")	Euclidean Distance, projection similarity, similarity score, cosine similarity, average distance, Kolmogorov-Smirnov, Mann-Whitney, Mood's Median, and Levene's test (LE)	120, 24, 31, 125, 5	Data Quality	Distance metrics quantify numerical distances between points rather than relying on syntactic likeness; for example, projection similarity assesses datasets by comparing their feature dimensions.	**Cosine Similarity**: Measures the cosine of the angle between two vectors. $$cosine\ similarity = \frac{D \cdot P_i}{\|D\|\|P_i\|}$$ where P_i is a feature vector of the known pattern, and D is the input data vector. Used to compare the orientation of two vectors regardless of their magnitude. **Jaccard Similarity**: Compares the similarity and diversity of sample sets. $$Jaccard\ Similarity = \frac{D\ \cap\ P_i}{D\ \cup\ P_i}$$ where D and P_i represent two sets. Ideal for determining the overlap between datasets or sets. **Euclidean Distance**: Computes the straight-line distance between two points in a vector space. $$Euclidian\ Distance = \sqrt{\sum (D_k - P_{i,k})^2}$$ where D_k and $P_{i,k}$ are components of the respective vectors. Measures absolute differences between feature vectors. **Statistical Tests**: From the referenced material, the following statistical tests are also highlighted for similarity measures([125]): **Kolmogorov-Smirnov (KS) Test**: A non-parametric test comparing two distributions. **Mann-Whitney (MW) Test**: A test assessing whether one of two samples tends to have larger values than the other. **Mood's Median (MD) Test**: Compares medians across

					multiple groups. **Levene's (LE) Test**: Evaluates the equality of variances for a variable across groups.
Data Type		111	Data Valuation Technique	Using decision-based valuation [111], data is classified into four categories: A for operational—frequent, short-lived, highly time-sensitive; B for one-time decisions—nonrecurring, high-accuracy, long realisation period; C for legal/safety—low immediate value but legally retained long-term; and D for research/innovation—uncertain future value with extended lifespan. Other categories can be used if desired.	Direct calculation.
Data Value (DV)	Data Criticality, Value of Information, Fixed Record Value, IP Value, Intrinsic Record Value	111, 129, 45, 81, 132, 56, 61, 5, 12, 15, 39	Data Valuation Technique	Do not confuse Price with Value. Data value denotes a data object's relevance to a specific consumer, shifting with application context and needs even when its quality remains unchanged. In distribution services, it quantifies the worth of a decision node within the Decision-Based Valuation framework by weighing cost and lifecycle. Closely related, data criticality measures how indispensable certain data is to business operations and decision-making, assessing its impact on processes, revenue, and service delivery.	To compute data value in the Decision-Based Valuation (DBV) method, the total value of a decision node VN is estimated based on first estimating each source's percentage contribution VR ∈ [0, 1] and its processing value VP; For instance, if a data source contributes 60% to the decision, VR = 0.6. Then each source's data value contribution VD is calculated (See [111] for further information). Alternatively, Completeness, Accuracy, Availability and Ubiquity can be used to calculate Value of Information enter via a Value of Information measure [39]: $$VI = \frac{Completeness \cdot Accuracy \cdot Availability}{Ubiquity}$$ Following again [111], the processing value (VP) accounts for the cost or value added during data processing. It is calculated using the formula: $$VP = 1 + \frac{t_p}{t_a}$$ where t_p is the time spent processing the data and t_a is the time spent acquiring it. The contribution of each data source to the decision node's value is calculated using the formula: $$VD = \frac{(VN \times VR)}{VP}$$ where VD is the value contribution of the data, VN is the total value of the Decision Node, VR is the percentage contribution of the data, and VP is the processing value. Distributing the value of the decision node among multiple data sources can be approached in two ways: **Even Distribution**: The value is evenly distributed across all data sources, regardless of their individual contributions. **Weighted Distribution**: A larger portion of the value is allocated to data sources that have a more significant impact on the decision node's outcome. A hybrid approach can be employed. This nuanced distribution requires a thorough understanding of the contribution and relevance of each data source. If a decision node relies on a single dataset, the calculation becomes more straightforward, as the entire value of the decision node is attributed to that dataset. This methodology highlights the necessity of integrating comprehensive metrics, such as completeness, accuracy, availability, and ubiquity, to ensure a robust and context-sensitive valuation of data.

					Furthermore, approaches based on Age of Information [45, 56]; Wasserstein Distance for Data Quality [81], and Entropy [132] can be found and studied (refer to these texts for alternative data valuation approaches).
Data Value Ratio (VR)		111	Valorisation Technique	The Data Value Ratio (VR) represents the proportion of a decision node's value attributed to a specific data source.	It is calculated as part of the value contribution (VD) using the following formula: $VD = VN \cdot VR / VP$ VD: Value contribution of the data source. VN: Value of the decision node, reflecting its overall value in the decision-making process. VP: Processing Value Ratio, accounting for the impact of processing on the data's value. The VR determines the percentage contribution of a specific data source to the decision node's overall value. Alternatively, if the decision-making process involves only costs, it can be calculated as (not defined in [111] but directly related in this work): $VR = VD / TotalCost(CT)$
Data Centre Performance Effciency	DCPE	95	Operational Efficiency	Metric used to evaluate how effectively a data centre utilises energy to perform useful work.	$DCPE = \frac{Useful\,Work}{Facility\,Energy}$
Data Centre Performance Per Energy	DPPE	95	Operational Efficiency	Measure of how efficiently a data centre converts energy into computational performance.	$DPPE = DCiE \times GEC \times ITEE \times ITEU$
Data Center Power Density	DCPD	95	Operational Efficiency	Metric used to measure how much power is consumed per rack in a data center. It helps data center operators assess how efficiently they are utilizing their rack space and power capacity.	$DCPD = \frac{rack\ power\ consumption}{number\ of\ racks}$
Data Center Productivity	DCP	121	Operational Efficiency	Metric that evaluates how effectively a data center converts its consumed energy into useful computational work. Unlike efficiency metrics like PUE, which focus on energy distribution, DCP emphasizes the actual output or performance of a data center in relation to its energy consumption in an area or general components.	$DCP = \frac{Useful\,Work\,Output}{Total\,Facility\,Energy}$
Data Center Space Efficiency	DCSE	95	Operational Efficiency	Assesses how effectively a data center utilises its available physical space. Efficient space utilisation is crucial for optimising operational costs, energy consumption, and overall performance.	DCSE = Rack Unit Space Utilisation / Floor Space Utilisation DCSE = Overall Unit Space Utilisation / Floor Space
Data Center Workload Power Efficiency	DWPE	120	Operational Efficiency	Bridges the gap between infrastructure-level and workload-level energy efficiency. It is the ratio of the energy consumption of the IT equipment to the energy consumption of the entire data centre.	$DWPE = \frac{E_{IT}}{E_{Total}} W_{eff}$, where E_{IT} represents the energy consumption of the IT equipment. E_{Total} denotes the total energy consumption of the data center. W_{Eff} is the workload-specific efficiency factor.
Data Centre Energy Productivity	DCeP	95, 121	Operational Efficiency	The DCeP essentially defines the data centre as a blackbox—power goes into the box, heat comes out, data goes into and out of the black box, and a net amount of useful work is done by the black box. In other words, quantifies useful work compared to the energy it requires. It can be calculated for an individual IT device or a cluster of computing equipment.	$DCeP = \frac{Useful\,Work\,Done}{Energy\,Consumed\,Over\,Time}$

Data Robustness	Shapley Robust	2	Operational Efficiency	Data robustness ensuring that datasets maintain their value and usability across various prediction tasks.	In [2], robustness is calculated using a Shapley-based framework to ensure fair value allocation in a data marketplace. It identifies and penalises replicated or redundant datasets using similarity metrics, reducing their contribution weight. Robustness is evaluated by simulating replication scenarios and measuring the system's ability to maintain fair value distribution and resist manipulation, demonstrating effectiveness against adversarial behaviors.
∆-T Per Cabinet	T - Per Cabinet	116	Operational Efficiency	Measures the temperature difference between the air entering and exiting a server cabinet, reflecting cooling efficiency and equipment health; it is aggregated here with the similar "Temperature per Cabinet" metric, which also tracks cabinet-level thermal conditions.	$\Delta T_{cabinet} = T_{outlet} - T_{inlet}$
Demand	Score, Number of Data Consumers	94, 15	KPIs and Metrics for Data Monetisation	The demand or score for data is defined differently depending on the system under consideration and thus, the strategy to be implemented. For example, for DaaS demand for data is conceptualised as buyers' willingness to pay for specific types of data. This demand is measured using a scoring mechanism where buyers rate the desirability of different categories of data. Also, different models, such as Hotelling, can represent system and demand. Independently, this is a new domain in which no clear estimation of demand has been defined.	For a broader perspective, Demand can be inferred from metrics such as data usage rates, download frequency, or the economic contribution estimated from the data's deployment. Then: $Demand = f(usage\ rates, download\ frequency, economic\ contribution, quality, service, \dots)$
Deployed Hardware Utilisation Efficiency	DH-UE	95	Operational Efficiency	Measures the efficiency of deployed servers by comparing the minimum number of servers required to handle peak compute load to the total number of servers deployed.	$DH - UE = \frac{Minimum\ Number\ of\ Servers\ Required\ for\ Peak\ Load}{Total\ Number\ of\ Servers\ Deployed}$
Detail		39	Data Quality	There is no clear specification of Detail for its estimation in reference. Nevertheless, purely technical accuracy considerations could imply how finely data is broken down or how much descriptive information is captured in each data point. Then key aspects of data can incorporate: **Granularity**: The extent to which data is broken down into smaller, specific components. High granularity means more detailed data. **Descriptive Attributes**: The number of attributes or fields that describe an entity or event. More detailed datasets have a larger number of attributes that provide rich, descriptive context. **Precision**: How exact the data is. For numerical data, this might refer to how many decimal points are included, for temporal the description of timestamps, while for categorical data, it refers to the specificity of categories. Other business-related measures can also be considered.	As as KPI, through the combination of key aspects.
Differential Privacy		127, 30, 130, 74, 115	Data Governand and Compliance	Same as Inferential Privacy, but in the opposite direction. Also as defined in [130], a metric used to gauge the level of privacy preserved in data when shared or sold in marketplaces. Differential privacy provides a rigorous, quantitative way to limit how much information about any one individual (or data point) can be inferred from the output of an algorithm. At a high level, it says: "If you add or remove a single individual's data from the dataset, the algorithm's	Please refer to [30,130] for methods to derive and implement noise. In practice this is usually achieved by adding carefully calibrated random noise to the algorithm's output.

				output distribution shouldn't change too much". Differential privacy is a widely adopted concept to quantify the privacy loss in private ML algorithms.	
Discoverability	Linked to Effectiveness (Platform), Platform Performance	93, 83, 44	Innovation and Growth	As defined in [93], this metric describes the ability of a dataset to show the content of information and be able to search or discover datasets. For a platform, measures the platform's ability to meet the user's information needs accurately. This KPI evaluates how effective the search functionality is in returning relevant datasets based on user input.	As a miked metric component (or KPI), it can be measured as an aggregation of key components, including: **Searchability** – The extent to which data or content can be found using a search engine or filtering mechanisms. **Metadata Usage** – The presence of structured metadata that improves indexing and retrieval. **Content Transparency** – Whether the dataset displays sufficient information to be identified and assessed before access. **Navigation Efficiency** – How easily users can browse and locate relevant content on a platform. Others such as visibility, indexing, metadata and organisation can also be considered.
Downloads	Download Frequency	92, 23, 15, 37	Customer and Market Oriented	The number of times users have clicked to retrieve a dataset.	Direct measurement.
EAFDL		54	Data Quality	EAFDL measures annual data loss as a fraction of total stored data.	$EAFDL = \frac{E(H)}{MTTDL \cdot U}$ where: U is the total amount of user data stored in the system. E(H) is the expected data loss, calculated as $E(H) = c \prod \alpha_j$ where c is the amount of data per device, and α is the fraction of rebuild time completed during subsequent failures. MTTDL is the mean time to data loss.
Ease of Measurement		16	Data Valuation Technique	Ease of Measurement gauges how readily a data value dimension can be quantified using existing methods, tools or frameworks. It reflects the effort and complexity required to evaluate that dimension—helping you focus on those that are straightforward to assess and flagging ones that may need extra work or context.	In [16], Ease of Measurement is not explicitly calculated as a numeric metric. Instead, it is described qualitatively based on the existence of established frameworks that can be used to assess a data value dimension. Here's how it is implicitly evaluated: **Presence of Established Metrics or Standards**: Dimensions with established, widely used metrics—like timeliness or data quality—are easier to measure (e.g., via the W3C Data Quality Vocabulary or DaVe). **Dependence on Complex Models or Context Dimensions** that require business process models or understanding of outcomes, dependencies, and context are harder to measure. For example, business utility or impact depends on understanding specific contributions of datasets to operational efficiency or profit. This makes these dimensions harder to quantify. **Scoring Likert-Style Assessments**: Subjective dimensions like competitive advantage are scored on a 0–4 Likert scale, while concrete questions (e.g., "What happens if you lose this data?") are inherently easier to measure than abstract ones.
Easy-to-Use	Easy to use, ease of use, search (for	136, 110, 44, 118, 23, 50, 77, 25	Data Quality	The degree to which a system, product, or interface is designed to enable users to efficiently, effectively, and satisfactorily achieve their goals with minimal effort and complexity, ensuring accessibility and user-friendliness in operation and learning.	Easy-to-Use can be measured through various methods:

	queries), (user) friendliness				**Heuristic Evaluation**: Assessing user interface designs against usability principles, such as navigation simplicity and feedback mechanisms [44, 50]. **Usability Testing**: Observing users completing predefined tasks to measure efficiency (time and effort), effectiveness (accuracy and completeness), and satisfaction [44, 50]. **User Feedback**: Collecting survey responses to evaluate learning curve, intuitiveness, and user satisfaction [118, 77]. **Performance Metrics**: Monitoring error rates and task success rates to gauge ease of system use [110,136]. **Weighted Criteria Analysis**: Using methods like Analytic Hierarchy Process (AHP) to evaluate and aggregate scores for ease-of-use dimensions [50, 77]. Key metrics include task completion time, error rate, user satisfaction ratings, and the learning effort required [44, 50, 110].
Economic Efficiency	Data Business, Characteristic Index	46, 20	KPIs and Metrics for Data Monetisation	This index gauges data criticality for business processes by weighting its operational relevance—higher weights signal greater importance. Economic efficiency ensures that the gains from data quality metrics exceed the costs of applying and improving them.	Refer to citation for further information but it could be estimated from the overall economic efficiency or data value V, calculated as a sum of calculated values V_i weighted ω_i (like a KPI). Each characteristic indexed i is defined as follows: **Data Business**: measures the business relevance of data. It is determined based on expert evaluations that assign weights to reflect the importance of data in supporting specific business objectives. **Data Application**: assesses the utility of data in different applications. This is also evaluated through expert analysis, considering the significance of data in decision-making processes. **Data Security Level**: quantifies the sensitivity and confidentiality of data. Weights are assigned to reflect security requirements, as determined by expert judgment. **Data Update/Access**: computed from operational logs that track the frequency of operations such as additions, deletions, modifications, and queries. The frequency values are processed to calculate an index value. **Data Creation**: reflects the age or recency of the data and is weighted based on its importance for current business needs. Finally, the incremental benefit of considering data quality in decision-making can be assessed using the following formula: $\Delta E = E(a2, DQ, P2, S) - E(a1, DQ, P1, S)$ Here, E(a2, DQ, P2, S) is the expected payoff when data quality (DQ) is considered, and decision alternative a2 is chosen. Conversely, E(a1, DQ, P1, S) is the expected payoff when data quality is ignored, and decision alternative a1 is chosen. The term ΔE quantifies the economic improvement achieved by integrating data quality considerations into decision-making processes.

Elapsed Time		53, 47, 91	Operational Efficiency	Total time taken to complete an operation.	$ET = T_{end} - T_{(start)}$
Electronics Disposal Efficiency	EDE	95	Operational Efficiency	Assesses the disposal of decommissioned Information and Communication Technology (ICT) assets.	$EDE = \frac{Total\ Weight\ of\ Responsibly\ Disposed\ ICT}{Total\ Weight\ of\ Decommisnioned\ ICT}$
Encryption Time (ET)	Decryption Time (DT)	31	Operational Efficiency	Measures the time taken to convert plaintext data into an encrypted format using encryption algorithms. It's important in cloud and data security to understand how quickly data can be secured before transmission or storage.	Direct measurement or, alternatively, could be estimated based on the estimation of different components, such as $T_d = f\ (A, K, D, S)$ where: T_d = Decryption Time, A = Algorithm complexity (e.g., AES, RSA, Blowfish), K = Key size (e.g., 128-bit, 256-bit), D = Data size (amount of encrypted data), S = System performance (CPU, memory, disk speed).
Energy Data Centre	EDC	101	Operational Efficiency	Total energy consumption of a data centre, encompassing energy usage of both IT hardware and supporting infrastructure.	$EDC = \sum E_i$ where E_i represents the individual energy consumption of each component in the system.
Energy Efficiency	Power to Performance Effectiveness (PPE), PUE, DCiE, Data Centre Performance per Energy (DPPE), DCEP, Communication Network Energy Efficiency (CNEE), Network Power Usage Effectiveness (NPUE), Energy Proportionality Coefficient (EPC) and Other Metrics	32, 86, 21, 72, 122, 99, 48, 101, 95, 135, 121	Operational Efficiency	Measures how efficiently the system handles workloads concerning energy consumption. This ratio is a key metric for operational efficiency. Please, refer to each term in this table for further information. Power usage Effectiveness - PUE compares total facility energy to IT-equipment energy, with Data Centre nfrastructure Efficiency - DCiE as its inverse. Japan's Green IT Council's Data Centre Energy Productivity - DCPE metric merges GEC, PUE, IT Equipment Energy Efficiency - ITEE and IT Equipment Utilisation - ITEU for an all-around efficiency score. Power to Performance Effectiveness - PPE measures power per performance, DCEP tracks useful work per energy, CNEE gauges energy per data unit, NPUE isolates network-component energy, and EPC assesses how device energy scales with workload.	Refer to the specific metric for estimation or check references.
Energy ExpenseS (EES)		95	Operational Efficiency	Quantify how the Energy ExpenseS have been altered (i.e. increased or decreased) compared to a baseline scenario after the equipment is upgraded or the introduction of flexibility mechanisms.	$EES = \frac{\sum C_{baseline}(t) - \sum C_{flexible}(t)}{\sum C_{baseline}(t)}$ Where the summations are over the total number of time intervals considered. $C_{baseline}(t)$ represents the energy cost at time interval t under the baseline scenario. $C_{flexible}(t)$ represents the energy cost at time interval t under the flexible energy management strategy.
Energy reuse faction (ERF)	Energy reuse effectiveness (ERE)	32, 86, 21, 72, 122, 99, 48, 101, 95, 135, 121	Operational Efficiency	The ERF, defined in ISO/IEC 30134-6 / EN 50600-4-6, determines the share of the total energy consumption that is reused.	$ERF = 1 - \frac{ERE}{PUE} = \frac{Energy\ Reused}{Total\ Facility\ Energy}$
Energy Waste Ratio	EWR	95	Operational Efficiency	Quantifies the proportion of energy in a data centre that does not directly contribute to IT operations.	$EWR = 1 - \frac{1}{PUE}$

Entropy	Shannon's Entropy, Heterogeneity, information entropy, Additional information Value (AIV), Joint Entropy, individual entropy, information score	8, 27, 128, 120, 131, 71, 114, 132, 133, 48, 12	Data Quality	Entropy quantifies a dataset's randomness or information content—such as spatial and temporal diversity in travel-time predictions. Additional Information Value measures how much entropy reduction a transformation adds. Joint entropy captures the combined uncertainty of two variables, while conditional entropy is the uncertainty remaining in one once the other is known.	**Shannon Entropy**: Shannon entropy measures the average uncertainty in a dataset: $H(x) = -\sum P(x)\, log_2 P(x)$ where $P(x)$ is the probability of an event x. Shannon entropy is foundational in information theory, assessing the information content or disorder in a dataset. **Information Entropy**: quantifies the informational content or uncertainty in datasets. It is widely used as a metric for data valuation, pricing, and quality analysis. **Joint Entropy**: measures the combined uncertainty of two or more variables. It is essential for understanding relationships and dependencies between datasets. $H(x,y) = -\sum\sum P(x,y)\, log_2 P(x,y)$ **Conditional Entropy**: Represents the uncertainty of one variable given another, i.e. evaluates how much one variable reduces the uncertainty of another. $H(x\|y) = \sum P(x,y)\, log_2 P(x\|y)$ **Data Information Entropy**: Assesses the information richness of datasets, particularly in applications like data pricing and quality analysis. Higher entropy indicates more uncertainty or greater information content. **Relative Entropy (Kullback-Leibler Divergence)**: Relative entropy measures the difference between two probability distributions, defined as: **Additional Information Value (AIV)**: Additional Information Value (AIV) quantifies the added informational value by reducing entropy when integrating new data or attributes. **Heterogeneity**: Heterogeneity describes the diversity or variability in data. While not as explicit as entropy, it aligns with entropy principles as a measure of data variety.
Error Rate / Ratio / Count	Uplink / Downlink error rate, trouble tickets, Inter-Server Error Rate (ISER), failure rate	36, 112, 51, 40, 122, 55, 15, 37	Operational Efficiency	Quantifies the proportion of failed operations within a system, network, or process over time. It is a key metric in service reliability, communication networks, machine learning, databases, and cloud computing to assess stability and efficiency.	$Error\ Rate = \frac{Number\ of\ Failed\ Transactions}{Total\ Transactions}$ $Error\ Rate = \frac{Number\ of\ Incorrect\ Data\ Entries}{Total\ Number\ of\ Data\ Entries}$ $ISER = \frac{1}{N(N-1)}\sum\sum BER_{ij}$ where BER is the Bit Error Rate (please refer to [36] for more information on the last expression). Examples of applications involving the bit error rate (BER; see [36]) and other rates are found in IT service monitoring to track success-rate swings and system faults via trouble-ticket counts [112], in machine learning and data marketplaces to assess model accuracy and data quality [40], in databases and big data analytics to detect anomalies and ensure data integrity [51, 55], in cloud and communication networks to evaluate latency, bandwidth and efficiency through

					Uplink/Downlink Error Rate (UDER) and Inter-Server Error Rate (ISER) [122], and in data value chains to gauge system reliability and governance effectiveness [15, 37].
Extensibility		117	Innovation and Growth	Ability to add new features or devices without disrupting existing systems, linked to scalability.	Please refer to [34].
FAIRness Score		73	Data Quality	Evaluates datasets' compliance with the FAIR principles. More information can be found at https://www.go-fair.org/fair-principles/.	Please refer to (Automated tools like CkanFAIR compute this score).
Field Value		12	KPIs and Metrics for Data Monetisation	Denotes the relative importance of a single dataset attribute to a business, model or system. It can be gauged in several ways, including market-based methods (data prices and benchmarks), utility-based measures (impact on outcomes and usage frequency), cost-based calculations (acquisition, storage and maintenance expenses), content-based criteria (uniqueness, completeness and accuracy), risk-based assessments (opportunity cost and loss of information value) and model-based techniques like Shapley values, which quantify each field's contribution to predictive accuracy.	Market-based, utility-based, cost-based, content-based, and risk-based methods, as well as model-based techniques (e.g., Shapley Value), are used to calculate field value.
Fixed to Variable Energy Ratio	FVER, DC-FVER	95, 116	Operational Efficiency	Metric designed to assess the proportion of fixed (constant) energy consumption relative to variable (dynamic) energy consumption.	$FVER = 1 + \frac{Fixed\ Energy}{Variable\ Energy}$ The "1" serves as a reference point, indicating a perfectly efficient data centre where all energy consumption is proportional to the workload.
Format	Format compliance, codification, conformity, available formats	106, 63, 118, 23, 124, 77, 80, 14	Operational Efficiency	Measures adherence to specified data formats to ensure data structure and quality, often calculated as conformity rate.	$Conformity\ Rate = \frac{Number\ of\ Conforming\ Entries}{Total\ Number\ of\ Entries}$
Granularity	Data Frequency, Abundance	75, 30, 23, 120, 78	Data Quality	Refers to the level of detail or precision of the data being collected, stored, and analyzed (or space between time stamps for dynamic data). It indicates how finely data is broken down into its components and can significantly impact the quality and usefulness of the data for various applications.	Direct measurement.
Green Energy Coefficient	GEC	95, 121	Operational Efficiency	Metric that quantifies the proportion of energy used in a data centre that comes from renewable energy sources. It is used to assess the sustainability and environmental impact of data centre operations.	$GEC = \frac{Energy\ from\ renewable\ sources}{Total\ Facility\ Energy\ Consumption}$
Grid Utilisation Factor	GUF	116	Operational Efficiency	Assesses the extent to which a data centre relies on power from the electrical grid versus on-site power generation.	Direct measurement.
Growth Rate		116	Innovation and Growth	Measures the increase in the number of records within a dataset over a specific time. This metric tracks the expansion of data volume and provides insights into data collection and acquisition trends. A higher growth rate indicates enhanced data generation or collection capabilities, while a stable or declining rate may signal data input limitations. It is essential for planning scalability, storage, and future data resource value.	$Growth\ Rate = \frac{Number\ of\ New\ Records}{Time\ Period}$

Funded by the European Union

Hop Distance (HD)	Uplink/Downlink HD (UDHD), InterServer HD (ISHD)	122	Operational Efficiency	Measures the number of intermediate devices data traverses in a network, affecting efficiency and latency.	This document categorises HD as a communication metric and presents it as a direct measurement rather than a derived value.
HVAC System Effectiveness	HSE	95, 116	Operational Efficiency	Helps to measure the overall efficiency of a data centre's cooling system.	Unit Energy/HVAC Energy
Inclusiveness		68	Innovation and Growth	Measures how well a strategy or data-driven initiative invites and leverages contributions from a broad spectrum of stakeholders while still meeting its essential goals. It highlights the value of integrating diverse perspectives and inputs to drive greater innovation and value without sacrificing quality or performance. Key benefits include: **Diverse Insights and Innovation**: Organisations ensure they gather a wide range of viewpoints and expertise, which fuels better decision-making, sparks creative solutions, and delivers stronger outcomes. **Enhanced Stakeholder Engagement**: A focus on inclusiveness deepens involvement from customers, employees, and partners, building trust, loyalty, and collective ownership of results. **Maintaining Quality Amid Participation**: Strikes the right balance between broad participation and high performance, ensuring that every contribution adds value rather than detracts from core objectives.	Can be estimated as key components aggregation. Additionally, it can be formulated depending on the system, as a ratio. For federated learning or distributed data: $Inclusiveness = \frac{Number\ of\ Valid\ Contributors}{Total\ Number\ of\ Contributors}$ For croporate contexts: $Inclusiveness = \frac{\#\ stakeholders\ actively\ engaged}{Total\ Potential\ Stakeholders}$
Inferential Privacy		30	Data Governance and Compliance	Quantifies the probability that an adversary can correctly infer a private parameter from observable data. It aims to ensure that even with observed data, the adversary's ability to infer private information is significantly limited.	Please refer to [30] for further information.
Information Content (IC)		56	Data Quality	Quantifies the amount of unique or novel information in each data packet. It is based on the probability of an event captured by the packet, where less probable (more unique) events hold higher information content. Data packets with low redundancy, often representing unexpected or rare events, have a higher IC value, making them more valuable for real-time and relevant data applications.	Formula for Information Content: $IC(p) = 1 - P(r_{carried} \mid R_{history})$ Where P is the probability, $r_{carried}$ is the data reading in the packet, and $R_{history}$ represents prior readings in the data stream.
Information Diffusion	Data Distribution, Information Distribution	26, 124	Learning and Growth	Addresses the flow of information across various entities, such as individuals, organisations, or systems, and how effectively, quickly, or broadly it can be shared or accessed. Key components include: **Scarcity**: The availability or rarity of the information; scarce information is less likely to spread widely. **Sharing**: The extent to which information is openly shared or kept private. Open data and social sharing platforms promote wider diffusion (i.e. linked to data governance). **Infrastructure**: The technological systems and networks (e.g., databases, cloud systems) that enable information transfer.	Measured based on the key components or use of a proxy, as defined in [26] using the standard deviation of data distribution among nodes: $\sigma = \sqrt{\frac{1}{N}\sum(x_i - \mu)^2}$

Funded by the European Union

				Channels: The media or platforms (e.g., email, social media, enterprise systems) through which information flows, is presented, or advertised.	
Information Frequency		111	Data Quality	Measures how often information is updated, accessed, or used. It is more about the rhythm or rate of interaction with the information.	The explicit calculation of Information Frequency (IF) is represented as: $$IF_r = \frac{FT - \sqrt{(FN - FI)^2}}{FT}$$ Where IF_r is the Raw Frequency Component, FT is the Frequency Tolerance of the Decision Node, FN is the Frequency Requirement of the Decision Node, and FI is the Supplied Frequency of the information. The IF is determined as: IF = (1 if $IF_r \geq 0$ or 0 if $IF_r < 0$) This implies that the True Frequency Component is binary: It is valid (IF = 1) if the raw frequency component is within acceptable limits. It is invalid (IF = 0) if the raw frequency component falls outside acceptable limits.
Integrity	Reliability, Data Prevention, Data Source, Corroboration	39, 11, 46, 23, 76, 129, 124, 100, 134, 55, 6, 43, 25	Data Quality	Data integrity ensures the accuracy, consistency, and reliability of data throughout its lifecycle, from creation to storage and retrieval. Maintaining data integrity is crucial for ensuring the data is trustworthy and useful for decision-making, analysis, and operations.	Can be treated as a KPI computed from several facets of data quality. Alternatively: **Deviation metrics** quantify the proportion of entries conforming to expected values [11]: Deviation Integrity = 1 − Number of Deviations / Total Entries **Gap analysis** focuses on missing data in contexts where completeness is paramount [39]: Integrity Score = 1 − Number of Gaps / Total Expected Entries **Consistency verification** in relational systems ensures referential and semantic correctness [23, 46], for example: Referential Integrity = Valid References / Total References. **Tamper-detection** leverages cryptographic proofs (e.g. blockchain) to measure unaltered blocks [6, 43]: Integrity = Tamper-Free Blocks / Total Blocks **Expert-driven frameworks** assign weights w_i to criteria with scores s_i [23]: $$Integrity\ Score = \sum w_i s_i$$ In large-scale deployments, simple completeness checks yield [76, 55]: Integrity = Complete Records / Total Records **Reliability-based** integrity may also be expressed via Mean Time to Data Loss (MTTDL) and Expected Annual Fraction of Data Loss (EAFDL), which assess how frequently and severely data loss events occur [54].
Internal Rate of Return (IRR)		3	KPIs and Metrics for Data Monetisation	Internal Rate of Return (IRR) is a financial metric used to evaluate the profitability of an investment. It represents the annualised rate of return that makes the net present value	The IRR is the discount rate that satisfies NPV = 0. At this rate, the investment breaks even, with the costs exactly matching the benefits in terms of their present value. Formula for NPV:

				(NPV) of all cash flows (both incoming and outgoing) equal to zero. In other words, IRR is the discount rate at which the present value of an investment's costs equals the present value of its benefits.	$NPV = \sum\left(\frac{C_t}{(1+IRR)^t}\right) - I_o$ Where C_t is the cash flow at time t, I_0 is the initial investment cost and t is the time period.
Interoperability	Compatible, Integration Capabilities	75,11	Operational Efficiency	Ability of different systems, applications, or dataset to seamesly be integrated and/or work together. It ensures that data can be shared, understood, and utilised across various platforms and environments. Key aspects to estimate this as a metric or KPI include Concordance of Data Quality and Governance factors (e.g. same Granularity or similar to a settled standard or Schema).	Interoperability can be expressed as a weighted composite of key dimensions (as a normal KPI): $Interoperability = \sum w_i \times x_i$ where w_i are weights reflecting each dimension's importance and the components x_i represent data-format compatibility, system integration readiness, adherence to exchange standards, semantic alignment, and workflow interoperability.
IT Equipment Energy Efficiency	ITEE	95	Operational Efficiency	The ITEE metric focuses on the useful work related to IT equipment (as opposed to DCPE which focuses on the data centre as a whole).	$ITEE = \frac{Useful\ Work\ Done\ by\ IT\ Equipment}{Total\ Energy\ Consumption\ of\ IT\ Equiment}$
IT Equipment Utilisation	ITEU	95	Operational Efficiency	The ITEU metric focuses on the ratio of actual energy consumption to the rated energy consumption of IT equipment, providing insight into the energy utilisation efficiency of the IT infrastructure.	$ITEU = \frac{Total\ Energy\ Consumption\ by\ IT\ Equipment\ (actual)}{Total\ Energy\ Consumption\ of\ IT\ Equiment\ (Rated)}$
IT Power Usage Effectivenes	ITUE	101	Operational Efficiency	The ITUE is a PUE-type metric for the IT equipment rather than for the data centre.	The compute components include CPU, memory, and other operational-units-based energy consumption. This excludes cooling, power supplies, and voltage regulators. $ITUE = \frac{Total\ Energy\ Into\ IT\ Equipment}{Total\ Energy\ Into\ Compute\ Components}$ Additionally, $TUE = PUE \times ITUE$
Latency	Uplink/Downlink communication latency, interserver communication latency, database access latency, transaction finality time, link/downlink communication latency (UDCL)	36, 47, 51, 91, 97, 19, 90, 122, 56, 102, 6, 5	Operational Efficiency	Measures the delay between the time a request is made and the time a response is received, typically measured in milliseconds (ms). It can also refer to the time it takes for a system to respond to a service (e.g., transaction finality time as noted in [6]).	Latency = Time of Response − Time of Request
Learnability		83	Customer and Market Oriented	Assesses how easily users can learn to use the platform and perform tasks such as data search and visualisation. Learnability is evaluated based on user testing results, focusing on how simple and understandable the data and interface are for new users.	User testing/score system results based on how simple and understandable the data and interface are for new users.
Leave-One-Out (LOO)		8	Data Valuation Techniques	A method for estimating the data value by measuring the difference in performance when a specific data source is different. The LOO value is calculated by comparing the prediction accuracy of a model trained with and without the specific data source (see equation).	LOO(n_i) = v(SN) − v(SN − n_i) Where n_i is the data source/point under analysis

Licensing	License Compliance, Free License, Licensing restrictions	4, 23, 14	Data Governance and Compliance	Measures the proportion of datasets published under specific licensing terms, such as open licenses.	Binary Score system for compliance or the use of Frameworks and standards. Frameworks, like the Open Data Barometer. These frameworks imply a quantitative measurement, which would naturally involve calculating proportions or percentages to assess compliance with licensing standards.
Lifecycle	Shelf life of data	111, 129, 15, 37	Innovation and Growth	The lifecycle refers to the length of time a source of information is expected to remain active. Information is perishable, and after the lifecycle ends, the node (source of information) is assumed to no longer generate value.	Follow [111] for a description of the estimation based on valuation techniques.
Lineage		55	Data Governance and Compliance	Refers to the entire history of the data, including its origin as well as all the transformations, movements, and processes it has undergone throughout its lifecycle.	$Lineage\ Completness = \frac{Elements\ with\ Lineage}{Total\ Number\ of\ Data\ Elements}$
Location YardStick Score		57	Data Valuation Techniques	Measures the contextual value of location data based on its contribution to specific analytical tasks, such as trajectory estimation.	$\text{LYS}(x_{;}\ y, D, M_T) = \max\{0, d_t(M_T(D), y) - d_t(M_T(\{x\} \cup D), y)\}$ Where T is the analytical task for evaluating the value of location data (e.g., trajectory estimation), x is the new location data point, y is the correct or actual outcome for task T , D is the existing dataset (excluding x), M is the model performing task T, d_t is the distance function measuring the difference between the model's output and the correct answer.
Loss and Missed Opportunity Costs	Data root cause remediation	11,129	Data Valuation Technique	Corresponds to revenues and profits lost due to poor data quality. For instance, inaccurate customer email addresses can result in lower revenues as acquired customers cannot be reached for advertising campaigns.	Captures financial losses from process errors or inefficiencies: Loss Cost = Total Direct Losses + Indirect Costs, where "Direct Losses" are quantifiable impacts (e.g. refunds, penalties) and "Indirect Costs" are less tangible effects (e.g. customer dissatisfaction, operational delays) [11]. **Missed Opportunity Cost** represents unrealised revenue due to inaccessible or low-quality data: Missed Opportunity Cost = Potential Revenue − Actual Revenue, with "Potential Revenue" from missed opportunities (e.g. untapped markets) and "Actual Revenue" in the same period [11, 129]. **Root Cause Remediation** assesses systemic error-prevention via COBIT 5 maturity levels (Partially Achieved: 15–50 %, Largely Achieved: 50–85 %, Fully Achieved: 85–100 %) [129], yielding reduced losses and improved efficiency. **Total Cost** integrates both components: Total Cost = Loss Cost + Missed Opportunity Cost.
Loss of Information Value (LIV)	Revenue Loss, linked to Replacement Cost	39, 41, 12, 16	Data Valuation Technique	A metric quantifying the financial impact of losing or compromising information, encompassing both the cost of reacquiring or replacing the information and the cumulative income lost over time due to its unavailability.	LIV = Acquisition Price + Σ_{time} Income Loss + Compliance Costs + Operational Inefficiency Costs Each one of the terms can be neglected as needed to represent the system under question. **Acquisition Price**: Represents the Replacement Cost, which includes expenses for reacquiring or regenerating the data. This could involve purchasing data from external sources or recreating it internally using resources such as labor, technology, and time. **Income Loss:** Total income lost over a period (T) due to the unavailability of data. It includes missed revenue opportunities, reduced customer satisfaction, and delays in service delivery.

Funded by the European Union

					Compliance Costs: Costs related to regulatory penalties, legal liabilities, or fines incurred as a result of non-compliance due to missing data. **Operational Inefficiency Costs**: Poor data management or unavailability can lead to workflow delays, higher corrective costs, and reduced productivity. These impacts reflect indirect costs which are distinct from efficiency losses and often fall under overhead or unexpected operational costs.
Maintainability		11, 118	Operational Efficiency	Maintainability evaluates how easily data systems, databases, and datasets can be managed, updated, cleaned, and enhanced over time. It involves modularity, scalability, ease of updates, error management, documentation, automation, and adaptability.	No clear definition of its estimation or calculation, nevertheless as described in [11], it can be calculated as a metric as "Number of pages with missing meta-information", thus linked to data quality and governance considerations.
Maintenance Frequency		105	Operational Efficiency	The frequency with which a dataset requires maintenance, directly impacting operational costs.	Direct measurement.
Market Adjustment Factor	Discount Price, Full Price	131, 45	KPIs and Metrics for Data Monetisation	A coefficient that adjusts the base market price depending on external market conditions. Reflects factors like market demand, competition, and dynamics influencing the final sale price.	This relationship is expressed as: $$price = \alpha \times price_{estimation}$$ where α represents the Market Adjustment Factor, and $price_{estimation}$ denotes the market-assessed price (refer to MVI).
Market Value of Information (MVI)	Average financial contribution per record	39, 15, 16, 37	KPIs and Metrics for Data Monetisation	This metric calculates the potential income from selling, leasing, or sharing information. It incorporates time, price, exclusive price, and discount rate.	There is no specific formulation on the provided mansucript, neverheless, the values can be defined based on the different components and data value. The MVI price is influenced by several variables including scarcity (S), timeliness (T), additional information value (ΔH), data quality (Z), and costs (C), and can be formulated as: MVI = f (S, T, ΔH, Z, C).
Mean Time to Data Loss	MTTDL	54	Data Quality	A traditional metric estimating how long a system will operate before experiencing data loss. Useful for comparing redundancy schemes and estimating system reliability.	Direct estimation from records or MTTDL is calculated as: $$MTTDL = \frac{E(T)}{P_{DL}}$$ where $E(T) = \frac{1}{n\lambda}$ is the expected time between failures (where n is the number of devices and λ is the failure rate), P_{DL} is the probability of data loss, derived from the transition probabilities $P_{e \to e+1} = (n - e).\lambda.R_e$ where R_e is the rebuild time at exposure level e. Please, refer to [54] for more information.
Metadata	Contextual Information, Documentation features, Profiling	111, 60, 63, 44, 113	Data Quality	Describes the required information in detail, including preferences for format and medium.	May be assessed via a Boolean or scored system based on format adherence or metadata presence. More advanced schemes [63] incorporate completeness (all required fields populated), accuracy (faithful representation of the data), consistency (use of standardised formats and vocabularies) and transparency (context on source, purpose and limitations).
Moderation		80	Data Quality	The Moderation metric tells you how many of your data points lie within the extreme ends of what you'd expect if the data were perfectly Gaussian (i.e. normally distributed).	$$Moderation = \frac{1}{N}\sum \frac{\lvert Y_j^M \rvert}{T}$$

Funded by the European Union

					where N is the number of variables, T is the total number of observations, Y_j^M is the subset of moderated values for variable j.
Mutual Information		71, 22, 5, 12	Data Quality	Measures the mutual dependence between two variables, quantifying the amount of information shared. Applications include feature selection, data compression, and pricing.	Please refer to [71] for its calculation.
NDG		66, 108	Data Governance and Compliance	Metric related to Traceability. Total number of simple decisions that trace up two simple goals in the goal hierarchy.	Please refer to [66] for more information.
NDGI		66, 80	Data Governance and Compliance	Metric related to Traceability. Total number of simple decisions that trace up to simple goals and trace down to the information requirements.	Please refer to [66] for more information.
NDI		66, 108	Data Governance and Compliance	Metric related to Traceability. Total number of simple decisions that trace down to the information requirements.	Please refer to [66] for more information.
Net Present Value		111, 3	KPIs and Metrics for Data Monetisation	Determines the future cash flows of an asset to calculate its value.	$NPV = \sum(\frac{C_t}{(1+IRR)^t}) - Io$ Where C_t is the cash flow at time t, I_0 is the initial investment cost and t is the time period.
Network Metrics		120	Operational Efficiency	These represent metrics that are specific to data centres operational efficiency and include: Diameter Stretch, Path Stretch, Maximum Relative Size, and Network Utilisation.	**Diameter stretch** is the ratio of the modified to the original network diameter: Diameter Stretch = New Network Diameter / Original Network Diameter. **Path stretch** compares altered and original shortest-path lengths between any two nodes: Path Stretch = New Path Length / Original Path Length. **Maximum relative size** captures how much the network's scale (nodes, links, memory) has grown: Maximum Relative Size = New Network Size / Original Network Size. **Network utilisation** is the fraction of bandwidth actually in use: Network Utilisation = Total Used Bandwidth / Total Available Bandwidth
Network Traffic Overhead		6	Operational Efficiency	Measures the bandwidth and resource consumption associated with data transactions, influencing costs and scalability.	Direct measurement.
NGD		66, 108	Data Governance and Compliance	Metric related to Traceability. Total number of simple goals that trace down to simple decisions.	Direct measurement.
NGI		66, 80	Data Governance and Compliance	Metric related to Traceability. Total number of goals at higher levels that trace down to information requirements at lower levels.	Direct measurement.
NID		66, 80	Data Governance and Compliance	Metric related to Traceability. Total number of information requirements that trace up to simple goals in goal hierarchies.	Direct measurement.
NIG		66, 80	Data Governance and Compliance	Metric related to Traceability. Total number of information requirements at lower levels that trace up to simple/complex goals.	Direct measurement.

Funded by
the European Union

Node Value	NV	111	Data Valuation Technique	A decision node in DBV marks a point in an organisation's workflow where a particular piece of information informs a choice; its value is the information's contribution to that decision, weighted by data quality.	The Node Value (VN) quantifies the contribution of a decision node and is calculated as: VN = Vδ × Qf Where Vδ = (Vmax + Vmin) / 2 represents the average decision (Value Range), and Qf = IA + IF is the quality factor based on the accuracy (IA) and frequency (IF) of the information. Although costs, including acquisition (CA), processing (CP), and maintenance (CM), are considered in other aspects of the DBV framework, they are not subtracted directly in this calculation of VN but could be considered.
Number of Sensitive Field		69	Data Governance and Compliance	Measures the number of data columns governed by policies in a data store.	Direct measurement.
Objectivity		11, 17, 25	Data Quality	Objectivity refers to the degree to which data and its analysis are free from bias, personal opinions, or subjective interpretations. Ensures conclusions based on data are valid, reliable, and replicable. Key aspects include unbiased data collection, consistent application of methods, and evidence-based conclusions.	Apart from processes to reduce the incidence of bias in data collection, methods to evaluate their existence can be used. For example: Non-parametric cohort analysis compares metrics (e.g. accuracy, FPR, TPR) across groups without distributional assumptions. **Statistical parity** checks P (Positive | A = a) = P (Positive | A = b), **Distributional skewness** computes (Observed Count − Expected Count) / Expected Count to reveal over- or under-representation. **Equalised odds** requires P (Y* = 1 | Y = y, A = a) = P (Y* = 1 | Y = y, A = b), and **Disparate impact** uses the "80 % rule" P (Outcome | A = a) / P (Outcome | A = b) < 0.8
Open Data Barometer	ODB	4	Data Valuation Technique	The Open Data Barometer (ODB) is a global index created by the World Wide Web Foundation to measure how governments are using open data to promote transparency, innovation, and social impact. It evaluates three main areas: readiness, which looks at how prepared governments and societies are to support open data; implementation, which assesses the availability and quality of published data; and impact, which examines how open data benefits politics, the economy, and society. By combining these factors, the ODB ranks countries and highlights opportunities for improvement.	Using expert surveys, case studies, and data analysis, the Open Data Barometer (ODB) assigns each country a score based on criteria such as open licensing, machine readability, and data accessibility, revealing both successes and gaps in open-data adoption. Policymakers, researchers, and advocates rely on the ODB's readiness, implementation, and impact scores—detailed in its methodology at https://opendatabarometer.org/leadersedition/methodology/.
Openness	Sharing	65, 70, 124, 93, 15	Data Governance and Compliance	Openness measures the degree to which datasets provide a confirmed open license and format. The metric evaluates factors like open licensing, format compatibility, metadata quality, and ease of access.	Frameworks such as [104] can be used.
Operational Cost (OPEX)	Maintenance Cost, System Cost, Storage Cost, contractual Costs, Labor Costs, Utility Costs, Transaction Fees, Publishing Cost,	111, 105, 41, 96, 6, 96, 81, 12, 37	KPIs and Metrics for Data Monetisation	Operational Expenditure (OPEX) includes the recurring costs necessary to keep an organisation running. These cover data-related activities such as storing, organizing, backing up, securing, and updating data. Costs vary depending on dataset size, tools, and organisational needs. In data centres, OPEX also includes hardware maintenance and contractual costs.	Direct estimation. Operating Expenditure (OPEX) can be represented as an equation that aggregates all recurring operational costs incurred over a specific period. The general formula is: OPEX = Maintenance Cost + Storage Cost + Contractual Costs + Transaction Fees + Application Cost + Other Recurring Costs

	Service Cost, Application Cost, Cost			**Contractual** costs are fixed payments to third-party providers for services like data processing, analysis, or infrastructure. These may involve cost-plus terms based on data volume or service time and can represent major expenses, especially when contractors work on-site. Managing these is key to data monetisation profitability. **Data maintenance** costs include: 1) Storage: Cloud or physical hardware. 2) Management: Tools and personnel to organise and access data. 3) Backup/Recovery: Systems to prevent data loss. 4) Security: Measures against breaches and corruption. 5) Cleansing/Updating: Ensuring data remains accurate. 6) Compliance: Meeting legal and regulatory standards. 7) Optimisation: Upgrades for system performance. **Labour** costs can include hardware installation, staff for data handling and system maintenance, consulting services, and training expenses. **Utility** costs cover electricity, cooling, equipment maintenance, and other overhead such as safety gear, transport, and dedicated server or office spaces used solely for Big Data operations.	Each term can be estimated separatedly, for example, Maintenance Cost = (Cm · f · Size), where Cm is the maintenance cost per unit size, and f is the maintenance frequency. Storage cost can simply be calculated as Storage Cost = Cs*Size, where Cs is the cost per unit size of data.
Other Utilities		95, 116, 121, 67, 72, 122	Operational Efficiency	Other Utilities (e.g. Thermal and Air management metrics) fall under Operational Efficiency alongside HVAC performance measures. Key indicators include the Recirculation Index (RI), which quantifies how much hot exhaust air is recirculated rather than expelled; the Capture Index (CI), which assesses how efficiently the cooling system directs exhaust back into the intake without mixing with supply air; and the Data Centre Cooling System Efficiency (DCCSE), the ratio of cooling infrastructure power to total IT load. See the cited references for a complete list of metrics.	Please refer to the provided references for a full list.
Oversight	Audit	129	Governance and Compliance	Assesses an organisation's capability to monitor and validate compliance with governance standards, processes, and policies. It includes mechanisms for periodic audits to ensure data integrity and proper governance implementation.	Even if not specified, common metrics related to audit can be evaluated. For example, to calculate a data oversight metric, key components could include: **Audit Frequency:** Percentage of planned audits conducted. **Audit Coverage**: Percentage of systems or processes reviewed. Coverage = Audited Systems / Total Systems **Compliance Rate**: Percentage of compliant processes. Compliance Rate = Compliant Processes / Reviewed Processes **Resolution Rate**: Percentage of issues resolved on time. Resolution Rate = Resolved Issues / Total Issues

Funded by the European Union

					Repeat Findings Rate: Percentage of recurring issues. Repeat Findings Rate = Repeated Issues / Total Issues Combine these components using weighted averages as a KPI:
Ownership		41, 23	Data Governance and Compliance	This metric evaluates the outright ownership of the data set, including any licensing restrictions and service agreements. Ownership impacts the value of the data based on who controls it and under what terms it can be used or shared. Assessed based on licensing terms and ownership rights detailed in service agreements.	Ownership can be calculated based on three components [41]: **Licensing (L)**: Full ownership: 1, Limited restrictions: 0.5, Highly restrictive: 0. **Service Agreements (S)**: No dependence: 1, Moderate dependence: 0.5, High dependence: 0. **Outright Ownership (O)**: Full ownership: 1, Partial/shared: 0.5, None: 0. The Ownership score is then calculated as an aggregated metric
Payment-Accuracy Tradeoff		127	Data Valuation Technique	The Payment–Accuracy Tradeoff as a KPI measures the balance between the cost of compensating individuals for sharing their private data and the accuracy of the insights gained from that data. It reflects how much payment is required to achieve a desired level of accuracy in data analysis, considering the privacy constraints.	The core idea is that there's a fundamental trade-off between how accurate the collected data can be and how much one must pay people to share it. Please review [127] (complex forumation) for more information.
Performance per Watt	PpW	95	Operational Efficiency	Measures the actual energy efficiency of every device in the data centre and how it is used.	The PPW approach uses a relative performance indicator for each individual asset. This indicator is calculated by the types of hardware and capabilities learned from an asset inventory of that device. This Performance Indicator (PI) is a simple measurement for getting relative performance of the device in question. PpW = (PI × Avg Device Utilisation/Watts). For example, check https://www.datacenterknowledge.com/energy-power-supply/pue-is-dead-the-case-for-performance-per-watt.
Plausibility	Credibility, Believability, Match between system and the real world	39, 11, 60, 136, 107, 137, 44, 124, 134, 114, 17, 25	Data Quality	Plausibility in the context of data refers to the degree to which the data is reasonable, credible, and aligns with expectations based on known relationships, patterns, or rules. It assesses whether the data makes sense within the context it's being used, ensuring that it reflects real-world situations or follows logical assumptions.	As mentioned in report [33], the assessment carried out focuses on the data points that present extreme values which are a possible consequence of errors in the compilation of the reporting obligations. They use a scoring system to define the credibility of outliers. Further information in [114] is found on how to estimate a plausibility score. Basically, a plausibility score is computer for each record by looking at each of its attributes and asking "Is this value reasonably close to the bulk of the data for that attribute, and does it actually exist?", then a binary value is given. a gap is given for the plausibility based on distributions or other considerations. In other words, the spread can be measured in two ways: the ordinary standard deviation, or a more robust version (the interquartile range divided by 1.35) that downweights extreme values. Together, this provides a simple count of how many of a record's values look like they come from the core of the data distribution.
Policy		69, 129	Data Governance and Compliance	Policies are high-level principles or rules that an organisation establishes to guide decision-making and behavior.	Check Data Principles for further information.
Power Density Efficiency	PDE	95, 67		A variation of PUE that can provide insight into the improvements to both the IT equipment and the supporting cooling system. The proposed metric enables evaluation of the impact	Please refer to [67] for the estimation equation.

				of physical changes inside the racks on energy efficiency, which is not possible using the common metrics above.	
Power Usage Effectiveness (PUE)	Carbon Usage Effectiveness (CUE), Total Energy Usage Effectiveness (TUE), Data Centre Efficiency (DCE), Data Centre Infrastructure Efficiency (DCiE), SPUE, pPUE, P UE1−4, SI-POM, H-POM	32, 86, 21, 72, 122, 99, 48, 101, 95, 135, 121	Operational Efficiency	PUE is probably the most well-known data centre efficiency metric and is defined in ISO/IEC 30134-2 / EN 50. SI-POM and PUE are essentially equivalent metrics used to evaluate the energy efficiency of a data centre's infrastructure by comparing total facility power consumption to IT equipment power consumption. The IT H-POM metric, on the other hand, focuses on the efficiency of power conversion within IT equipment, aiming to minimise the overhead power relative to the actual compute power used. Directly connected to PUE is CUE which measures the carbon emissions associated with the energy consumption of a data centre. It quantifies how much carbon dioxide (CO2) is emitted per unit of energy used by IT equipment	$PUE = \frac{Total\ Facility\ Energy\ Consumption}{Equipment\ Energy\ Consumption}$ $DCiE = \frac{1}{PUE}$ $CUE = \frac{Total\ CO2\ Emissions}{Equipment\ Energy}$ $CUE = PUE \times CEF$ $TUE = PUE \times ITUE$ For the rest of PUE, check corresponding ISO/IEC references.
Precision		111, 11, 118, 25	Data Quality	Precision refers to the level of detail with which data is captured, measured, and represented. It describes how specific the data values are and impacts the reliability of results and analyses. Further concepts are provided in the main document.	Direct measurment.
Privacy Budget		40, 115	Data Governance and Compliance	The privacy budget quantifies the amount of privacy loss tolerated in a system. It is denoted by ϵ (epsilon), which controls the trade-off between privacy and accuracy in differential privacy algorithms. A smaller ε provides stronger privacy guarantees at the expense of data utility. By treating ε as a budgeted resource, organisations can balance data utility against individual privacy guarantees in a transparent, mathematically grounded way.	Calculated using the parameters ϵ (privacy loss) and δ (failure probability) of differential privacy. It determines the noise level added to computations, based on the sensitivity of the function (Δf), which measures the maximum change caused by a single individual's data. Common mechanisms include the Laplace mechanism and the Gaussian mechanism. Please check references for further information.
Privacy Level	Privacy sensitivity, propensity score	127, 35, 74, 115, 125	Data Governance and Compliance & KPIs and Metrics for Data Monetisation	Measures the privacy risk or leakage in data reporting. It helps assess or control the trade-off between privacy, data utility, and the cost of compensating individuals for privacy loss. A lower propensity score indicates synthetic data closely resembles real data without revealing sensitive information.	The privacy level ϵ in differential privacy is defined by the inequality: $Pr[M(D) \in S] \leq \epsilon \cdot Pr[M(D') \in S],$ where M is the mechanism, D and D' are datasets differing by one element, and S is a set of outputs. Smaller ϵ values provide stronger privacy by limiting how much the presence or absence of a single individual affects the output a single individual affects the output.
Process Failure Costs		11	Data Valuation Technique	Poor quality data causes processes to fail. For example, inaccurate mailing addresses cause correspondence to be misdelivered, algorithms to fail, or ML/AI component to have low accuracy or misled results.	As a recommendation (not based on reference): To estimate process-failure costs, first list which processes suffer from poor data quality and the errors behind each failure. Quantify direct costs—rework labor, wasted materials, fines—and indirect costs like lost opportunities, churn, and reputation harm. Track time spent fixing issues and convert it to monetary value, use historical records for guidance, and supplement with surveys or interviews for wider impacts. Finally, sum all costs and check the total against industry benchmarks or past data for validation.
Processing Value Ratio		111	Data Valuation Technique	Determines the value added through processing data into information. It is calculated differently based on the type of value transfer (Type H or Type L).	Type H (High Value Retention): Data typically gains between 5% and 10% value after processing. (i.e. Vp = 1 + 0.05 to Vp = 1 + 0.1) Type L (Low Value Retention): The value gained is proportional to the ratio of processing time (t_p) to acquisition time (t_a), calculated as Vp = 1 + t_p / t_a

Protection Expense		69	Data Governance and Compliance & Data Valuation Technique	Cost to apply protection measures (e.g., encryption, access control) to specific data stores. Includes costs of implementing security controls, maintenance, and resource allocation.	Direct estimation.
Proximity		5	Data Quality	Proximity is important in applications like fog computing and habitat monitoring, where the location of events or sensors affects data valuation.	Direct measurement.
Quality Factor (QF)		111	Data Quality	Links overall quality of information, including accuracy and frequency, to its potential value for business innovation and growth.	QF = Accuracy + Frequency + (Others; such as Completeness + ...)
Quality of Service (QoS)	Data connected to service levels, service characteristic index, Service Level Agreement	105, 129, 20	Customer and Market-Oriented	Refers to the performance level and reliability expected by clients when accessing datasets. Key points include latency constraints (ensuring acceptable latency for reconstructing datasets) and adherence to Service Level Agreements (SLAs), which outline expected performance metrics such as response times and availability.	Quality of Service (QoS), a core KPI measured against SLAs, combines technical and user-focused measures: the Service Characteristic Index [20], Data Quality Metrics [129], Service Performance and SLA Compliance checks [105], and binary pass-fail tests of critical performance criteria [129].
Quantity of Private Projects		28	Innovation and Growth	Represents the quantity of private projects/services/datasets available online.	Direct measurement.
Quantity of Public Projects		28	Innovation and Growth	Represents the quantity of public projects/services/datasets available online.	Direct measurement.
Range		80, 46	Data Quality	Quantifies the proportion of data values that fall within predefined lower and upper bounds, reflecting the validity of data within expected limits.	The Range metric ([80]) is calculated using the following formula: $range = \frac{1}{N}\sum\frac{\lvert Y_j^*\rvert}{T}$ where N is the number of variables, T is the total number of observations, Y_j^* represents the subset of values for variable j that lie within the predefined lower and upper bounds.
ReconstructionCost	Packet Recovery Score	105, 56	Data Valuation Technique	Reconstruction Cost refers to the financial or computational effort needed to restore lost or corrupted data, typically from backups or redundant sources. It reflects the cost-effectiveness of recovery in systems focused on data integrity and availability. Packet Recovery Score, often used in networking contexts, measures how effectively lost or corrupted packets are recovered during transmission. It indicates the resilience of systems like IoT or wireless sensor networks against packet loss [56].	$Ce \approx (r \times sp + w \times sq),$ where r is the read cost, w is the write cost, and sp, sq are the sizes of the parent and child datasets, respectively.
Regulatory Compliance	Compliance Cost	134, 40, 107, 62, 100	Data Governance and Compliance	Measures whether the data complies with relevant legal and regulatory standards. Estimated financial cost to business of not keeping data for compliance/regulation purposes. It can be defined as a quantitative or qualitative measure that evaluates the extent to which data adheres to the predefined rules, standards, or requirements. Compliance metrics focus on ensuring data consistency, reliability, and alignment with organisational or regulatory frameworks.	Compliance can be calculated using formulas that assess adherence to defined standards. A basic metric is: $Compliance = \frac{N_c}{N_t}$ where N_c is the number of compliant entries, and N_t is the total assessed. When multiple standards apply, a weighted formula provides more accuracy:

					$$compliance = \frac{\sum w_i c_i}{\sum w_i}$$ Here, c_i is the compliance score, w_i the weight, and n the number of criteria. Quantitative metrics may include adherence to standards (e.g., ISO 20022 [100]), compliance costs (e.g., audits and penalties [40]), and risk scores for non-compliance [62]. Qualitative measures involve governance and ownership assessments during integration [129], often tracked via audits or frameworks like COBIT [107].
Relevance	Decision Support Capabilities, Relevance Factor, Priority Score, Importance, existence	111, 39, 75, 11, 103, 136, 118, 50, 113, 55, 20, 61, 17, 5, 25, 12	Data Valuation Technique	Refers to the usefulness of data for business processes, ensuring data is meaningful, reusable, and adaptable to changing demands. Includes relevance factor, priority score, and existence checks for completeness.	Can be measured like Utility in which the objective or (operational context) is different, i.e. specific for only one business instead of a domain. Different approaches can be used to estimate Relevance, for instance, by using the DBV approach specified in [111].
Renewable Energy Factor (REF)	On-site Energy Fraction (OEF), On-site Energy Matching (OEM)	119	Operational Efficiency	Measures the proportion of a data centre's total energy consumption that is sourced from renewable energy. OEF evaluates the share of a data centre's energy needs met through on-site renewable energy generation. OEM assesses how well the on-site renewable energy generation aligns with the data centre's energy consumption patterns.	$$REF = \frac{E_{ren}}{E_{DC}}$$ where: E_{Ren} is the total energy derived from renewable sources (kWh), E_{DC} is the total energy consumed by the data centre (kWh). $$OEF = \frac{\int R(t)dt}{\int L(t)\,dt}$$ where: $R(t)$ is the renewable power produced on-site at time t, $L(t)$ is the total power load of the data centre at time t, $$OEM = \frac{\int \min(R(t), L(t))dt}{\int L(t)dt}$$ $min(R(t), L(t))$ represents the amount of on-site renewable energy that is directly used to meet the load, $L(t)$ is the total data centre power load over time.
Reputation	Popularity	11, 117, 124, 100, 55, 17, 5	Market Penetration	Refers to the perceived quality, reliability, and trustworthiness of data, impacting decision-making and governance. Key aspects include quality, governance, transparency, and ethical use.	Assessed through both quantitative and qualitative methods, including historical accuracy, trust scores, system reliability, popularity metrics, error response, standards compliance, and social feedback such as user ratings and endorsements. Combining these factors yields a comprehensive reputation score [107].
Response Time		119, 53, 91, 50, 76, 102	Operational Efficiency	Total duration from the initiation of a request to the system's response. It encompasses multiple factors such as processing time, communication latency, database access latency, transaction processing speed, and network-induced delays.	Response Time = $t_{response} - t_{request}$
Responsiveness	Time Metrics, Speed	51, 137, 118, 39, 11, 44, 50, 13	Operational Efficiency	Responsivenes is more than response time. Relates to the responsiveness of a system or service, measuring data processing speed and timely updates. Speed refers to the raw	It is computed as an aggregated KPI from its sub-components. A simplistic approach can use: $$Responsiveness\ =\ t\ at\ Query\ Completion\ -\ t\ at\ Query\ Start$$ And for speed:

Funded by the European Union

				processing capability of a system—such as how quickly a database query is executed—responsiveness emphasises the system's ability to react promptly to user inputs or requests.	$Speed\ =\ \frac{Number\ of\ Operations}{Processing\ Time}$
Return on Investment (ROI)	Profit, Estimated Benefit to Business from Using Data	111, 3, 132, 133, 15	KPIs and Metrics for Data Monetisation	Measures the profitability or efficiency of an investment, comparingnet gains to costs.	$ROI = \frac{Value\ Produced - Cost\ Incurred}{Cost\ Incurred}$
Revenue	Economic Benefits	42, 2, 50	KPIs and Metrics for Data Monetisation	Refers to total income from normal business operations.	$R\ =\ Price\ per\ Unit\ \times\ Number\ of\ Units\ Sold$
Risk Cost	Regulatory Risk	69, 16	Data Governance and Compliance & Data Valuation Technique	Measures the financial impact of a potential data breach, including fines, legal fees, reputation damage, and operational disruption. Risk Cost is defined as the total monetary impact incurred by an organisation if a data breach, compromise, or loss occurs. It is computed by quantifying various financial and reputational damages and multiplying these by the number of records at risk.	$Data\ Risk\ Cost\ =\ (Penalties\ +\ Litigation\ Costs\ + Reputation\ Damage + \frac{Notification}{Recovery\ Costs}$ $+\ Operational\ Disruption)\ \times\ Number\ of\ Records\ at\ Risk$ Furthermore, a scale can also be used for quick assessment of risk cost based on organisational thresholds: 0: Negligibly small costs. 1: Low costs (e.g., notification only, no significant penalties). 2: Moderate costs (e.g., some legal and recovery expenses). 3: High costs (e.g., litigation and major operational disruptions). 4: Intolerably high costs (e.g., catastrophic penalties, reputation damage).
Risk Score	Risk Management Index	75, 69	Data Governance and Compliance	Refers to the systematic process of identifying, assessing, and mitigating risks associated with data assets. Risk score is calculated as a weighted sum of multiple risk factors, including sensitivity level and protection percentage, to reflect the overall risk level of a data store.	Risk is calculated by combining weighted conceptual and architectural attributes of data. Conceptual factors include business value, sensitivity level, policy coverage, and potential loss, while architectural factors assess protection, exposure, access, and usage patterns. These are normalised and aggregated into a risk score using a weighted formula. This method aligns with ISO 31000 and can be enhanced with FMEA techniques, incorporating severity, occurrence, and detectability to refine risk prioritisation.
Rival Access Loss		12	Data Valuation Technique	Estimates the financial cost to a business if competitors gain access to its data.	In the absence of a direct formula, a qualitative approach is used, evaluating data by its accessibility (open, restricted, proprietary), strategic value, uniqueness, and timeliness. Open data generally has minimal impact, while proprietary data entails higher costs due to its exclusivity and value. Stakeholder input from surveys or interviews helps classify data impact as low, medium, or high. A flexible scoring model can be expressed as: $Cr\ =\ f\ (I)\ +\ f\ (U)\ +\ f\ (T)\ +\ f\ (A)$ where $f(I)$, $f(U)$, $f(T)$, and $f(A)$ represent the importance, uniqueness, timeliness, and accessibility of the data, respectively. This framework supports both qualitative evaluation and the foundation for future quantitative analysis.
Runtime	Processing Time	102, 110	Operational Efficiency	Refers to the time taken to execute tasks or processes in a data system. Runtime overhead evaluates performance improvements or degradations relative to a baseline.	$Runtime\ =\ T_{end} - T_{start}$ $Processing\ Time\ =\ T_{CPU}\ +\ T_{I/O}$

Funded by the European Union

Satisfaction	Feedback (user), user's satisfaction, user attitude, behaviour, business user satisfaction, Degree of satisfaction, Business User Satisfaction	83, 106, 35, 137, 44, 118, 23, 52, 84, 77, 15, 37	Customer and Market-Oriented	Measures user satisfaction with the platform interface, layout, and data presentation. Combines objective metrics like accuracy and relevance with subjective feedback to gauge satisfaction.	While not all approaches are directly derived from the cited manuscripts, related discussions can be found within them. Readers are encouraged to consult the original sources for full methodological details: **TOPSIS** (Technique for Order Preference by Similarity to Ideal Solution): Satisfaction is computed as the relative closeness of a system's performance to an ideal solution [83,106]. **SERVQUAL**: Satisfaction is evaluated by comparing user expectations (E) with perceptions (P) across service quality dimensions: $SERVQUAL = \sum(P_i - E_i)$ Widely applied in services and logistics, this method highlights gaps between expected and perceived performance [137, 118]. **Satisfaction Rate**: For usability studies, satisfaction is the proportion of users who complete tasks successfully: $Satisfaction\ Rate\ = \frac{Successful\ Respondents}{Total\ Number\ of\ Respondents}$
Scalability	System Concurrence Processing Capabilities, Elasticity	105, 117, 27, 110, 50, 76, 24, 90	Operational Efficiency	Evaluates the ability of a system to handle increased data volumes. Elasticity allows dynamic allocation of resources to meet demand, ensuring performance and cost-efficiency.	As a KPI, it can be linked to the main components referred in the Taxonomy Figure (Redundancy, Interoperability, Extensibility, Adaptability) or use, alternatively, changes in metrics under different circumstances, depending on the system type: Throughput [90, 117]. Response Time [50]. Velocity [117, 50]. Concurrency [50, 90]. Elasticity [76, 117]. Baseline Comparison: Performance compared to a baseline, such as R2D2 processing TB-scale data within 5 hours, showcasing scalability [105].
Scarcity		124, 131, 25, 12	KPIs and Metrics for Data Monetisation	Measures the availability or rarity of information. Scarce information may have limited distribution, impacting its monetisation potential.	Although no explicit formulation is provided in the reviewed texts, scarcity can be inferred as inversely related to the accessibility of similar datasets. One approach defines scarcity as $S\ =\ 1 - \frac{A}{T}$ where S is the scarcity score, A is the number of accessible datasets similar to the target, and T is the total number of datasets within the market or domain. A higher S indicates greater exclusivity and limited access [25]. Scarcity may also be expressed through a demand-supply ratio (S = D / Sp), where D is the demand and Sp is the supply of comparable datasets. A higher ratio signifies increased scarcity, emphasizing the dataset's limited availability [12].
Schema		11, 24	Data Governance and Compliance	Defines the presence or absence of desired attributes, including clarity, comprehensiveness, flexibility, robustness, and precision of domains.	Schema as a metric is measured by evaluating several aspects of a schema's structure, consistency, and alignment. Key approaches include: **Completeness**; **Flexibility and Adaptability**: Assessing the schema's ability to handle semistructured or schemaless data formats, often using techniques like schema evolution tracking or flexibility scoring.; **Semantic Alignment**: Using ontology-based or graph-based methods to determine how well schema

					elements relate to semantic relationships and integrate with other datasets.; **Standardisation**; **Interoperability**; **Error Detection and Correction** [11, 24]. Readers are encouraged to review [18] where different metrics specific for tracking Schema are included.
Security Composite Efficiency Indicator		62	Data Governance and Compliance	Combines multiple efficiency indicators to assess overall system performance in cybersecurity. Metrics include equipping coefficient, technical readiness, and penetration testing resilience.	$E = (E_U + E_{TG} + E_{US} + E_K + E_{HP}) / 5$ where E_U: equipping coefficient with cyber defence means, E_{TG}: technical readiness coefficient, E_{US}: equipping coefficient with serviceable cyber defence means, E_K: staffing coefficient with IT system administrators, E_{HP}: staffing coefficient with service personnel.
Security Level Index	Security, Access Security	11, 50, 23, 62, 134, 20, 61, 17	Data Governance and Compliance	Measures the protective measures implemented to safeguard data from unauthorised access and breaches. Key aspects include confidentiality, integrity, availability, and compliance with regulations.	As noted in [20], the Security Level Index (SLI) is not defined through a specific formula, but its evaluation draws from multiple frameworks and indicators. According to [62], security is assessed using operational metrics like incident response time, threat detection, and vulnerability repair. In [17], IoT security is estimated via trust scores, incorporating historical reliability and access control. [61] Integrates security into overall data value through metrics such as encryption strength and compliance, while [134] focuses on access control in high-volume environments. A generalised SLI can be constructed using the metrics outlined in [95], including: Average Comparisons per Rule, Defense Depth, Detection Performance, Vulnerability Exposure, Firewall Complexity, Latency, and other network-level indicators.
Server Compute Efficiency (ScE)		95	Operational Efficiency	Designed to assess the efficiency of compute resources within a data centre. The metric aims to evaluate how effectively servers perform computational tasks relative to the energy they consume.	To measure server efficiency, some tools like the Server Efficiency Rating Tool (SERT, https://www.spec.org/sert/) are available.
Service Agreement		41	Governance and Compliance	A formal contract defining data ownership, licensing restrictions, compliance standards, and usage rights, evaluated for clarity, regulatory alignment, and flexibility to support data utility and accessibility.	Not defined in [84], but could be linked as a Boolean defining the existence or nonexistence of Service Agreements within the data. Further levels could be incorporated based on the level of the restrictions and standards involved.
Shapley Fairness	Fairness Metrics	53, 2, 73	Data Quality	Related to allocating resources proportionally in systems like VMs or ensuring data fairness, e.g., reducing biases.	Shapley Fairness allocates rewards based on the Shapley Value and fairness principles like efficiency, symmetry, null player, additivity. Chek references and Shapley value for information.
Shapley Value			Data Valuation Technique	A baseline metric to measure the importance of each data provider, dataset, or feature in a coalition of data sources. Reflects the marginal contribution of each data source to the overall system, providing a fair value and reward mechanism.	$\phi_i = \sum \frac{\lvert K\rvert!\,(\lvert N\rvert - \lvert K\rvert - 1)!}{\lvert N\rvert!} \cdot [v(S \cup [i]) - v(S)]$ Where ϕ_i is the Shapley value for player (data source) i, N is the set of all players (data sources), S is a subset of players excluding i, $v(S)$ is the value (e.g., forecasting accuracy) obtained from subset S of players, $v(S \cup [i]) - v(S)$ is the marginal contribution of i.
Social Welfare	Social Benefits	115, 50	Innovation and Growth & Customer and Market-Oriented	Represents the difference between the aggregate utility of all model requesters and the total privacy cost of all data owners in a data marketplace.	Please refer to [115] for more information and equations.

Funded by
the European Union

Space Cost	Disk Occupation	126, 51	Data Valuation Technique	Amount of storage space required to store data, which can vary depending on the file format and storage mechanisms used. The cost can be linked later to the cost associated with maintaining such information.	Direct measurement.
Space, Wattage, and Performance. (SWaP)		32, 86, 21, 72, 122, 99, 48, 101, 95, 135, 121	Operational Efficiency	Evaluates the efficiency of a system by considering its performance output relative to the space and power it consumes.	$SWaP = \frac{Performance}{(space \cdot power)}$
Statistical Parity		82	Innovation and Growth	A fairness metric in ML ensuring equal probability of inclusion in the positive predicted class for sensitive groups. Synthetic data can be adjusted to meet statistical parity requirements.	$SPD = P(\hat{Y} = 1 \mid A = minority) - P(\hat{Y} = 1 \mid A = majority)$ Where $\hat{Y}$ is the Model predictions and A is the Sensitive attribute group.
Stochastic Divergence	Identity-based Exact Match, Jensen-Shannon Divergence, Wasserstein distance	24, 81	Data Quality	Stochastic Divergence is a measure of the difference or similarity between statistical distributions, focusing on their probabilistic or statistical characteristics. It encompasses metrics that assess how closely two distributions align or diverge.	Jensen-Shannon Divergence assesses similarity by averaging the Kullback-Leibler divergences of two distributions relative to their mean. It is symmetric and bounded between 0 (identical) and 1 (maximally different), making it well-suited for comparing soft attribute distributions. Another measure is the p-Wasserstein distance, which captures the minimal cost of transporting one distribution into another. Check references for mathematical formulation.
Storage Cost	Cost of data Storage	105, 81	Data Valuation Technique	Storage costs cover the infrastructure, management, and upkeep of data systems, including hardware, cloud services, energy use, and compliance. Poor management of redundant or unused data can greatly inflate these costs, especially in data lakes and enterprise systems [81,105].	$Storage\ Cost = Cs \times Sv$ where Sv is the size of the dataset and Cs represents the cost charged by the storage provider (e.g., Azure, AWS) per unit size for a specific storage tier.
Stranded Power Capacity Per Rack (SPCR)		116	Operational Efficiency	Allocated but unused power within individual server racks.	$SPCR = Budgeted\ Rack\ Power - Actual\ Rack\ Power$
Structure	Data Structure	11, 124, 24, 15	Data Quality	Defines the organisation and storage format of data, enabling efficient access, modification, and management.	Structure is assessed in terms of how well data elements are interconnected, standardised, and harmonised within an information system. (i.e. require Governance Specifications). Use a Boolean to mark conformity.
Success Rate / Ratio / Count		112, 68	Operational Efficiency	Proportion of operations, tasks, or processes successfully completed relative to total attempts. Evaluates system reliability and effectiveness.	$Success\ Rate = \frac{Number\ of\ Successful\ Operations}{Total\ Number\ of\ Operations}$
Support		120	Operational Efficiency	Proportion of records in the dataset that satisfy both the antecedent and consequent of a rule, highlighting common patterns.	Direct evaluation; use a Boolean to mark conformity.
Synchronisation		1	Data Quality	Refers to the process of ensuring that two or more data sources (like databases, devices, or systems) have the same, up-to-date information Can be explicitly defined in quantitative terms based on its role in assessing the alignment and coherence of systems or datasets.	Synchronisation can be quantified using the following approaches: **Time offset** (∆T) measures the discrepancy in time alignment between the source (Tsource) and the target (Ttarget). It is defined as: ∆T = |Tsource − Ttarget|

Funded by the European Union

					Synchronisation error quantifies the average deviation in synchronisation over a set of n observations. It is calculated as: $Syncrhonization\ Error = \sum \lvert \Delta T_i \rvert$ **Maximum synchronisation error** can be expressed as: Max Error = max(|ΔTi|)
Syntactic Similarity	Levenshtein Distance, edit distance, cosine similarity, Q-gram distance, semantic similarity, etc.	128, 120	Data Quality	Metrics to assess how similar data values are in terms of their syntax, e.g., based on character similarities. Levenshtein Distance is used as a primary metric to calculate this similarity. Additional metrics include Edit Distance, Cosine Similarity, Q-gram Distance, and Jaccard Coefficient.	**Levenshtein Distance** computes the minimum number of edits (insertions, deletions, or substitutions) needed to transform one string into another. It is particularly effective for identifying near-duplicates in data fields. For example, strings with a distance threshold ≤ 2 may be considered similar. **Cosine Similarity** measures the cosine of the angle between two text vectors (e.g., term-frequency vectors) and is useful for comparing multi-word fields or document-level data. A similarity score ≥ 0.8 often indicates a strong match. **Q-gram Distance** divides strings into overlapping substrings of length q and calculates distance based on matching substrings. This method is particularly effective for identifying partial matches.
System Capacity	Memory, CPU, bandwidth, storage capacity	119, 51, 110, 97, 86, 21, 72, 122, 20, 102, 95	Technology and Infrastructure	Refers to the maximum amount of work, load, or use that a system can handle without significant degradation in performance. Includes CPU, memory, storage, and bandwidth capacity. This information can be used also to generate ratios for System Utilisation metric.	Direct measurement.
System Robustness	System Stability	68, 50	Operational Efficiency	Measures the ability of a system, model, or process to remain stable and perform well despite disturbances, faults, or unexpected inputs. Key aspects include system stability, resilience to adversarial attacks, error handling, and generalisation.	Use a weighted combination of key components. Consider at least a combination of Stability, Resilience, and the ability to maintain functionality despite errors, faults, processes, and attacks. An example is included in [68], robustness in the martFL architecture is calculated as the percentage of malicious data providers (DPs) whose low-quality or adversarial local models are successfully excluded during model aggregation.
System Utilisation	CPU utilisation, memory utilisation, disk utilisation, deployed hardware utilisation (DH-UR), etc.	112, 119, 32, 47, 51, 95	Operational Efficiency	Measures how efficiently and effectively a system's resources—such as CPU, memory, storage, and network are being used during a given period.	$Utilization = \frac{Times\ or\ Quantity\ of\ resourse\ Used}{Total\ Available/monitoring\ time\ or\ Quantity}$ $DH - UR = \frac{Number\ of\ Active\ Servers\ Running\ Live\ Applications}{Total\ Number\ of\ Servers\ Deployed}$ The complexity of the representation is contextual and, as mentioned in [95], their values can be treated as links to KPIs, since most of them represent ratios or percentages (See as examples Table 7 of [95]).
The Green Index (TGI)		95	Operational Efficiency	Metric offers flexible green benchmarking. While defined using performance-per-watt, it can incorporate any energy-efficiency metric. TGI focuses solely on IT equipment power but can be extended to include cooling infrastructure.	As mentioned in [95], it can be estimated as performance-per-watt metric: $TGI = \frac{Useful\ Work}{Component\ Energy\ Consumption}$

Funded by the European Union

The Value of Information for Business (VIB)		39	Data Valuation Technique	Assesses how useful information is for business processes, focusing on accuracy, completeness, relevance, and delay in receiving the information.	$VIB = Completeness \times Accuracy \times Availability / Ubiquity$ or $VIB = Accuracy \times Completeness \times Relevance / Delay$
Throughput	Transaction Processing Speed	119. 47, 51, 91, 90, 61, 6, 5	Operational Efficiency	Measures the number of requests a system processes over a specific period (e.g., requests per second, minute, or hour).	Throughput is calculated based on the total data volume transferred and the time taken for the transfer process. $R_P = \sum B \times Size(b_i) / T$ Where R_P is the pipeline throughput. B is the number of batches transferred. $Size(b_i)$ is the size of each batch b_i., T is the total time taken from the start to the end of the pipeline processing. Additionally, batch throughput is defined as: $R_b = \frac{Size(b)}{l_b}$ Where l_b is the batch latency, which includes both queuing delays and processing time.
Timeliness	Data Freshness	39, 11, 10, 4, 70, 46, 60, 98, 103, 136, 107, 118, 23, 38, 124, 77, 113, 114, 55, 45, 61, 17, 80,14, 43, 5, 25, 48, 16	Data Quality	Measures the availability of data when needed, ensuring data is up-to-date and accessible within an appropriate timeframe. Includes frequency of updates, and speed of availability.	The most relevant timeliness metric is data freshness. It is the delay since the last business action recorded in the database. Hence, it can be used directly. Since data timeliness measures the currency of a copy of data stored in a database, it shoud consider the timestamps of business actions and the time when the data pipeline (ETL process) loaded the data into the database, data warehouse or data lake. Another method for calculating timeliness involves summing decay rates across attributes $Timeliness = \sum [\exp(-decline \cdot age(i,j)) \cdot a_{ij})]$ where decline is the decline rate for the attribute j, $age(i,j)$is the age of the attribute value a_{ij}, and a_{ij} is the indicator for the presence of the value. Alternatively, the currency of data refers to how promptly data becomes available relative to when it is needed: $Timeliness = \max\left[1 - \left(\frac{age\ of\ the\ data\ value}{shelf\ life}\right), 0\right]$ Where Age of the Data Value is the time difference between when the real-world event occurred (data creation) and the timeliness assessment, Shelf Life is the maximum time a data value remains up-to-date, Exponent (s > 0) is a parameter determined by experts to control the sensitivity of the timeliness metric.
Total Equipment Utilisation (TEU)		32, 86, 21, 72, 122, 99, 48, 101, 95, 135, 121	Operational Efficiency	Could be considered within System Utilisation Metric but given its broad use, has been set apart. It encompasses the overall facility's resource usage. It includes cooling, power distribution, and other supporting infrastructure components.	$TEU = \frac{Total\ Useful\ Workload\ (including\ cooling, power, etc.)}{Total\ Facility\ Equipment\ Capacity}$

Traceability	Addressability, NGD, NDI, NDG, NID, NDGI, verifiability, provenence documentation, audit trail coverage, compliance rate	11, 66, 55, 108	Data Governance and Compliance	Refers to the ability to track and verify data throughout its lifecycle. It involves understanding data lineage—how data flows from its source through various transformations to its final use—and provenance, which captures the origin and history of the data, including any changes made. Maintaining audit trails ensures that all modifications are recorded, including who made them, when, and what was changed, supporting accountability and transparency. Traceability also plays a key role in meeting regulatory and governance standards by demonstrating data integrity and responsible data handling. Additionally, addressability defines how the origin of data can be identified and, if needed, contacted or referenced. Other metrics useful to estimate traceability includes Data Lineage Completeness measures the percentage of data elements for which lineage information is available. Provenance Documentation assesses the percentage of data elements with documented provenance. Audit Trail Coverage evaluates the percentage of data changes that are recorded in an aud it trails. Compliance Rate measures the percentage of data processes that comply with established governance and regulatory standards.	Calculating data traceability can be complex, as it involves multiple dimensions of data management. However, a combination of metrics can help assess the level of traceability in data systems. Below are some potential metrics and a formula to calculate a traceability score: **Lineage Completeness** = Elements with Lineage / Total Number of Data Elements **Provenance Documentation** = Elements with Provenance / Total Number of Data Elements **Audit Trail Coverage** = Number of Changes Recorded / Total Number of Changes **Compliance Rate** = Number of Compliant Processes / Total Number of Processes An overall traceability score can be calculated by combining these metrics. For example: **Traceability** = (Lineage + Provenance + Audit + Compliance Rate) / 4 This formula averages the scores of the individual metrics to provide a comprehensive view of data traceability within an organisation. Each component can be scored on a scale (e.g., 0 to 100), and the final score can help organisations assess their data traceability and identify areas for improvement. For additional details, refer to the descriptions of traceability requirements in [66] and [108].
Traffic Energy	Management and Monitoring Traffic Energy (MMTE)	36	Technology and Infrastructure	Accounts for the power required by sensors to generate data, the energy used for data transmission across multiple network hops, and the computational resources needed to store and process incoming information. Additinally, some specific component of data management can be measured, as defined in [36].	$MMTE = CNEE \times Management\ and\ Monitoring\ Traffic$
Traffic Ratio	Management and Monitoring Traffic Ratio (MMTR), Internal Traffic, External Traffic	36	Technology and Infrastructure	Represents the proportion of a specific type of network traffic relative to the total traffic in a network or data centre. It is used to analyse the composition of traffic, distinguish between different categories (e.g., internal vs. external, management vs. application-specific).	$MMTR = Total\ Data\ Centre\ Traffic\ /\ Management\ and\ Monitoring\ Traffic$ The traffic ratio for specific components: $Application\ Traffic\ Ratio = 1 - MMTR$
Transparency	Objective Measurement	63, 82, 129	Innovation and Growth & Market Penetration & Governance and Compliance	Refers to the ease of identifying, understanding, and interpreting data and processes. Includes clarity in presenting results and consistency in analysis.	In data governance, transparency is measured using structured assessment frameworks like DAMA-DMBOK Framework (Data Management Body of Knowledge), COBIT 2019 (Control Objectives for Information and Related Technologies), or ISO/IEC 15504 (Data Quality Standard) For data transformation and AI models, Transparency is assessed based on explainability, fairness, and model documentation.
Trustworthiness	Assurance, trust score	41, 118, 124, 29, 15, 37	Innovation and Growth-Oriented	Refers to the believability or trustworthiness of data, often based on its integrity, reliability, and consistency. From [29]: Trust Score (TS) is derived from reputation and credibility without involving a centralised authority.	A general equation for Trustworthiness (Trust Score, TS) can be formulated based on the components outlined in the literature. For example: TS = w1 · R + w2 · C + w3 · Q + w4 · T where: R = Reputation Score, derived from verified identity sta tus and historical trust ratings. C = Credibility Score, based on feedback and ratings from past interactions. Q = Data Quality Score, including completeness, accuracy, and consistency. T = Transaction History Score, evaluating the

					frequency and success rate of transactions. w1, w2, w3, w4 = Weighting factors, predefined based on importance.
Typicality		80	Data Quality KPIs or Metrics	Measures how well a data point aligns with expected or "typical" patterns within a dataset. It helps identify outliers or unusual events using statistical or machine learning methods. Use statistical distributions and statistical tests to check it.	z = X − μ / std Where X is the data point, μ is the mean, and std is the standard deviation.
Understandability		70, 11, 63	Data Quality	Reflects how clearly a dataset and its components—such as field names, units, and metadata—can be interpreted and used by the intended audience. It involves ensuring that field names are clear and unambiguous [70], units are explicitly defined and consistently applied [70, 11], and metadata offers sufficient context to describe data structure and relationships [63, 11]. It also includes identifying barriers to interpretation, such as inconsistent formatting or missing explanations [63, 70].	**Binary Scoring**: Individual fields or datasets are scored as either understandable (1) or not understandable (0). Aggregated scores provide an overall metric of understandability for the dataset [70]. **Qualitative Feedback**: User interviews and schema evaluations are conducted to identify ambiguities and evaluate schema clarity [11]. **Issue Analysis**: Specific barriers such as unclear metadata or inconsistent formatting are identified and categorised to improve dataset usability [63].
Uniformity			Defined in this work	Uniformity as a metric measures the consistency of data representation across datasets, systems, or within a single dataset.	Uniformity can be estimated by entropic considerations or as a metric that evaluates the proportion of data entries that adhere to a defined standard compared to the total number of entries in the dataset. The formula for Uniformity is given by: $Uniformity = N_{Entries} / Total,$ where $N_{Entries}$ is the count of data entries that conform to the defined standard, $Total$ is the total number of data entries in the dataset.
Uniqueness	Percentage of Duplicate Data, Concentration, Redundancy	39, 11, 70, 136, 63, 38, 64, 134, 9, 80, 43, 5, 25, 48	Data Quality	Ensures that each record in a dataset is distinct, avoiding duplicates. This also can be linked to data systems, in which the reference of Redundancy is linked to the number of repeated datasets that support System Robustness.	Depending on the type of uniqueness, it should include (1) Identification of duplicates: Ensuring no multiple entries exist for the same entity [11, 9] or (2) Maintenance of distinct records: Verifying that each record is represented uniquely within the dataset [39, 11]. Based on this, it can be defined as follows: $Data\ Uniqueness = Number\ of\ Unique\ Records / Total\ Number\ of\ Records$ $Percentage\ of\ Duplicate\ Data = Number\ of\ Duplicate\ Records / Total\ Records$
UPS Metrics	USF, UCF, UPFC, UPF, UPEE	116, 95	Operational Efficiency	Metrics related to the performance of Uninterruptible Power Supply. These metrics are general energy efficiency or performance metrics related to that unit, such as Energy Efficiency (UPEE), which measure of how UPS converts input power into usable output power while minimizing energy losses.	Different representation given the specific metric. Please refer to the provided references for more information.
Usability	Usage, ease of use, friendliness	11, 65, 136, 41, 110, 44, 118, 23, 50, 124, 77, 25	Data Quality	The ease and efficiency with which quality data (Data-Value, Accuracy, Integrity, Completeness) can be effectively accessed (Communnication, Accessibility, Timeliness), understood (Clarity), and correctly be utilised (Easy-to-use or Utilisation & Performance, Relevance or Utility) by users to accomplish specific tasks.	Estimated as a KPI by summing weighted main components from taxonomy or from those specified in the explanation of this metric. Furthermore, usability and utility are interlinked, with the difference that utility is related to data valuation techniques, while usability has been linked to a wider perspective.
User Frequency	User Count or Concurrent Users	28, 12	Customer and Market Oriented	Tracks the number of concurrent users accessing data, with an associated cost for each additional user.	Direct measurement (e.g. Count the number of active session IDs at a specific point in time and use it again comparing IDs).

Funded by the European Union

Utility	Relevance, Application characteristic index, retention	111, 10, 30, 136, 124, 94, 20, 16	Data Valuation Technique	Utility, in the context of data valuation, refers to how effectively a resource—such as a dataset or system—supports the achievement of intended goals. Unlike usability, which focuses on ease of interaction, utility measures functional contribution to outcomes. While related to relevance, utility is broader; relevance is domain-specific and tied to particular use cases, whereas utility reflects overall potential impact.	Utility, as specified in this work, could be represented mathematically as a continuous function U(i, j, k, l), where i denotes the organisational activities that could use the data, j represents the data/processes needed to generate information (e.g. if data is distributed and steps required to construct), k corresponds to the quality dimension (e.g., accuracy or timeliness), and l indicates the penalty terms (e.g. like costs and legal binding associated with the data). This formalisation allows for the systematic evaluation of utility as a function of the resource's effectiveness in achieving strategic and operational objectives. The general structure of relevance (to calculate the utility function is) R(i, j, k, l), whith the index previously explained. If a process or activity does not influence a dataset or vice versa, the utility value is defined as U (i, j, k, l) = 0. To evaluate the total utility, you can aggregate the relevance: $Utility = \sum\sum\sum\sum w_i w_j w_k w_l \, R(i,j,k,l)$ As a metric, utility can be assessed using Kolmogorov statistics, particularly in scenarios where utility reflects the preservation of a dataset's functional value after transformations such as obfuscation or anonymisation. The Kolmogorov statistic quantifies the similarity between the original and transformed data distributions. Additionally, utility and relevance can be evaluated through expert-based scoring systems that consider contextual factors. When formalised, such scoring can be converted KPIs.
Validity		46, 38, 134, 55, 43, 25	Data Quality	Refers to the compliance of the information with the business rules that describe it (e.g. the age of a person must be Integer). It seems similar to Veracity but it covers the aspect of accuracy and precision of the data concerning the intended use. Validity in [43] is defined as the adherence of data values to specified criteria, ensuring that each data entry is logically sound and meets predefined conditions. This is typically done by applying specific rules to the dataset. For example, a rule might state, "Asset Cost in any Asset record must be greater than zero" which would flag any entries with a non-positive asset cost as invalid.	Rule Application: Each record in the data table is checked against one or more validity conditions. These conditions can include checks on single fields (e.g., "Purchase Price > 0") or multiple fields (e.g., "Install Date is before Warranty Expiration Date"). Policy Evaluation: For each rule, the system identifies the number of records that satisfy the validity condition versus those that fail. $Validity\ Score = \frac{Number\ of\ Valid\ Records}{Total\ Recoreds\ Evaluated}$
Value Added	value-added, Diminishing Value (Opposite Direction)	11, 41, 130, 17	Data Valuation Technique	Value-Added and Diminishing Value metrics reflect the relevance and economic utility of data over time. While Value-Added emphasises the contribution of data to decision-making or operational efficiency, Diminishing Value highlights the temporal decline in data's utility, particularly critical in sectors like IoT and real-time monitoring systems, where up-to-date information is essential for maintaining competitive advantage [17, 41, 11].	Value-added can be estimated qualitatively or through cost-benefit analysis, linking data improvements to operational or economic gains. In contrast, diminishing value is assessed by monitoring the temporal decline in data utility, usage, or demand—particularly in real-time applications. These assessments often rely on user evaluations, scoring frameworks, or cost analyses as discussed in [41, 11]. As emphasised in [17], value is also influenced by how well data serves its purpose within a specific context. Moreover, data value attenuation can be modeled mathematically as a nonlinear decay process, where value decreases over time in inverse relation to rising entropy. Check references for more information.
Value of Privacy	Privacy cost	127, 115, 61	KPIs and Metrics for Data Monetisation	Minimum payment required for sharing data at a specific privacy level (ϵ). Balances privacy loss with data utility.	Check references for quantitative information.
Value Range		111	KPIs and Metrics for Data Monetisation	Describes the potential value achievable using required information. This value is assessed within a range, where moderation ensures values remain within reasonable or typical limits.	Vδ = (Vmax + Vmin) / 2 Where Vmax and Vmin represent the ranges from expert knowledge.

Funded by
the European Union

				Based on the DBV method, it incorporates an income approach for estimating benefits of intangible assets.	
Variety	Multifacetedness	49, 41, 24, 134, 114, 13	Data Quality	Refers to the diversity of data types and formats. Variety emphasises the complexity and richness of data attributes for comprehensive analysis.	Variety, or multifacetedness, refers to the diversity of data types, formats, sources, and features within a dataset. It can be measured using metrics such as the number of distinct data types, sources, and attributes, as well as schema variation and semantic diversity. Entropy-based measures and sparsity analysis further help quantify variability within categorical features or across datasets, providing a comprehensive assessment of variety [49, 24, 134, 114].
Velocity	Frequency Parameter	134, 20, 48, 119, 111	Data Quality	Velocity describes the rate at which data is generated, ingested, and processed over time.	Direct measure. Furthermore, in [111], velocity can be related to the frequency parameter.
Views		92, 37	Market Penetration	Represents the number of times a dataset's page is accessed by users.	Direct measurement.
Visualisation		106, 23	Customer and Market-Oriented	Refers to the graphical representation of data using charts, graphs, maps, and diagrams. Effective visualisations improve clarity, accuracy, and relevance for quick insights and decision-making.	Although there is no universally defined reference framework, visualisation can be assessed by direct user's assessment, then account for the variety of visualisation types, customisation options, interactivity (e.g., filtering, real-time updates), and clarity of design. Additional considerations include dataset integration and accessibility compliance. These factors can be measured, both quantitatively and qualitatively, and combined into broader usability frameworks. A sample metric can be calculated as follows: Visualisation Score = W_1(Dataset Coverage) + W_2(Feature Variety) + W_3(Usage Metrics) + W_4(Expert Ratings) Where Wi corresponds to the weight assigned based on the importance of each metric.
Volatility		11,102	Data Quality	Volatility, as a data quality metric, refers to how frequently data changes or how long it remains valid before becoming outdated. It is closely linked to timeliness and currency, and may also indicate how often data deviates from expected values or business rules. In statistical and financial contexts, volatility is typically measured as the standard deviation or variance over time, reflecting the magnitude of fluctuations in a variable's value.	Three Possibilities $Volatility = \frac{Storage\ Time - Update\ Time}{Total\ Time}$ $Volatility = \frac{Number\ of\ Changes\ from\ a\ reference}{Total\ Time\ of\ Observation}$ $Volatility = \sqrt{\frac{1}{N-1}\sum(x_i - \mu)^2}$ Where N is the number of data points, x_i are the data points, and μ is the mean.
Volume	Quantity, Entries	75, 49, 106, 51, 8, 107, 41, 120, 134, 55, 13, 17, 15, 37	Data Quality	Represents the total amount of data available for analysis, influencing reliability, pattern recognition, and machine learning. Adequate quantity must pair with high quality for actionable insights.	Direct measurement.
Wasserstein	Wasserstein Distribution	81	Data Quality	Measures the difference between two probability distributions, quantifying how much the empirical distribution deviates from the true distribution.	Please refer to [81] for information.

Water Usage Effectiveness	WUE	95	Operational Efficiency	It measures the amount of water a data centre uses in relation to the energy consumed by its IT equipment.	$WUE = \frac{Total\ Water\ Usage}{Equipment\ Energy}$ $WUE_{workload} = \frac{Total\ Water\ Usage}{Useful\ Work\ Output}$
Weighted Coverage Function		27	Data Valuation Technique	This function assigns weights to database instances, adjusting prices for queries based on the seller's inputs. It reflects information disclosure.	The weighted coverage function measures how sensitive a query is to changes in a dataset. It works by comparing the query's output on the original dataset to its output on slightly modified versions, called neighbouring instances—these differ from the original by one or more records. Each of these neighbouring datasets is assigned a weight based on its relevance or likelihood. The function sums the weights of all neighbouring instances where the query result changes, capturing how much influence individual data points have on the overall output [27].
Winning Rate		42	Customer and Market-Oriented	Winning rate is defined in the context of online ad allocation as a **fairness metric**. Measures the percentage of queries a data provider successfully serves, relative to the queries they are eligible to serve. Higher rates indicate valued offerings.	$Winning\ Rate = \frac{Number\ of\ Queries/Impressions\ Won}{Number\ of\ Queries\ /\ Non-zero\ Bids\ Participated}$