Project No. 957117

Project acronym: EVIDENT

Project Title:

bEhaVioral Insgihts anD Effective eNergy policy acTions

Deliverable 5.3

Data Documentation

Programme: H2020-LC-SC3-EE-2020-1
Start date of project: December 01, 2020
Duration: 36 months

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 957117

Document Control Page

Deliverable Name	Data Documentation
Deliverable Number	D5.2
Work Package	WP5
Associated Task	Task 5.2
Covered Period	M18-M30
Due Date	February 2023
Completion Date	31st of January 2023
Submission Date	12nd of February 2023
Deliverable Lead Partner	CERTH
Deliverable Author(s)	Georgios Sidiras (CERTH), Christos Ntoumanopoulos (CERTH), Tilemahos Efthimiadis (JRC), Panagiotis Sarigiannidis (UOWM), Anna Triantafyllou (UOWM), Athanasios Liatifis (UOWM), Karagiannidis Georgios (UOWM), Fragulis Georgios (UOWM), Karypidis Paris (DUTH)
Version	V1.0

Dissemination Level		
PU	Public	**X**
CO	Confidential to a group specified by the consortium (including the Commission Services)	

Document History

Version	Date	Change History	Author(s)	Organisation
0.1	12th of August, 2022	Table of contents, Initial version	Georgios Sidiras	CERTH
0.4	19th of December, 2022	Initial version	Tilemahos Efthimiadis	JRC
0.6	3rd of January, 2023	Contributions to Sections 2, 3 and 4	George Sidiras	CERTH
0.7	10th of January, 2023	Contributions to Sections 2, 3 and 4	Panagiotis Sarigiannidis, Anna Triantafyllou, Athanasios Liatifis,	UOWM

			Karagiannidis Georgios, Fragulis Georgios	
0.8	16th of January, 2023	Section 5	Paris Karypidis	DUTH
0.9	20th of January, 2023	Section 6	George Sidiras	CERTH
1.0	31st of January, 2023	Review comments	George Sidiras	CERTH

Internal Review History

Name	Institution	Date
Paris Karypidis	DUTH	29th of January, 2023
Tilemahos Efthimiadis	JRC	30th of January, 2023

Quality Manager Revision

Name	Institution	Date
Dimosthenis Ioannidis	CERTH	January 30, 2021

Table of Contents

List of Figures

List of Tables

Acronyms

Acronym	Explanation
AI	Artificial Intelligence
API	Application Programming Interface
CERIF	Common European Research Information Format
CERN	European Organisation for Nuclear Research
CLDQ	Cost-effect of Low Data Quality
CMDQM	Comprehensive Methodology for Data Quality Management
CMM	Capability Maturity Model
CRIS	Current research information system
DMAIC	Define, Measure, Analyse, Improve, Control
DMP	Data Management Plan
DOI	Digital Object Identifier
DQA	Data Quality Assessment
DQAF	Data Quality Assessment Framework
DQMHD	Data Quality Methodology for Heterogeneous Data
DX.Y	Deliverable X.Y
EAD	Encoded Archival Description
EHR	Electronic Health Record
EOSC	European Open Science Cloud
ERA	European Research Area
ETL	Extract transform load
EU	European Union
GDPR	General Data Protection Regulation
HIQM	Hybrid Information Quality Management
IT	Information Technology
KPIs	Key Performance Indicators
MIQA	Methodology for Information Quality Assessment
ML	Machine learning
NOACD	National Open Access Centre Desks
OODAM	Observe-Orient-Decide-Act Methodology for Data Quality
ORP	Other research products
PDQA	Practical Data Quality Approach
PhD	Doctor of Philosophy
PID	Persistent Identifier
R&D	Research and Development
R&I	Research and Innovation
RDM	Research Data Management
ROAR	Registry of Open Access Repositories
ROI	Return on Investment
TBDQ	Task-Based Data Quality Method
TDQM	Total Data Quality Management

TIQM	Total Information Quality Management
TX	Task X
URN	Uniform Resource Name

Executive Summary

The EVIDENT project explores the role of behavioural interventions that enable energy consumers to make more efficient energy choices. Data is one of the drivers of this change enabling entirely new opportunities. Accessing and interconnecting data is vital to drive organisational growth and innovation. All data, especially those related to scientific data, are retrievable to be reused to contribute to new research or even to continue old ones. As well as ensuring that people are in control of their information is needed to build trust among stakeholders in the data economy.

This deliverable is the outcome of Task 5.2 "Data documentation" and aims to present the data collected and utilised in the context of the EVIDENT project. It is the first out of two deliverables, along with Deliverable 5.4 "Updated Data documentation", that presents the key indicators of data quality, such as non-response and attrition, discusses how they affect the interpretation of the results, proposes ways to make the data comprehensible, replicable and publicly available, and, finally, presents approaches to address data anonymity.

Deliverable 5.3 constitutes a document that reviews methodologies and presents ways to enhance data quality, handle non-response and attrition, make datasets publicly open and deal with data anonymization. A complete data catalogue among the available dataset will be reported in Deliverable 5.4 ‘Updated data documentation’.

1. Introduction

The data that organisations have in their possession is a critical asset for designing and offering quality services that meet the needs of their customers. The loss or leakage of such data can have severe consequences, even affecting the organisation's viability. This is one of the main reasons why companies aim to secure and protect their data. Therefore, the value created is not only from the data itself but from its use. For example, the same data information can be used for sending invoices, marketing and sales purposes, etc.

Data-driven applications, along with machine learning (ML) technologies, have been integrated into every aspect of an organisation, shaping the services and customer experience. Such examples consist of automated assistants, personalized healthcare, and real-time optimized planning.

This deliverable is focused on data quality by outlining a framework for defining and monitoring specific quality metrics according to an organisation's needs. Emphasis is placed on understanding the criticality of quality data towards the innovation and growth of an organisation. Additionally, the deliverable explores various data elements that can be used for data documentation. Finally, the opportunities arising from sharing of research data are investigated.

1.1 Purpose of the Deliverable

This deliverable aims to explore the data documentation for Evidence-Based Policy [1] and to describe the processes and tasks implemented in the EVIDENT project. All types of data can be considered a piece of evidence [2]. The documentation method provides the basic structure and how the data is organized. Data documentation can be achieved with metadata standards usually associated with forms describing the relevant data.

1.2 Relation with other Deliverables and Tasks

The deliverable receives as inputs D3.3 'Data collection and management' and D5.1 'Impact evaluation plan and policy measures'. In addition, the deliverable is closely related to all deliverables of Work Package 5 - 'Policy measures', since it discusses the technical aspect of how the collected data should be processed to be easily discoverable by the research community and be used for policy briefs.

1.3 Structure of the Document

This deliverable is structured as follows:

- Section 1 serves as an introduction to this deliverable.
- Section 2 provides an overview of the relevant literature with respect to data quality and key data quality indicators.
- Section 3 lists and describes various data documentation approaches.
- Section 4 outlines techniques for enhancing the discoverability and anonymisation of research data.
- Section 5 presents the EVIDENT directions for documenting the data.
- Section 6 concludes this deliverable.

2. Data Quality and Key Data Indicators

High-quality data contribute towards offering high-quality services to customers, complying with regulations, facilitating strategic decision-making processes, and increasing overall operational efficiency. Organisations seek to use data to obtain insights that will eventually generate profits. Data is the foundation of various applications and systems that deal with different business functions in an organisation. Likewise, it plays an essential role in organisations' applications related to business intelligence, data mining or even decision support. Data also help in data-driven enterprise resource planning and customer relationship management systems. Big data is continuously formed under various applications and systems in organisations. Since data and information form the basis for decision-making, they need to be carefully managed to ensure that they are timely, accurate, complete, and easy to locate. Moreover, data should be further distilled into information and managed effectively to create revenue. Nevertheless, low-quality data can introduce several problems, which is why organisations try to manage their data effectively. To highlight the impact of having high-quality data, the business classification of data should be examined.

2.1 Categorizations of Data

As shown in Figure 1, data can be categorised into master, reference, transactional, and historical data. Additional information about metadata is presented in Section 3.3 Metadata Standards .

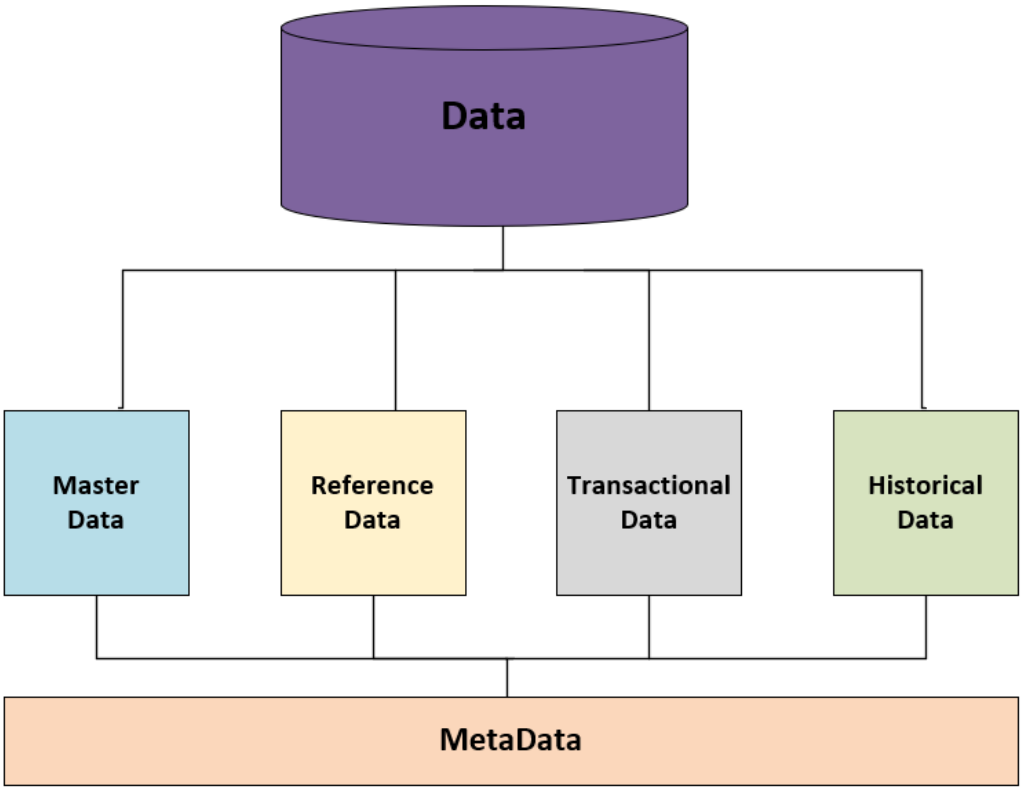

Figure 1: Categorizations of data

2.1.1 Master Data

Master data are essential business information that describes an organisation's core entities and supports its operations. They also play a significant role in the daily digital processes of the organisation, by defining the key business characteristics, such as customers, products, and accounts [6]. One aspect of master data is that they do not include any transactions. However, they are usually used in transactions multiple times. Any errors in master data can have a significant impact on the costs. For example, supposing the address of a customer is incorrect, the orders and sales will be affected. Therefore, a high error rate in master data leads to higher costs.

2.1.2 Reference Data

Reference data are sets of allowed values corresponding to descriptors referenced and shared with other systems, applications, and repo-like data from master data transactions. Reference data have become valuable by extending and reusing references. Such examples are country codes, post codes, and product codes. Reference data must be separated from master data, usually consisting only of a list of allowed values and text descriptions describing the values.

2.1.3 Transactional Data

Transactional data describe organisational events and are commonly contained in large data sets. These data represent an organisation's relevant internal and external events, such as orders or payments. Transactional data are information captured during the actual trading process. These entries are associated with the reference data and have a time dimension.

2.1.4 Historical Data

When a transaction is completed the transactional data are converted into historical data. These data contain events and are vital in terms of security and predictability. Organisations that collect historical data periodically check and compare it with older data to see if there is a problem. Finally, the type of data can be used for forecasting purposes.

2.2 Overview of Data Quality

Data quality refers to the ability of the data to meet the business needs, as well as the assessment of the data's suitability to serve the business's purposes in each context. The overall objective is 'fitness for purpose'. From a business perspective, high data quality is achieved when the data meet the consumers' needs. High-quality data are free from errors and feature the required characteristics that are summarised in Table 1 [37].

Table 1: The characteristics of the data make them suitable for use

Characteristics	Preferred characteristics
Complete	Pertinent
Valid	Contextual
Correct	Easy to read
Unique	The right level of detail
Current	Unambiguous
Reliable	Easy to understand
Consistent	

Evaluating data quality involves considering two distinct elements, which are intrinsic data quality and contextual data quality. The intrinsic data quality focuses on the inherent characteristics of the data, such as its accuracy, format, and ease of access. These characteristics are typically independent of the situation or environment in which the data is being used. For instance, when considering demographic information, it is important to ensure that data elements such as age and salary are accurately represented as numerical values and cannot contain negative numbers.

The second aspect is contextual data quality**,** which is associated with the content, the purpose of the data, and the decisions to derive from the data or even to define the intended recipient of the data. Data quality does not have a single standard, even within a single organisation.

Data quality depends on the context where the data will be used and on compliance with the applicable requirements. To maintain a provable level of data quality, an organisation must determine the degree of compliance. The data quality should then be determined, knowing that the minimum effective level is constant. The DMAIC model, that is define, measure, analyse, improve, and control, is a useful asset for evaluating and improving data quality.

2.3 Data Quality and Cost

Data errors often start with minor errors, such as an incorrect product code or date format. However, these errors can spread exponentially as they are distributed through the systems since an organisation's data are not static, but linked to other information systems. As data move, they affect systems differently, as well as their involvement in business processes [3], [4]. Therefore, the more connected the systems are, the greater the impact of errors on the data quality.

To accurately examine the significance of poor data quality, we need to consider its impact and how it negatively affects business users and customer satisfaction, as well as the increasing operational costs [5]. Poor data quality does not reflect the real situation in an organisation and contributes to establishing a misleading perception of the organisation. These data may be also involved in the formulation of key performance indicators (KPIs) for decision-making processes. Ensuring completeness and high-quality data is necessary for meeting a project’s or an organisation’s objectives. Some indicative examples of bad data quality are depicted in Figure 2.

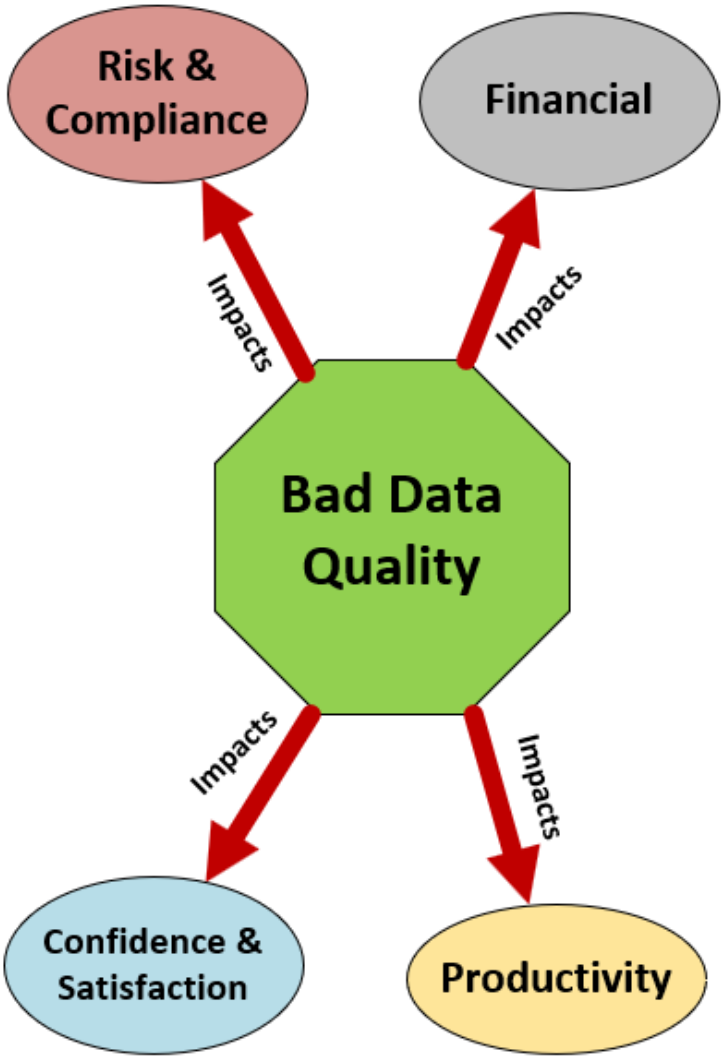

Figure 2: Impact of bad data quality

The cost of low-quality data can be very high. A process that produces data, which is consumed by different business processes within an organisation, is usually located far away from the point of data entry into the system. As data move between processes, the cost of correcting them increases. The phrase "1:10:100" is usually adopted when describing the cumulative effect of the cost of fixing an error as it moves through various stages (Figure 3).

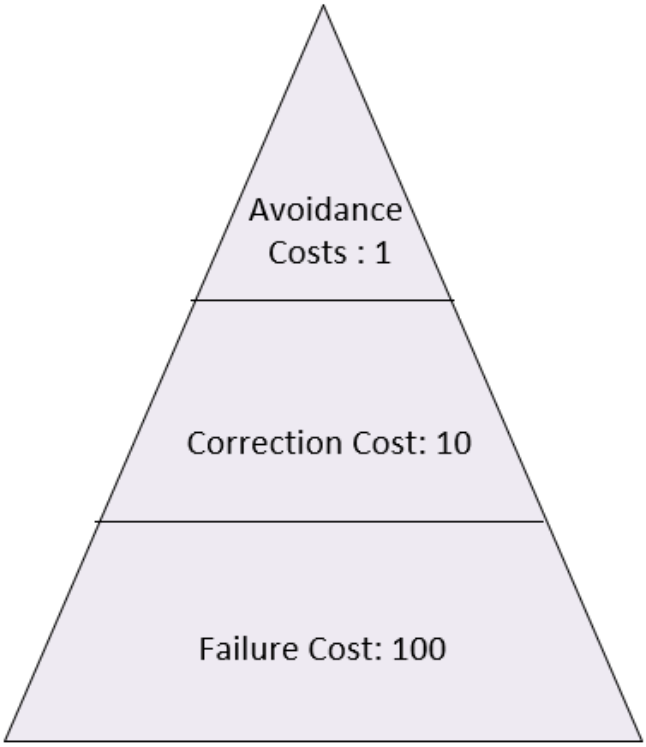

Figure 3: Prevention, Correction and Failure costs

It would be unrealistic to declare that perfect quality can be achieved in any domain, which means that certain compromises are unavoidable. The level of data quality that is suitable is determined by the significance of the data, the needs of the business, as well as the cost and time required. The cost of resources and time to attain and sustain the desired quality level should be weighed against the return on investment and the advantages obtained from that specific quality level. In this direction, the questions that need to be addressed are as follows:

- *What are the minimum acceptable data quality levels?*
- *It is necessary to take corrective actions if the error rate is low?*
- *Is the required standard of data quality worth the resources needed to attain it?*

In data quality management, it is not necessary to have a completely error-free data set. Instead, there is a certain threshold of acceptable quality levels that should be identified and accepted. This threshold is determined by considering the specific needs and requirements of the data set in question. The criticality of the data is a crucial factor in determining the minimum acceptable quality level, as some data sets may be more sensitive and require a higher level of accuracy. This acceptable quality level can be established by evaluating the data within its intended context of use. Once this level is established, ongoing monitoring should be performed to ensure that it is maintained and not impacted by new or altered business processes.

2.4 Data Quality Dimensions

Data quality must be measured in order to evaluate, maintain, and control it. Quantifying data quality is essential for data quality management. As mentioned in the previous subsection, the data quality level depends on the particular business case. It is not a one-dimensional example of data accuracy but has many characteristics and constraints, such as completeness, consistency, currency, and

timeliness. Therefore, one or more dimensions of data quality must be measured to define the data quality. This depends on the context, situation, and task for which the data will be used.

Table 2: Commonly cited data quality dimensions

Data Quality Dimension	Data Quality Dimension
Accuracy	Sufficiency
Reliability	Usability
Timeliness	Usefulness
Relevance	Clarity
Completeness	Comparability
Currency	Conciseness
Consistency	Freedom from bias
Flexibility	Informativeness
Precision	Level of details
Format	Quantitativeness
Interpretability	Scope
Content	Understandability
Efficiency	
Importance	

Table 2 summarizes several widely used dimensions of data quality. The quantification of data quality can be enabled by leveraging data quality dimensions. Additionally, the quality of data is closely related to various levels of organisational data, including data elements, data values, and data fields at the lowest level, data records and data sets at intermediate levels, and database tables at higher levels, with enterprise-level data warehouses occupying the highest level of hierarchy. This progression from lower to higher levels represents an increasing level of complexity and aggregation. As the levels of data hierarchy are increased, the difficulty of quantifying data quality also increases. Data quality is a measure of the condition of the database, including factors such as accuracy, completeness, consistency, and reliability. The importance of data quality is also increased with data processing related to business operations and the increasing use of data for decision-making. Inaccurate data are often identified as the root cause of business failures, incorrect analyses, and poorly designed business strategies.

The following **data quality factors** have been identified:

- **Data accuracy** refers to data consistency with reality and is considered the primary measure of data quality. There are two accuracy characteristics, namely form and content, where the form eliminates ambiguities about the content. For example, storing a date record (e.g., 1/5/2023) using a U.S. format to a database using European standards will result in a wrong

timestamp (i.e., 5/1/2023). In this case, the data content was correct, however, the data format led to inaccurate information.

- **Data completeness** is an elementary data quality dimension. In a database, "data are present" is synonymous with non-empty values in the table's data field, while "data are absent" means empty or zero values in the data field. Additional values, such as "unknown" or "not applicable", can be also used to represent missing data.
- **Data consistency** refers to data values being the same for all occurrences of an application and must be synchronised with all organisational entities. The layout and appearance of the data should be consistent across the complete range of data relating to the data entity.
- **Absence of duplication** of data records. The unique data quality dimension is the opposite of the data quality repeatability assessment.
- **Data currency** pertains to the time-based aspects of data quality and is defined as the pace at which data is updated to meet the current needs and demands of the business. This can be evaluated by determining the regular frequency at which specific data items are expected to be updated and assessing their timeliness. Another key aspect that impacts data currency is volatility, which refers to the frequency of changes to the data over time. The higher the volatility of the data, the lower its timeliness becomes. The significance of timeliness as a data quality dimension depends on the level of volatility present in the data.
- **Data conformity** refers to the compliance of the data format to the standards established by an organization. Validity or compliance means that the data aligns with internal or external standards, guidelines, or standardized data definitions, including metadata definitions. By comparing the data elements and metadata, it is possible to assess the level of conformity.

Having high-quality data offers numerous benefits. Firstly, it reduces the costs associated with identifying and correcting insufficient data, thereby avoiding errors that can increase operational expenses. Secondly, it leads to increased accuracy in analytics applications, which leads to better decision-making and an increase in sales. Lastly, it leads to more efficient time management, allowing data management teams to focus on more productive tasks.

The nature of data quality has expanded with cloud computing, artificial intelligence (AI) and machine learning big data systems and privacy on protection laws such as the European General Data Protection Regulation (GDPR). Data quality dimensions allow us to relate to the different perspectives from which data quality can be viewed, as depicted in Figure 4.

Several dimensions of data quality have been identified, which can be used within an organisation to evaluate the quality of its data in terms of compliance with specifications, suitability for usage, and the provision of appropriate data to the necessary business users at the right time. These dimensions of data quality assist in understanding the various perspectives from which data quality can be viewed. While these dimensions of schema quality do not directly affect the data itself, they do impact the structure, representation, and type of data and can have a significant impact on the quality of the data in the long term.

An organization can use several data quality dimensions to assess its data quality concerning compliance with specifications, suitability for the intended purpose, or timely delivery of accurate data

to the appropriate business users. The data quality dimensions contribute to understanding the diverse viewpoints from which data quality can be examined. Although the schema quality dimensions are not inherent qualities of the data, they affect the data's composition, representation, type, and overall sustained quality.

Figure 4: Data quality dimensions [37]

2.5 Measurement of Data Quality

Utilizing the dimensions mentioned earlier to assess data quality aids in determining the condition and degree of effectiveness of the data. It is important that identifying crucial data based on business requirements and impacts, and measuring data quality utilizing the appropriate dimensions are fundamental to achieving high data quality levels. As data quality is expansive, and organizations have numerous data elements stored in various systems, it is impractical to evaluate every aspect or dimension of data quality. However, the most pertinent data quality dimensions are contingent on the usage framework. It is essential to select the appropriate data elements and components to ensure maximum benefit. This decision should be based on economic benefits, priorities, impacts,

business needs, data criticality, benefits, and the types of projects in which the data elements will be employed.

It is recommended to take into account data profiles before embarking on data-related projects. All data quality dimensions are not equally vital for all objectives. Prioritization of some data quality dimensions over others is necessary. For instance, data security may not be significant for handling all data elements, but it becomes crucial for sensitive information such as customer financial data, social security numbers, or patient health information. As a result, when evaluating the quality of a dataset containing social security numbers, data security is evaluated. When measuring data quality, one should consider that each case is different and depending on the business plan as well as the budget, the appropriate metrics will be selected. For example, when measuring data quality, accuracy is an essential element, but other data quality dimensions, such as accuracy, integrity, uniqueness, coverage, and completeness can also be used as key performance indicators. Measurement outputs should be reviewed to establish that data quality issues need further investigation and correction or that the data reflect a valid business case.

Therefore, when conducting surveys and tracking participation and responses, it would be beneficial to include metrics on non-participation, or how we could manage responses with some metrics, we inevitably end up losing non-response to the survey and disengagement. This not only reduces the data set, thus reducing the impact of the study but also has the potential to introduce bias. Below we examine non-response and attrition rates [6].

2.5.1 Non-response

Non-response happens when a subject does not respond to a survey either partially (item non-response) or entirely (unit non-response). Unit non-response reduces the sample size and the power of the study [7], [8]. Significant differences between survey respondents and non-respondents can cause non-response bias, which is a type of attrition rate. This leads to the following types of problems:

- The reduction of sample size lowers the statistical power and reduces the efficiency of estimates, as well as the precision.
- Bias in estimates when non-response is selective (non-random)

The *non-response error* occurs due to the absence of a complete information collection for all units of the selected survey. Non-response error impacts the survey results in two ways. First, reducing the sample size or the quantity of collected information in responding to a particular question leads to larger standard errors. Secondly, and perhaps more importantly, a bias is introduced to the extent that the underlying distribution of several characteristics of non-respondents differs from those of respondents of a selected survey [9], [10]. There are several ways of dealing with non-response such as reporting information on non-response and assessing response rates and types of non-response. Furthermore, information on non-responders and evaluation of the response mechanism. Finally, the identification of a strategy to address non-response in the analysis.

Non-response is an important indicator, while high non-response values indicate a high probability of the data not being reliable. Table 3 shows the large possibilities that usually affect the results [11].

Table 3: Range of Possible True Percentages

	When Response Rate is:				
	90%	70%	50%	30%	10%
If 50% of the responders provided a particular answer; true value if everyone in sample responded could range from	45% - 55%	35% - 65%	25% - 75%	15% - 85%	5% - 95%

Although the response rate can be easily determined, its impact on the data is unknown. For instance, if there is a new round of data collection with relevant questions from the original survey, the results will be added to the original dataset. The data will be evaluated as a sample of non-responders receiving poor treatment. If half of the non-responders follow, then the responders in this data collection phase will be weighted by a factor of 2, which will be combined with the original data, meaning the results can be according to the equation below [11].

$$Adjusted\ Response\ Rate = \frac{Phase\ I\ Responses\ +\ 2\ \times\ Responses\ From\ Phase\ II}{Initial\ selectable\ sample}$$

In case data on non-responses are not available, the survey is likely to be biased against the contents. One approach would be to increase the level of worry for researchers to collect data on the non-response bias. Rather than increasing survey participation, it would be more efficient for part of the research effort to be about the no-response sample designed to assess key areas in which responders differ from non-respondents. Rather than expanding that, data from different sources can be used to how respondents answered and examine various parameters. When the response rate is high, there is little chance for error due to-nonresponse. On the other hand, when the response rate is low, the probability of significant error due to nonresponse is increased.

2.5.2 Attrition Rates

Attrition is mainly occasioned by non-communication or denial by sample members. On the one hand, reasons for dropping out due to non-contact are known to be quite different from those due to refusal. In surveys, the response rate otherwise referred to as the completion rate or return rate, is the number of respondents to the survey divided by the number of people in the sample. It is normally expressed in the form of a percentage. The primary measure is also used in direct marketing to refer to the number of people who responded to an offer [12].

Attrition bias occurs when the pattern of attrition in the sample is not random. The variables affecting the attrition may be associated with the outcome variable of interest, such as education, health, or household economic well-being. However, withdrawal bias will occur more formally should the error term in the interest equation be correlated with the error term in the selection or withdrawal equation. From this perspective, attrition bias depends on the particular model, as the correlation between the error terms depends on the exact specification of the model [12].

2.6 Data Quality Frameworks

This section presents a comparative analysis of various data quality frameworks, focusing on the definition, assessment, and enhancement methods that can be implemented in diverse business

environments. The authors of [13] have identified twelve data quality frameworks, as summarized in Table 4.

Table 4: Data Quality Frameworks

Year	Reference	Name
1998	[14]	Total Data Quality Management (TDQM)
1999	[15]	Total Information Quality Management (TIQM)
2001	[16]	Cost-effect of Low Data Quality (CLDQ)
2002	[17]	Methodology for Information Quality Assessment (MIQA)
2002	[18]	Data Quality Assessment (DQA)
2006	[19]	Comprehensive Methodology for Data Quality Management (CMDQM)
2006	[20]	Hybrid Information Quality Management (HIQM)
2009	[21]	Practical Data Quality Approach (PDQA)
2011	[22]	Data Quality Methodology for Heterogeneous Data (DQMHD)
2013	[23]	Data Quality Assessment Framework (DQAF)
2016	[24]	Task-Based Data Quality Method (TBDQ)
2017	[25]	Observe-Orient-Decide-Act Methodology for Data Quality (OODAM)

In [14], the author proposes a framework for managing data quality as a product, which entails a series of steps to ensure that the data meets the requirements of its intended users. The proposed approach involves identifying user-specific data quality requirements, establishing metrics for measuring the quality of the data, developing a plan for quality improvement, implementing controls to ensure data quality, monitoring and reporting on the quality of the data, and continuously improving its quality. Furthermore, the importance of aligning data quality with an organisation's goal is highlighted. Also, the limitations and challenges of data quality management methods are discussed.

The author of [15] highlights that having poor data quality can lead to increased costs, as well as reduced profits and competitiveness. To this end, it is important to adopt a preventative and proactive approach to data quality management, instead of applying corrective measures. In addition, the role of data governance in ensuring data quality is discussed, while practical recommendations for implementing data governance procedures are outlined.

The author of [16] presents a flexible approach for defining, measuring, and improving data quality by introducing a framework focused on understanding the value of data quality. Furthermore, the author outlines several data quality rules and approaches for consolidating enterprise knowledge.

Lee et al. [17] propose a methodology for the assessment and benchmarking of information quality. The methodology encompasses an information quality model, a questionnaire to measure it, as well as analysis techniques for interpreting the measurements. The analysis techniques are applied to analyse the gap between an organization and best practices. The analysis results are useful for determining the best area for information quality improvement activities. Finally, the proposed methodology is illustrated through its application to five major organisations.

In [18], the authors presented objective and subjective data quality assessment methods, which can assist in developing data quality metrics in practice. Based on these functional forms, illustrative metrics have been developed for important data quality dimensions. Finally, the authors proposed an approach for integrating objective and subjective methods and demonstrated its applicability in practice.

Batini et al. [19] introduced a thorough methodology for managing data quality that aims to incorporate and enhance the techniques and tools introduced in other frameworks. The proposed methodology is designed to be flexible, comprehensive, and easy to apply. Its flexibility allows users to select the most appropriate techniques and tools for each phase and context, while its comprehensiveness is achieved by considering existing techniques and tools and integrating them into a framework that can function in both intra- and inter-organizational contexts, and can be applied to all types of data. Additionally, the proposed methodology is straightforward since it is organized into phases, with each phase having specific tools, techniques and objectives.

The Hybrid Information Quality Management framework, presented in [20], aims to provide a structured approach to managing run-time error detection and correction. Specifically, it supports the identification of errors during the run-time phase of the process and implements corrective actions as required. Additionally, it introduces a new perspective on the traditional data quality management cycle, namely the user perspective. During the data quality definition phase, the main stakeholders of the business processes are identified, and the relevant dimensions are defined for each stakeholder class. The user perspective is established based on the importance and objectives associated with each quality dimension for each class of users, and this approach is used to examine the data quality challenges from an appropriate point of view.

In [21], Angeles and Garcia-Ugalde have developed a technique to inform users about the qualitative features of their data, the origin of the data, and how it is combined to ensure data trustworthiness based on quality. The proposed approach assigns quality scores to generated data by considering them as primary data sources, by comparing the available quality scores of their origins, or by aggregating the quality characteristics of all of their origins. Additionally, the quality of the data is enhanced by including a conflict resolution function and the code or formula employed for data integration, depending on the data granularity, along with a brief recommendation to users for trusting data based on the conflict resolution function employed.

The authors of [22] designed a quality assessment methodology for heterogeneous data that takes into account all types of data managed in an organization, including unstructured, semi-structured, and structured data. Moreover, they defined a meta-model for outlining the relevant knowledge managed in the proposed methodology. According to the methodology, the various data types are transformed into a common conceptual representation, while two data quality dimensions are analysed, namely the currency and accuracy dimensions.

The significance of a continuous improvement approach to data quality management and the crucial role of data governance in ensuring the ongoing quality of data are highlighted in [23]. Moreover, the author outlines a comprehensive methodology for measuring data quality, which involves defining the data quality criteria, selecting suitable metrics, and employing the appropriate quality analysis techniques.

In [24], the authors presented a new data quality assessment and improvement method for structured data, which is simple and practical so that it can easily be applied to real-world situations. The proposed method is able to identify potentially risky processes and suggest appropriate countermeasures. Towards achieving continuous improvement, an award system is integrated for assisting in the selection of the countermeasures. The proposed method is most appropriate for small and medium organizations.

Finally, in [25], the authors introduced the "observe–orient–decide–act" framework that aims to identify and enhance data quality through its continuous application. The proposed framework is adaptive and can be utilised across different application domains and organisation sizes. Although the framework does not involve any formal process for analysis and improvement, issues related to data quality are detected and visualised through routine reports and dashboards during the observe phase.

The data quality dimensions utilised in each of the aforementioned frameworks are summarised in Table 5. The CLDQ framework has the highest number of data quality dimensions (i.e., 35 dimensions), while DQA has 16 dimensions. Furthermore, the TDQM, MIQA, and TBDQ frameworks have 15 dimensions. An illustration of the occurrences of each data quality dimension in the framework is depicted in Figure 5. In more detail, twenty data quality dimensions are used in more than one framework. Also, completeness is used in most frameworks, followed by timeliness and accuracy.

Table 5: Data quality dimensions used in each framework [13]

Name	Data Quality Dimensions	Number of Dimensions
TDQM	Interpretability, amount of data, timeliness, objectivity, completeness, ease of understanding, access, security, accuracy, relevancy, concise representation, believability, reputation, value-added, consistent representation	15
TIQM	Accessibility, validity, precision, nonduplication, completeness, accuracy, definition conformance, derivation integrity, timeliness, usability, rightness, contextual clarity, equivalence of redundant or distributed data	13
CLDQ	• Data model: identifiability, naturalness, clarity of definition, simplicity, semantic and structural consistency, flexibility, obtainability, relevance, essentialness, attribute granularity, homogeneity,	35

	comprehensiveness, robustness, precision of domains • Data Values: consistency, timeliness, null values, completeness, accuracy, currency • Information Policy: security, unit cost, accessibility, redundancy, metadata, privacy • Presentation: representation of null values, consistent representation, format precision, appropriateness, use of storage, portability, correct interpretation, flexibility	
MIQA	Timeliness, understandability, believability, free-of-error, reputation, accessibility, objectivity, concise representation, consistent representation, ease of operation, interpretability, completeness, relevancy, appropriate amount, security	15
DQA	Accessibility, relevancy, understandability, completeness, consistent representation, appropriate amount of data, free-of-error, reputation, concise representation, interpretability, security, objectivity, believability, timeliness, value-added, ease of manipulation	16
CMDQM	• Structured: accuracy, currency, completeness • Unstructured: reliability, relevance, currency	5
HIQM	Accuracy, completeness, consistency, timeliness	4
PDQA	Accuracy, completeness, consistency, currency, timeliness, uniqueness, volatility	7
DQMHD	Accuracy, currency	2
DQAF	Completeness, timeliness, consistency, integrity, validity	5
TBDQ	Consistent representation, security, interpretability, completeness, relevancy, accuracy, ease of understanding, access, concise representation, believability, amount of data, objectivity, reputation, value-added, timeliness	15
OODAM	Speed, volume	2

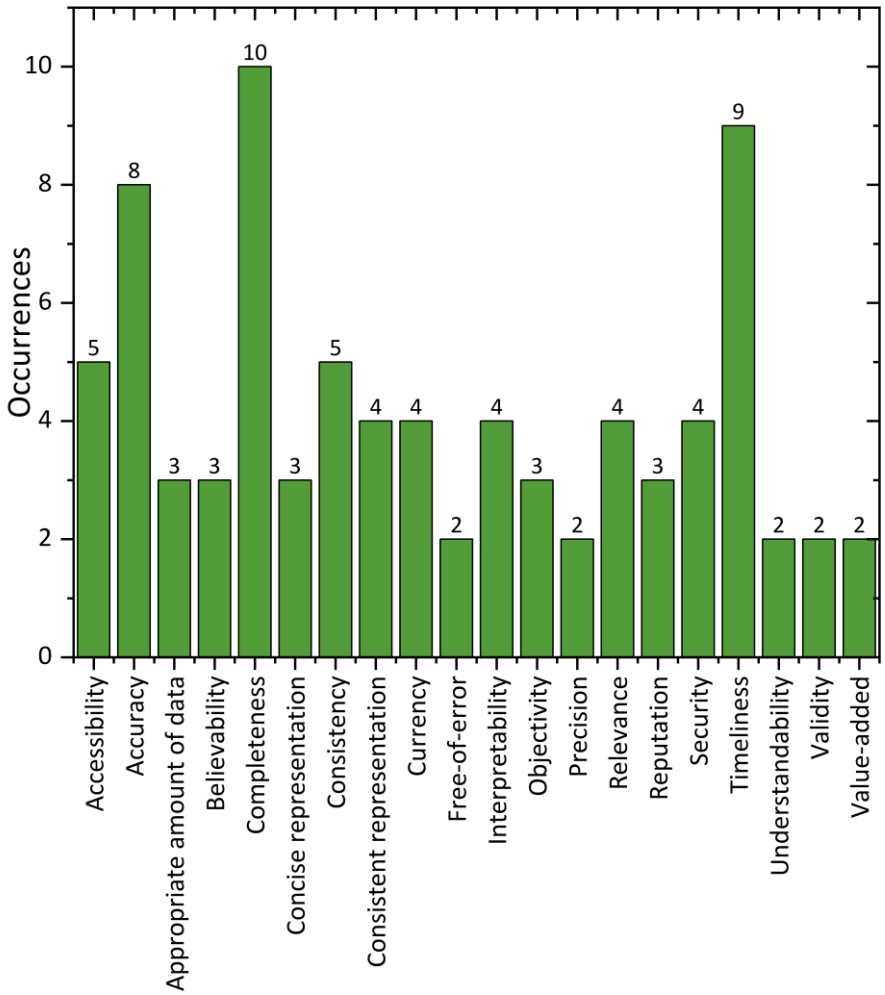

Figure 5: Data quality dimensions occurrences [13]

2.7 Data Quality Strategy

The data quality strategy involves defining the actions to achieve the goal and organizing the resources so that the implementation is conducted with the necessary operational resources. On the data side, we need to define the objectives, what data is essential and how the data will be used and what level of quality we need. Actions and processes should be determined to manage the data. A data quality strategy establishes the framework to solve problems about data quality and provides recommendations to minimize the distance from the threshold set. Establishing a data quality strategy will need a data maturity assessment to be conducted between different business structures in the organisation to specify how data is managed. The difference between data strategy and data quality strategy can only be semantic. Big data brings several challenges, and managing similar data brings several challenges. Different departments in an organisation maintain the business data stored to serve the business needs, which is why it is common to have duplicate records, inconsistencies, and data integration issues. The disorder and chaos that exist when we miss data strategy are not always visible. However, we can see it when we have dirty data, missing data, and timeless issues.

Similarly, without a data quality strategy, the decision-making process may be challenging. Even without a data quality strategy independently, the risk is for each department in an organisation to

develop its methods to manage and manage data as it is suitable for them. The absence of strategy gives the free pass to create their agenda and rules to perform their work.

2.7.1 The Capability Maturity Model

Data quality, as well as the maturity level of data quality, varies from organisation to organisation. Data quality maturity has been modelled after the capability maturity model (CMM) developed by Carnegie Mellon University [26]. However, the CMM needs to consider the maturity of an organisation concerning data management. The data quality maturity model defines five maturity levels starting from the initial chaotic level, where low-level practices and policies lead to continuous monitoring and improvement. Changes in people and process behaviours, tools, and technology characterize progress from low to high. Increasing data maturity and quality levels become part of the core practices, and we monitor them continuously. As we move from low to elevated levels, there is a significant reduction in risk and benefits from data quality management. Below we list the levels of data maturity.

Level 1: The Initial or Chaotic Level

The initial chaotic level is the lower level of the maturity data quality model. The absence of rules, policies and strategies for data management defines this level. There needs to be awareness of the issues resulting from the lack of data quality. The impact, practical solutions, and potential advantages of data quality management. This is where data is managed manually, even with spreadsheets. The same data can be in multiple databases and files in more than one location and with different formats and names.

Within this level, data quality is considered an added cost and not a return on investment (ROI). Daily data quality problems are encountered, changes are made at that moment, and issues often get rolled back and reworked; however, fixes are quick but repetitive. In organisations at this level, there are no roles and people responsible for the data, as there is no team to manage the data. The risk is high, and the benefits of data management are low at Level 1.

Level 2: The Repeatable Level

On Level 2, the organisation has partial knowledge of the impact of poor-quality data. It acknowledges the need for new processes to improve data quantity. Also, on this level, there is some basic management and information-sharing organisation from some rules, usually from good practices. In the beginning step, the data and format are analysed as the necessary fields are checked, for instance, fields that cannot have null values. Commonly the above is done in departments in an organisation and cannot be applied universally across the organisation. To move an organisation from Level 1 to Level 2 with respect to the data maturity state, the organisation must follow a data management policy and establish standards. The success of the Level 2 organisation will depend on the capabilities of the team responsible for data management.

Level 3: The Defined Level

Organisations at Level 3 clearly understand the value of good and high data quality; therefore, they try to be initiative-taking in their efforts towards maturity data quality. At this level, data management plays a major role in the organisation as the data is treated as an asset, just like the applications helping the organisation to make appropriate decisions. The level-to-level upgrade is characterized by documentation and establishing data management policy by the data quality team participating as core in the development team. The policies are tested with a test to ensure the data quality

requirements. Data quality is driven by business models where a business user knows how important the data is to help business functions in their decision-making. Over time data quality becomes part of the information technology (IT) department in the project. They have created a data management function to manage the data. Even the data quality team start to record good practices as it usually involves the whole organisation. From rules, data governance policies, processes, and services to data quality for applications and data quality validation.

Level 4: The Managed Level

On Level 4, data are considered assets and managed as strategic assets. Data quality is a major concern for the IT department, which applies metrics on performance and how well business needs are met. Tools with capabilities that take advantage of data management are implemented at the organisational level rather than the departmental level. Data quality management is proactive, with continuous measurement and data monitoring at all stages, which results in problems being identified early in the information life cycle by continuous measurement, monitoring of data quality, and assigning data quality to the generated bound at the organisation level. Continuous improvement is repeated in stages with processes that are continuously fed back and evaluated. An organisation can move to Level 4 when it leverages metadata, where the metadata will allow the data management group to maintain the enterprise data structure.

Level 5: The Optimized Level

It is the highest level an organisation can reach, where data and information are managed as assets, in the same way as finance, products and equipment. Data management becomes a business process rather than a technical tool at this stage. Data is enriched in real-time with additional data, such as market, geospatial, sociographic, and demographic data. Also, any unstructured information, such as policy documents, becomes subject to data quality control. Data quality maturity on this level, organisations use practices that have developed from lower maturity Levels 1 to Level 4 to improve data access, data quality and database performance continuously. According to Gartner, few organisations implement data quality, and typically organisations fall into Levels 1 and 2. Universally, few organisations are on Level 5. A Level 5 organisation embraces data quality and takes care of data quality processes such as metrics that assess the impact at the organisation's enterprise level.

Data quality aims to define business objectives and receive changes in business specifications. It should consider the people, processes, governance systems and technology needed to meet business objectives and priorities.

2.7.2 Data Quality Strategy: Preplanning

We will outline the process of creating the data quality strategy, namely where the data will come from, the team that will be formed, and a timeline. Figure 6 summarises the activities engaged in the preplanning.

Figure 6: Data quality strategy – Preplanning [37]

Timing should be taken into consideration before initiating the process of developing a data quality strategy. The duration of creating a data quality strategy can vary from two to six months, depending on the objectives being pursued and the availability of resources, key business partners, and relevant stakeholders. It is recommended to link the data quality strategy with the organization's planning and budgeting cycle. Therefore, planning should commence well in advance, approximately seven to eight months prior to the start of the budget planning session, to allow ample time for conducting enterprise-wide assessments, gap analysis, strategy formulation and review, stakeholder engagement, and obtaining executive approval.

As for the starting of the data quality strategy at the enterprise or business unit level, while defining the strategy at the enterprise level is considered good practice due to data crossing departmental and business unit boundaries, addressing a high-priority data issue in a specific business unit can lead to defining the strategy at the business unit level. The vision for the strategy and data quality must be established before the "design and deployment" phase, along with the explanation of its added value to the organization, and the determination of the specialists within the organization or external consultants to work on creating the strategy. The identification of key stakeholders and engagement with them before the initial planning and deployment phase is crucial. An agile, iterative approach, with stakeholder engagement throughout, is recommended over a waterfall approach, along with the setting of expectations, communication of inputs needed, addressing of concerns, eliciting of feedback, and incorporating of feedback.

In order to ensure the successful implementation of a data quality strategy, it is important to involve all relevant parties in the process through brainstorming sessions and individual meetings. This approach values the input of these parties, makes them feel involved and accountable for the changes and helps prevent any unexpected issues from arising. Forming a data quality strategy can take several months, so a flexible and iterative approach is recommended. This involves getting approval from stakeholders after each step, as each phase's results will inform the next. For example, proper assessments and recommendations should be done if the goal is not initially achieved. It is important to have meetings and activities throughout each phase to work towards the end goal, such as aligning the data quality strategy goals with the organization's corporate goals, assessing the current state, and prioritizing initiatives. Finally, a presentation of the outline should be prepared, reviewed, and approved by stakeholders to ensure a smooth process.

When constructing a data quality strategy for an organization, it is crucial to assess the present level of data maturity, remain aware of ongoing data quality efforts across various segments of the organization, and determine which areas are functioning effectively and where improvements are needed. Furthermore, it is important to contemplate the introduction of new data initiatives and how they will harmonize with existing data quality initiatives. A data quality strategy that is effective necessitates a thorough comprehension of the organization's data environment and the business value of its data, as well as an ability to extract the utmost benefit from the data for the business.

3. Data Documentation

This section refers to the structure research results should have and the additional features they may have to be effective. There should also be a mechanism for retrieval by opening the data so it can be easily identified. Consider that results get published on platforms such as Zenodo[1]. In this case, the platform's guidelines will help track them down. Understanding and making the data publicly available enhance transparency as it is easier to replicate and externally validate the impact results. In addition, the results will also depend on the guideline questions within the proposed thematic tests. Defining key evaluation questions will help us select and refine our evaluation's focus, enabling more efficient use of available resources. Having a concise list of questions can also be helpful when sharing the assessment with others.

In addition, understanding the data and opening it up improves accuracy because the results can be validated externally. Public access will also allow external researchers to analyse the same data, providing valuable information and learning for themselves and the original project. However, the privacy of research participants and the anonymity of all research subjects should be ensured in compliance with ethical and legal requirements. For example, data that could identify research participants (such as names, addresses, or location information) must be removed from publicly available datasets. This sensitive information can remain only in secure locations, accessible by authorized entities for specific purposes.

These digital data are stored in file formats, which often can be software. The software and file type usually depends on the primary purpose. For example, spreadsheet software if we want to represent data in a spreadsheet. This is because data tables have certain properties that the corresponding software supports. If such attributes are stored in a worksheet application, the users can expect the file format to preserve these properties or "important attributes". Conversely, if the table is created using a text editor, it is less likely that the software will support these properties. In the same way, a text editor will be more appropriate for formatting an article, for example, using a functional table of contents and adding page numbers. Formatting may, for example, depend on specific software. Software may become obsolete or support only certain versions of formats. It is also possible that certain formatting properties may exist only in the software used or even only in a particular version of that software. Files may also depend on the use of expensive or proprietary software. To exclude the risk of obsolescence and to ensure the accessibility and viability of important file properties, certain measures can be taken. One of these measures is the use of file formats with a high likelihood of remaining usable for many years. The Dutch National Expertise Centre and repository for research data published a guideline [27]. According to guidelines, the most appropriate file formats for long-term accessibility and sustainability should a) be frequently used, b) have open specifications, and c) be independent of specific software, developers or vendors.

The documentation process also includes data dictionaries, vocabularies, and readme files that can explain what the project data is and how the data can be collected, what the identifiers mean, and

[1] https://zenodo.org

how the data has been modified. In addition, to keep all documents associated with a project well organized and searchable, we should consider naming files, structuring directories, version control, and tagging files with specific keywords.

3.1 Unstructured Data

Typically, this type of data includes text, photos, audio and video files that have no defined structure and are difficult for an organization to manage. Unstructured data does not have a predefined model for which management can be done with a NoSql database. This data needs a hierarchical structure like a taxonomy. The taxonomy is the primary step to establishing a process for knowledge discovery. Figure 7 shows how with the assistance of the taxonomy, the unstructured data is transformed into a structure that can be transformed into knowledge.

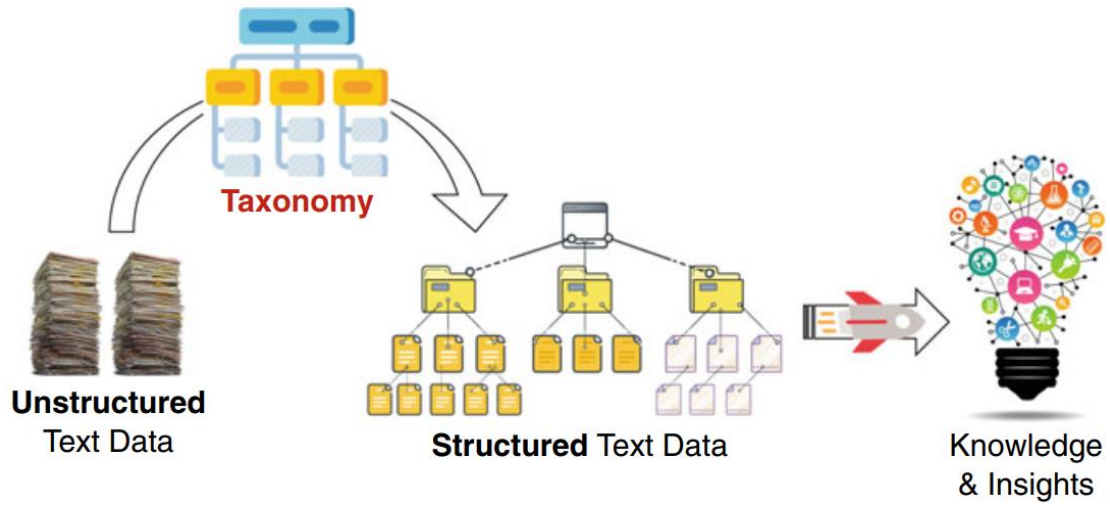

Figure 7: Converting unstructured text data into knowledge and using a taxonomy to structure raw text [28]

Taxonomies converge in a more abstract sense with the object they manage, whereas knowledge graphs contain more information. However, taxonomies contain additional information about the world they describe.

3.1.1 Creating taxonomy Data taxonomy

By organizing the data in a hierarchical format, it may be more efficient to start from an older taxonomy. We will examine the methodology, proposed by [29], the taxonomy should be enriched when our data changes. The creation of data taxonomies is related to the data we have available. First, we need to define a concept set and from this set, we extract terms and expand to a short list. Next, we construct the taxonomy. When we create the taxonomy we need to apply it to the data to validate its efficiency. Taxonomy development is a complex process, first, we need to define the characteristics of the domain of interest. Characteristic selection is a key problem for taxonomy development. These characteristics could be based on theory but might likewise be based on reality. The fundamental characteristics of the taxonomy are the meta-characteristics that we define at the beginning of the development. The meta-characteristics are more understandable than all the ones to be defined and are the foundation for selecting the later characteristics of the taxonomy. The purpose of the

taxonomy will be to be used so initially, we need to identify and define its intended users to be able to use it in the data. The selection of meta-characteristics must be done with careful consideration as it will strongly influence the results of the taxonomy. The method described is iterative and must have a termination condition. A Basic in the termination condition should fulfil the definition of the taxonomy and particularly by what it consists of and by its dimensions and whether it applies to all data. Table 3 shows the termination conditions as presented in [28].

Table 6: Termination Conditions

Termination condition	Comments
All objects or a representative sample of objects have been examined	If all objects have not been examined, then the additional objects need to be studied
No object was merged with a similar object or split into multiple objects in the last iteration	If objects were merged or split, then we need to examine the impact of these changes and determine if changes need to be made in the dimensions or characteristics
At least one object is classified under every characteristic of every dimension	If at least one object is not found under a characteristic, then the taxonomy has a 'null' characteristic. We must either identify an object with the characteristic or remove the characteristic from the taxonomy
No new dimensions or characteristics were added in the last iteration	If new dimensions were found, then more characteristics of the dimensions may be identified. If new characteristics were found, then more dimensions may be identified that include these characteristics
No dimensions or characteristics were merged or split in the last iteration	If dimensions or characteristics were merged or split, then we need to examine the impact of these changes and determine if other dimensions or characteristics need to be merged or split
Every dimension is unique and not repeated (i.e., there is no dimension duplication) Every characteristic is unique within its dimension (i.e., there is no characteristic duplication within a dimension)	If dimensions are not unique, then there is redundancy/duplication among dimensions that needs to be eliminated If characteristics within a dimension are not unique, then there is redundancy/duplication in characteristics that needs to be eliminated. (This condition follows from mutual exclusivity of characteristics.)
Each cell (combination of characteristics) is unique and is not repeated (i.e., there is no cell duplication)	If cells are not unique, then there is redundancy/duplication in cells that needs to be eliminated

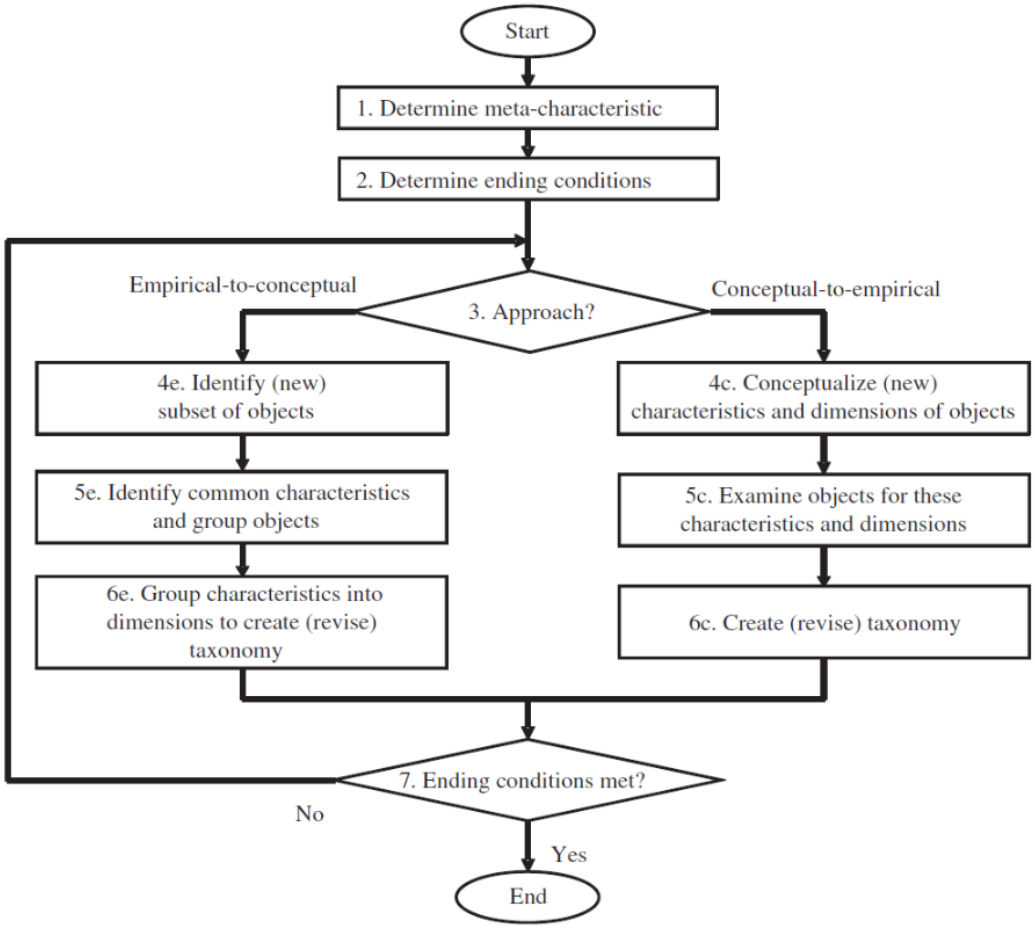

Figure 8: Taxonomy development methodology [30]

Figure 8 displays the steps used to create the taxonomy. We First start by identifying the meta-features, as mentioned above, taking into consideration the users and their expected utilization. Then specify the termination condition where we stop the process. The next step is whether to proceed with an empirical or a conceptual approach. A good practice would be that when we have data to choose from empirically, we could take an empirical approach example to run clustering algorithms on the data. Alternatively, if we know the domain, we can use both approaches. The empirical-to-conceptual approach identifies the conditional objects we want to include, these objects are usually known and identified by the literate review. The common features of the objects are then identified; as noted, the beginning is made with meta-features when we define the features they can be defined by the team and can be evaluated using statistical and graphical representations. The attribute group will create conceptual labels of the attributes and configure the dimensions. Each dimension will contain attributes and will be applied to the data. The conceptual to empirical approach starts without application to the objects but theoretically and empirically. It will then be applied to the objects.

3.2 Elements of Data Documentation

Data documentation includes various files that describe all the data used in a project. The more frequently used methods are described in the checklist below. Data documentation can be done using metadata templates, which are industry-specific forms that are widely used, or electronic workbooks, for example, which create metadata for the project while keeping the notes up to date. Documentation also includes data dictionaries, code books, glossaries, and retrieval files, which explain the project data, how it was collected, the meaning of abbreviations and how to modify the data. In addition, to keep all documents related to a project well organized and legible, attention should be paid to the naming of catalogues, the structure of the files, version control and the tagging of files with specific keywords.

Having proper documentation of your data makes it more easily understood by all parties involved. When variables, codes, and shortcuts are clearly defined, the overall quality of the data improves. Additionally, having readme files that explain the contents of folders can help reduce the chance of misinterpretation. Investing time in documenting your data during the project can save you time when it comes to publishing. This is especially important if the publication process takes a long time, as it can be easy to forget important details about the data collection and processing without proper documentation. High-quality documentation is also key for following open science principles, as it makes it easier to share your data with collaborators or fellow scientists after publication.

3.2.1 Data management software

It is mainly the software that allows us to create and manage databases, including metadata creation. Such software provides functionalities for a) making documentation easier, as it usually generates metadata by itself, b) offering methods to share and handle access, while having safe repository and inquiry tools, and c) providing methods for discovering errors (e.g., automatically detecting out of content inputs).

3.2.2 Data dictionaries

To avoid ambiguous terminology, we can use collections of descriptions and codes of procedures and calculations used in the project, and a dictionary of terms can be used. This could include a table with the researcher's name, the contents, descriptions, and the format of each dictionary. In broad terms, a data dictionary assists with the following:

- Creating the directory design as required.
- Dictionaries explain variables used in a dataset.
- Code books are packs of principles, algorithms and estimations utilised in a task.

3.2.3 Directory structure

A system directory will help with access control if there is sensitive personal data. When designing the hierarchical directory, we need to consider what kind of data there is and questions such as whether some of the folders will need access to certain projects or even how the folders are organized. It will take several clicks to locate what we need if we have a complex system, while fewer folders will result in searching among many files. Therefore, we must maintain a balance and avoid category overlaps. Of course, suppose we have sensitive data. In that case, we should be careful because the structured

metadata information may be visible to everyone. It is helpful to have a unique name on folders and files. Some considerations include:

- Making a folder structure to suit our task needs.
- Designing a transparent folder system that also controls access if we work with sensitive data.
- Striking a balance between external and deep folder hierarchy to save files is findable.

3.2.4 Tagging files

Tags are keywords that we assign to documents, which can be considered as labels are keywords set to files, making indexing, and searching files easier. A file can only be in one folder at a time, but it may have an infinite number of tags.

3.2.5 File naming conventions

Designing the nomenclature will be more effective if it begins at the launch of the project and the generation of a meaningful system outline. The same name should not be used twice to avoid loss of data. Therefore, a brief meaningful plan with unique names (in case of directory structure corruption) should be established.

3.2.6 Version control software

Dedicated software can be used for keeping track of changes to files, as well as maintaining a history of changes and which users made them. Also, version control software enables the return to an older version of a specific file.

3.2.7 Readme-files

Readme files are essential when sharing data as the recipients have a picture of the received files. Readme files provide details about data to ensure they are interpreted correctly. Usually, they contain details such as titles, authors/developers, definitions, dates, and file format descriptions.

3.2.8 Discovery metadata

The metadata allows the discoverability of the data that their generation is to understand the data. Metadata depends on where the data will be published and for how long. In general, metadata are included regardless of the nature of the shared data.

3.2.9 Research records

Research data may have a persistent identifier (PID). PIDs are permanent links that identify citable online sources, such as publications, datasets, and source code. The identifier remains the same even if the object's location on the internet changes. Widely used instances of PIDs include the Digital Object Identifier (DOI) and Uniform Resource Name (URN).

When publishing open data the following research records should be included:

- Data Management Plan (DMP)
- The license of use and reuse authorization
- Data handling agreements
- Methodology description
- Research plan
- Scientific publications from the data

3.3 Metadata Standards

In general, metadata can be considered as descriptors of data. They are labelled information that characterizes other data and is used to identify the data [31]. They constitute information that provides context about a particular piece of information, such as a file, document, or digital asset. Metadata can include a variety of details, such as the file size and format, author, date and time of creation, and any relevant keywords or tags that can describe the data context. Metadata is essential for organizing and managing large amounts of data, and it can be used for finding, accessing, and using information more efficiently. Furthermore, they play a crucial role in many processes, including digital archives, content management systems, and search engines [32]. An example of their use can be found in e-commerce applications, which assist customers by adding metadata to product categories to locate these products. Moreover, metadata can assist in product management. The metadata can be categorized into technical, business, and process metadata, as shown in Figure 9.

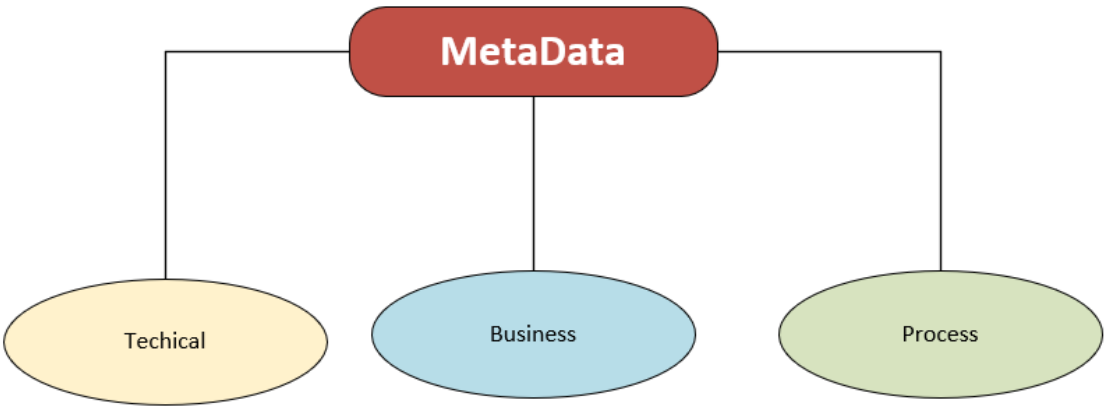

Figure 9: Metadata categories

Technical metadata describes various technologies that usually store data in repositories such as databases and files that will be used to access the data. An example could be a database or, similarly, a table's name and a data model's description.

Business metadata describes the functional non-technical view of the data and how the data are used by the organisation, which will add additional content giving value to the core data. Furthermore, business metadata can be in a different physical location of the data. Examples of such metadata could be business terms, business rules, privacy rules, security levers, and data quality rules.

Process metadata describes the results of various information system operations related to data creation and delivery. For example, extract-transform-load automatically processes data from

operations coming from scripts and keeps information such as the time of the last modification of the file and other information about files. Some organisations can collect data by transforming it into metadata and selling them. The collection and processing of these data concern companies that can use it to identify customers and their products, for example, which product they use and what service they can combine. There are also audit trail metadata, which are specific metadata that are not allowed to be modified but contain information such as the date the file was created, the owner of the file, and more. [6].

Metadata has been associated with activities that use the information to help improve the understanding and use of the underlying data. The metadata quality affects the data regardless of the category to which it belongs. The absence of common data standards between departments within an organisation and the lack of metadata leads to data quality issues [6]. Metadata is valuable to an organisation because it allows us to figure out the data and how to use it. Without understanding the data, they cannot translate it into information and thus, the data is useless. This is known as the problem of terminology. An example is in business vocabularies, where a frequent practice is that a department in an organisation may use the term 'customer' when there is a possibility that another department mentions 'client' when it refers to the same entity [6]. Specifically, when there is no data input template, we might encounter difficulties with our data and need additional processing to structure it. Another relevant example is the electronic health record of blood pressure. Some hospitals entered it numerically and others as text, creating difficulties. Usually, metadata is also associated with data quality because it supports the use of the data.

Similarly, structured information describes tracks, or facilitates access, use or management of an information resource. Furthermore, in structured information, the quality of metadata is the level of how informative the metadata is, including a clear statement of terms and conditions of use. Likewise, they should be based on content standards and perform the essential bibliographic functions of discovery, usage, provenance, currency, authenticity, and management. The National Information Standards Organization Framework Working Group has published a framework outlining six main metadata principles as follows [33]:

1. FAIR metadata serves community standards in a way suitable to the materials in the collection, users of the collection and current and potential future uses of the collection.
2. Having an adequate volume of metadata assists in achieving high interoperability.
3. Useful metadata utilises authority control and content standards to describe and collocate related objects.
4. Right metadata includes a clear view of the digital object's requirements and terms of use.
5. FAIR metadata supports the long-term curation and protection of objects in collections.
6. Metadata documents are objects themselves; thus, they should have the rates of eligible entities, such as unique identification, authenticity, management, achievability, and persistence.

3.4 Vocabularies

There is usually a direct correlation between the cost of creating metadata and the advantage to the user; it costs more to describe each item than to describe collections or groups of elements; also, it is more expensive to use a rich, complex metadata schema than to use a simple metadata schema; finally, it is more costly to apply formalised subject vocabularies and classification systems than to assign a few unchecked keywords. Development costs often lead to greater efficiency and

effectiveness for the business and the end user. Using a standard topic thesaurus or another controlled vocabulary, for example, can deliver greater precision and recall in search and may enable future features such as topic navigation and dynamic topic search. Any decision on which metadata standards to adopt and which levels of description to apply should be made in the context of the purpose of the organisation in creating the collection, the human and technical resources available, the users and intended use, and the approaches adopted in the research field or knowledge domain.

Organisations should keep in mind that multiple metadata systems may be necessary to meet their needs, depending on the type of collections they hold. Hence, it is advisable to consider using a combination of metadata systems. For instance, using Encoded Archival Description (EAD)[2] as a collection-level system for archival groupings with shared origins may be appropriate. Additionally, it is important to choose a careful combination of both published and collection-specific controlled vocabularies, which can be utilized as data values to fill in important access elements in the selected schemas.

[2] An XML standard for encoding archival finding aids, maintained by the Technical Subcommittee for Encoded Archival Standards https://www.loc.gov/ead/

4. Discoverability of Research Data and Data Anonymization

This section describes the methods that should be applied to research data to be effective. It explores how data can be made available for everyone and the positive benefits it can produce.

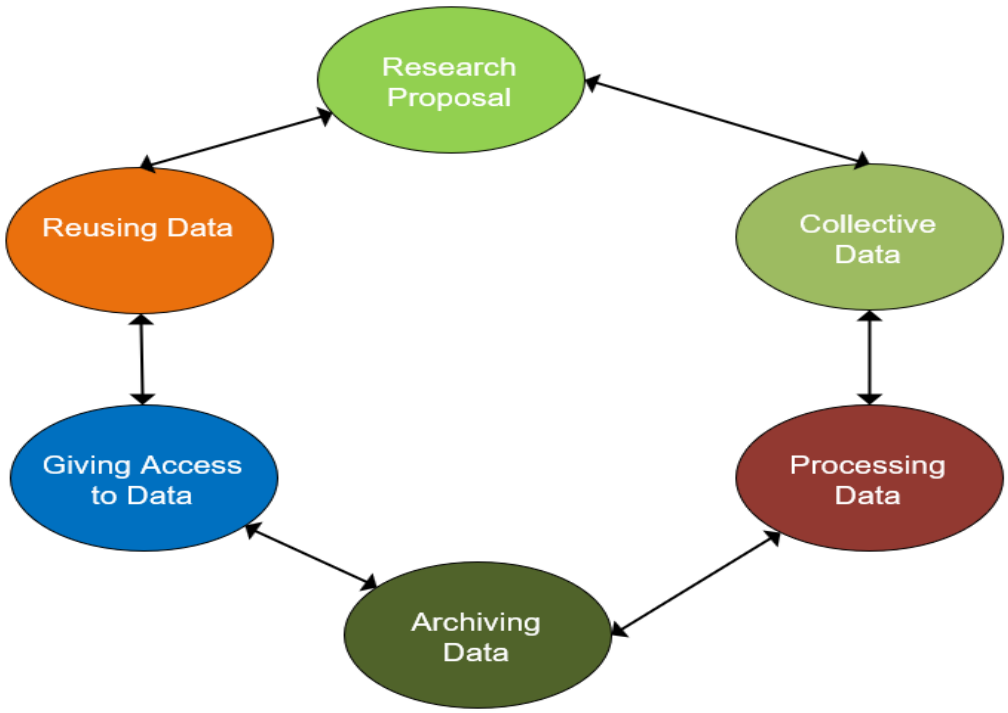

Figure 10: Research data lifecycle

Figure 10 illustrates the research cycle and procedures. Research Data Management (RDM) includes all methods and rules for managing, holding, and exchanging materials at every stage of the research workflow [34]. Good RDM guides scientists through processes that will improve, simplify, and enrich their research [35]. Researchers have a variety of innovative technologies and services at their service that can change how they manage their materials at every stage of their research. These range from software to collect and analyse materials, online data-sharing portals to publish and gain credit for data, cloud storage for institutions, and powerful but under-utilised features of familiar software [36]. RDM practices, training, and tools such as the DMP are applied at this stage. They help researchers better understand their materials' technical, legal and ethical aspects. For example, a reviewer notices that a result is not supported by the data or the sources used. In that case, the steps can be retraced and replicate the results over again. Bad RDM practices can lead to both losses of data, loss of research potential, loss of collaboration and funding opportunities and great disappointment.

Anonymity is another critical aspect to be considered when sharing data. Any information that could identify survey respondents (such as names, addresses, or location information) will be removed from the publicly available datasets [37]. This sensitive information will be kept secure and available only for authorized future data collection activities. How data sets are used to evaluate policy interventions becomes critical for data analysts to discuss how this may affect the interpretation of results. In research data, it will be valuable to provide metadata. Furthermore, the data can have a subject,

timespan, creation date, author, type, and size. Likewise, data will be provided in machine-readable formats, like *json*, *xml*, *xls*, and *csv*.

4.1 Research Data Management Policy

Adopting good RDM practices is a requirement for any project that produces or reuses research data. Also, it is a key part of the EU's open science requirements. The steps that should be followed are summarised as follows:

1. Prepare a data management plan and update it throughout the project.
2. Submit and maintain the data in a trusted repository and provide open access to it "as open as possible, as closed as possible".
3. Provide information (via the same storage) on any research outcomes or other tools and resources needed to reuse or validate the data.

In addition, beneficiaries must responsibly manage digital research data generated under the action ('data') in accordance with the FAIR. In D2.3 'Serious game implementation design', we presented the FAIR principles and the usage of Zenodo.

Data are valuable research products and assets that contribute to the knowledge economy. A high level of research data management is fundamental to both high-quality research and academic integrity. Therefore, the EVIDENT consortium has agreed on a research data management policy that clarifies the recommendations for proper data management. All those involved in scientific research are expected to implement these recommendations to the extent possible. The research community, in turn, continues to invest in a wide range of support to implement these recommendations [36]. Some of the guidelines used by the University of Leuven for research data management are listed below [38]:

1. Research data must be safely and sustainably saved and documented to ensure that the data can be accessed and retrieved when needed.
2. The metadata of research data must be logged to ensure that the data can be retrieved.
3. Deleting data must be explained and documented; documentation relating to the data should be kept, as this would impede the audit trial.
4. After the end of the survey, the related research data should be kept for at least ten years in a safe, secure, and durable manner for reproducibility, verification, and possible reuse.
5. All research shall be conducted following and considering existing contractual agreements, legislation, regulations, guidelines, or exploitation possibilities.

It would be further interesting to note the Lewis-Corral model that focuses on different types of functional activities, such as policymaking and training, with an underlying hierarchical concept. The proposed model maps the potential roles of the library in a 10-stage research life cycle model (at various points identifying possible partner services) [39]:

1. *RDM requirements collation - via testing (with academic departments)*
2. *RDM planning - researcher mentoring and advocacy at all levels (with doctoral training centres)*
3. *RDM computing - technical advice on data formats and metadata*

4. *Research data reporting*
5. *RDM training - training to researchers (with PhD training centres)*
6. *Licensing of research data*
7. *Research data assessment - guidance on which data should be retained.*
8. *Storage of research data (with IT services)*
9. *Access to research data*
10. *Impact of research data (with research support offices)*

The aforementioned model provides a framework for the management of research data. We start by collecting and validating the results. Then we design the research based on our results. With the support of information systems, we can build the infrastructure to host our data and further enrich the data with metadata. Next, we create documents from the research data. Similarly, we need to define the training for the researchers and establish the license for our data. Then, we need to determine which data we will keep in our collection to be stored in the infrastructure we have created and easily accessible to those who need it. Finally, it would be valuable to have the impact that this data will have on society. Figure 9 illustrates an example of ML model optimisation during an experiment. Knowledge is then generated, and the results of the experiments can be shared.

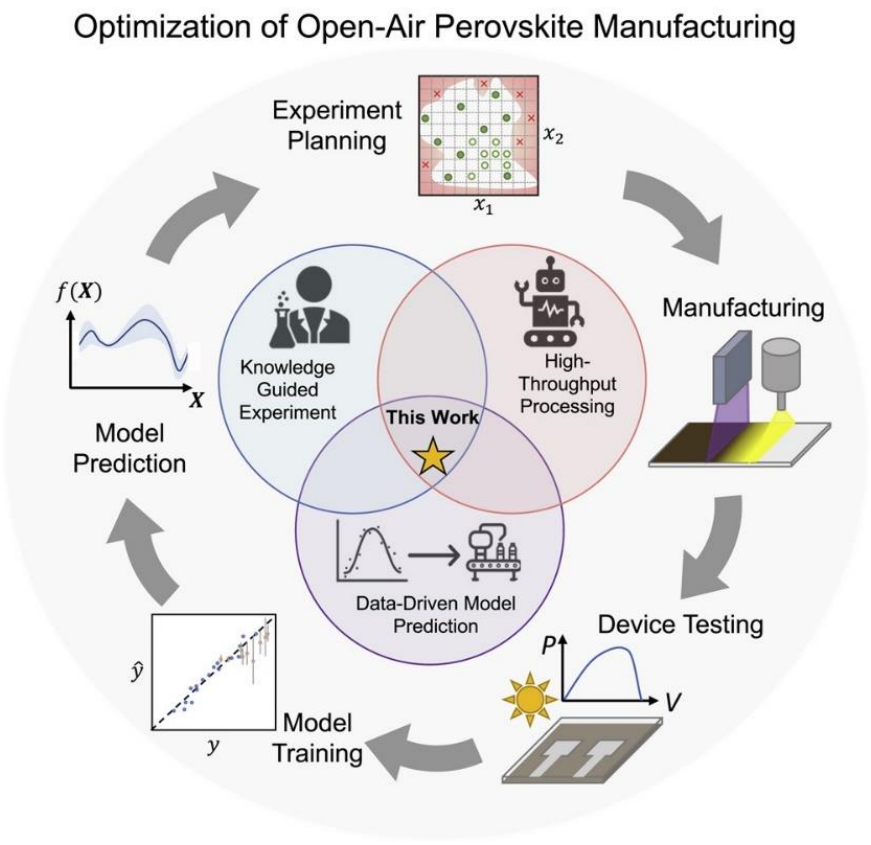

Figure 11: Example of ML model optimisation

The above is presented to understand the importance of data documentation in RDM. Another example is the European Open Science Cloud (EOSC), which has been recognised by the Council of the European Union (EU) as the pilot action to strengthen the new European Research Area (ERA). It is also identified as the data space for science, research, and innovation, which will be fully integrated with the other sectoral data spaces defined in the European Data Strategy [40].

4.2 Open science in Europe and discoverability

Open science involves sharing knowledge, data and information between researchers and tools as early as possible in the research and innovation (R&I) process openly and transparently, working with all relevant knowledge stakeholders, including academia, industry, public authorities, end-users, citizens, and the public-society at large. Open science can increase science and technology's quality, efficiency, and effectiveness. Furthermore, the impact of research and development (R&D) leads to a more remarkable ability to respond to the needs of the public and the wider public. Societal challenges and enhance society's confidence in the scientific system [41]. Some open science examples can be summarised as follows:

- Providing unrestricted access to the publication
- Adhering to the FAIR principles when sharing data
- Offering unrestricted access to data
- Providing information about outputs/tools/datasets for increasing reproducibility
- Offering open access to the results for facilitating reuse

4.2.1 Open access to scientific publications

The concept of open access to scientific research publications means that all users have free electronic access to these materials. Although there are no strict legal definitions of what is considered "access," it generally encompasses the basic rights to read, download, and print, as well as additional rights such as copying, distributing, searching, browsing, linking, tracing, and mining [42]. There are two main options with respect to open access.

- **Self-archiving/green open access:** The published article or the final peer-reviewed manuscript is archived (deposited) by the author, or a person representing the author, in an electronic repository before, at the same time as or after publication. Some publishers request that open access be granted only after a period of embargo.
- **Open access publishing / 'gold' open access:** an editorial is immediately published in an open-access format. In this model, payment of the publishing costs is shifted from subscribing readers.

On other occasions, the cost of open-access publications is paid by public grants or other funding models. In the context of research funding, open access requirements do not imply an explicit requirement to publish the results. The choice of publication is completely up to the awardees of the grant. Open access becomes an option only if publication is chosen for dissemination. Furthermore, open access does not affect the choice to commercialize the research results, for example, through patenting. The choice of whether to publish via open access should be made after the more general determination of whether to publish directly or to seek protection first.

4.2.2 OpenAIRE

The Open Access Infrastructure for Research in Europe (OpenAIRE) is the recommended entrance for researchers to specify which repository to select. It also provides support functions for researchers, including the National Open Access Centre Desks (NOACD). Other useful repository directories include the Registry of Open Access Repositories (ROAR) and the Directory of Open Access Repositories (Open-DOAR).

OpenAIRE aims to enable, encourage, assist, and support the scientific communication of open science in Europe. The infrastructure has been in operation for several years. It has effectively connected people, ideas, and tools to support the free provision, sharing, access, exchange, and exploitation of research results. In this direction, it offers a range of tools for education and communication on open access, enables knowledge exchange and provides the technical services required to facilitate and track open science publishing and the impact of research across geographical and scientific boundaries. OpenAIRE serves to complete a research graph whose objects are scientific outputs, institutions, funders, communities, bodies, and data sources [43]. The OpenAIRE Guidelines for Content Providers outline various research items, specifying main metadata fields, vocabularies, and protocols and presenting the best approaches from libraries, research data, and software while covering recent trends in research infrastructures.

For a publication to be considered machine-readable, it must be stored in a text format that can be easily processed by a computer. These publications should either be in a standard form or readily accessible to the public so that new processing tools can be developed. Ideally, the version submitted should match the published version. Scientific publication repositories can be electronic, and options for these repositories include foundation, thematic, and main repositories.[42]. OpenAIRE provides five guidelines for the management of research data depending on the type of data, as follows:

- The first category includes guidelines on how to manage publications. The purpose is to assist repository managers in making the metadata of all scientific publications and references to research projects accessible.
- The second category offers direction for data archive managers utilizing the description of research data. The standard data citation metadata schema is employed, with appropriate extensions of metadata properties and verified vocabularies provided. This also includes instructions on how to comprehend this metadata.
- The third category is aimed at software repository managers. They provide direct visibility to software as a 'reportable research product', are defined pragmatically, keeping mandatory properties to a minimum, focus on properties for reference (performance and access) and have the possibility for future property additions. In contrast, discovery properties for reuse are not considered.
- The fourth category is for other research products (ORP), which describes research products that differ from literature, data, and software, such as research services, protocols, and workflows.
- The fifth category concerns the Current Research Information System (CRIS) managers designed to capture integrated research information using the Common European Research Information Format (CERIF) standard. By implementing this, the movement and utilization of metadata within their systems as part of the OpenAIRE framework are facilitated. There are also suggestions for CRIS platform developers, including offering support functions for CERIF

managers and users. It is worth mentioning that the exchange of information between individual CRIS systems and the OpenAIRE infrastructure constitutes a point-to-point data exchange.

4.2.3 Zenodo

Zenodo is a repository designed and developed in the context of the European OpenAIRE project, which is managed by the European Organisation for Nuclear Research (CERN). It provides researchers with the ability to deposit various forms of digital research materials, including research papers, datasets, software, reports, and other research-related objects. A permanent DOI is issued for each submission, making the stored objects easily citable. Zenodo was established as a successor to the OpenAIRE to allow researchers in any subject area to comply with any open science deposit requirement absent from an institutional repository. It provides DOI on datasets and other submitted data that do not have DOI to facilitate work reporting and supports various data types and licenses. Zenodo is supported by CERN and operates on its high-performance computing infrastructure, which is primarily utilized for high-energy physics. It is built using Invenio, a free software framework designed for large-scale digital repositories.

4.2.4 Personal data

Personal data is related to the identification of a natural person who can be identified by reference to an identifier or even to one or more factors such as name, social number, location data, as well as physical, psychological, or genetic characteristics that can identify a person [44], [45]. The above definition leaves no space for incorrect interpretations of personal data. Any information that can be used on its basis or in combination with other information to identify and recognize a data subject. For example, in Europe, any information on any device ID that is unique and can determine the location and, in combination with subjects, is covered by the law on personal data. The GDPR law provides the framework to protect personal information and allow citizens to have control over their data. The data generated daily is massive, which hides knowledge waiting to be discovered. Healthcare organisations are a good example where we could exploit this data for better treatment. This data, of course, cannot be shared and studied without anonymization.

Knowledge discovery is the key to innovation; researchers discover data and extract new knowledge. Nowadays, vast knowledge is hidden in electronic traces of human activities. However, on the positive side, many worries have been raised about people's privacy, as the combination of available information can retrieve sensitive features. One way would be to delete the fields that have attributes. Nevertheless, an individual can be uniquely identified even if these attributes are removed. Some attributes that can identify an individual include credit card number, phone number, and social security number. Any information identifying one person from another can be used to re-identify anonymous data.

4.2.5 Attributes

Consider a Dataset T consisting of t records where each has several entities. The attributes are classified into four basic categories as follows [46]:

- **Direct identifiers** include attributes that can uniquely and directly identify an individual. Such identifiers include the name, social security number, phone number, and e-mail. In general, direct identifiers are removed before the application of the anonymization methods.
- **Quasi-identifiers** are attributes that can lead to the identification of an individual when combined with auxiliary information or other quasi-identifiers. Examples of quasi-identifiers are gender, race, area code, and age. These identifiers are usually suppressed or generalized during the anonymization process.
- **Sensitive attributes** refer to information which an individual usually wants to remain hidden. Examples of such attributes are diseases, salary, and political or religious views. Based on the purpose of the analysis these attributes may be retained.
- **Non-Sensitive attributes** consist of any attributes that are not included in the aforementioned categories. Weight, height, and hair colour are examples of such attributes.

4.3 Data Anonymisation

Anonymization is transforming a dataset *T* into its anonymised equivalent *T** so that privacy concerns are addressed. The original dataset may pass through several transformations in order to achieve anonymization and generate the anonymised one. In this direction, several techniques can be applied. For further details, we can refer to [47][48][49].

4.3.1 Suppression

While anonymizing datasets, suppression methods are frequently utilised for the purpose of removing or concealing specific identifying information [50]. These approaches have the potential to be efficient in protecting the privacy of individuals while at the same time enabling researchers and analysts to derive actionable insights from the underlying data. The practice of removing direct identifiers from a dataset, such as names and addresses, is one of the most popular suppression techniques. This method may be efficient in removing the most sensitive information from the data, but it may not be sufficient to prevent re-identification through other means, such as cross-referencing with other data sources. Nevertheless, it may be sufficient to remove the most sensitive information from the data.

Suppression methods can be categorized into record suppression, value suppression, and cell suppression. In more detail, record suppression involves removing entire records from the dataset that contain sensitive information that could potentially identify an individual. For example, if a dataset contains information about medical procedures that could identify an individual, record suppression would involve removing the entire row of data that corresponds to that individual's record. This method has the potential to successfully protect users' privacy, but it also carries the risk of losing crucial data in the process.

Value suppression involves replacing specific sensitive data values with more general or less sensitive values. For instance, the age value of a person can be omitted entirely or generalised to a range of ages rather than being included in the record in its exact form. This method is useful in cases where

particular data values could reveal private information about specific individuals, while at the same time enabling the preservation of other important information within the dataset.

Cell suppression involves masking individual data values within a record that could potentially identify an individual. For example, if a dataset contains information about medical procedures that could be used to identify an individual, cell suppression would involve masking or removing the specific medical procedures that correspond to that individual. This method is useful when only specific data values within a record are sensitive and should be suppressed, while other values should be retained.

4.3.2 Bucketization

Bucketization is a data anonymization approach that groups individual values into ranges, or "buckets". This technique is frequently employed in data science and statistics to safeguard the privacy of individuals in sensitive data sets. Bucketization can be used to conceal sensitive information, such as age or income, by grouping them into a range that represents a wider population [51], [52]. For example, instead of expressing an individual's actual age, it may be reported as a range (e.g., "15-20", "21-30", "31-50", etc.).

A benefit of bucketization is that it can assist preserve the usefulness of the data set while protecting individual privacy. In comparison to other anonymization techniques, such as random perturbation, bucketization can lower the amount of noise contributed to the data set. In addition, bucketization can be used with other anonymization methods, such as suppression and generalisation, to build a more comprehensive strategy for protecting privacy. A before-and-after example of the applying bucketization method is shown in Table 7 and Table 8.

Table 7: Bucketization example: Original data

Age, Gender	Gender	zipCode	GroupId
40	Male	57003	1
25	Female	56321	1
50	Female	10312	2
30	Male	42152	1
32	Female	57001	2

Table 8: Bucketization example: grouped data

GroupID	Gender	Count
1	Male	2
1	Female	1
2	Female	2

4.3.3 Permutation

The permutation is a typical anonymization method that includes shuffling or rearranging the identifying information in a data set to safeguard individuals' privacy [53]. With this method, the

original values are replaced with new values chosen at random from the same collection of values, but in a different sequence.

The permutation is an effective method for anonymizing sensitive data since it prevents the original data from being reconstructed from the anonymized data. Nevertheless, it is essential that the permutation technique is well-designed and that the anonymized data set is accurate and consistent with the original data set [54]. In addition, it is crucial to evaluate and assess the trade-off between data privacy and data utility, since loss of useful information can occur.

4.3.4 Perturbation

The perturbation method adds random noise or changes to the original datasets in order to ensure the individual's privacy. This method is usually applied in cases where the original datasets need to be preserved for analysis or research purposes, but where the disclosure of sensitive information must be minimized [55], [56]. The perturbation method can be further categorized as follows:

- **Noise addition** in the numeric records, we add to the existing values a random range of values.
- **Data swapping** is used for numeric and categorical values where values are exchanged between records.
- **Synthetic data generation** is based on an initial data set where a mathematical model is created that reproduces anonymized data. Because the reproduction of the data is random many argue that this data is not useful. A solution to the above is partial synthetic data and Hybrid data [57], [58].
- **Micro-aggregation** is a process where aggregated values of attributes are generated to have low risk in tracking. Its implementation is carried out in two phases the first is data partitioning and the second is partition aggregation [59]. The publisher creates several data sets derived from the original ones. Then different data sets are created from the originals and each partition represents the aggregated values.

4.3.5 *k*-anonymity

The k-anonymity algorithm is a widely used anonymization method and is based on the concept of creating groups or clusters of records with similar attributes [49]. According to the algorithm, for every row of records, the total number of records in the equivalence class to which it belongs should not be less than *k*. This indicates that there are at least *(k-1)* records with the same value in the semi-identity column. As a result, for any record in the k-anonymized dataset, the probability of associating a record with an individual is *1/k*.

This technique has the drawback of being susceptible to link attacks and unable of preventing attribute leaks. Also, the individual information in a k-anonymized dataset is vulnerable to two attacks method, namely homogeneous attributes and background knowledge.

4.3.6 *L*-diversity

L-diversity is a data anonymization method that aims to protect privacy in datasets by decreasing the granularity of data representation [60]. An equivalence class satisfies the *L*-diversity algorithm if the set of sensitive data corresponding to all records in that class has *L* acceptable values. If all equivalence classes in the dataset fulfil the *L*-diversity algorithm, then the dataset itself satisfies the algorithm.

A dataset that conforms to the *L*-diversity algorithm has significantly reduced data leakage risks compared to its *k*-anonymity equivalent [61]. Even though the *L*-diversity model takes into account

the diversification of sensitive attribute values in the equivalence group, it is susceptible to similarity attacks and skewed attacks. Consequently, an adversary can still infer the value range or sensitivity of an individual's sensitive information.

4.3.7 *t*-closeness

The *t*-closeness method was developed to further increase l-diversity by preserving the distribution of sensitive areas [62], [63]. t-closeness restricts the amount of individual-specific information an observer may acquire by requiring that the distribution of a sensitive attribute in any equivalence class is similar to the attribute's distribution in the entire table. *t*-closeness, unlike *k*-anonymity, eliminates attribute disclosure. *t*-closeness demands that the distribution of a sensitive characteristic in any comparable class closely resembles its distribution in the overall table.

To achieve *t*-closeness to an equivalent class, the gap between the distribution of a sensitive attribute and the distribution of the same attribute over the whole table cannot exceed a threshold *t*. In addition, for a dataset to be *t*-close, all equivalence classes must also be *t*-close. In addition, it is required that the distribution of a sensitive characteristic in any equivalence class closely resembles its distribution in the overall table. The primary downside of *t*-closeness is that it does not prevent identity revealing.

4.3.8 Differential privacy

Differential privacy aims to protect sensitive data while enabling the extraction of useful information from it [64], [65]. Differential privacy ensures that whether or not an individual's information is included in the dataset, it has almost no impact on the released dataset. To achieve this, it adds noise to the data before releasing it, making it difficult to determine if a particular individual's data is included in the dataset. The level of added noise should be carefully determined in order to balance privacy protection and the usefulness of the released dataset.

4.4 Information Loss Metrics

The loss of information that occurred during the anonymisation can have a considerable impact on the whole process. The biggest concern of the publisher is to find the balance between privacy and the maximum information that can be shared. To this end, various information loss metrics have been proposed, that can be categorized into genera-purpose, specific-purpose, and trade-off-purpose.

4.4.1 General-purpose

The resulting data set should be as close as possible to the original one. The minimal distortion (MD) metric measures how many times the identifiers were generalized [47]. Another general metric is iloss, according to which we assign to and will attribute with 1 and 0.

4.4.2 Specific-purpose

During the anonymization, we can see the loss of information we have. Specifically, generalization and also with suppression can affect the data mining process. As an example, we mention the Classification Metrics (CM) presented in [66] to measure the error resulting from suppressed or generalized records each error is charged with one unit.

4.4.3 Trade-off-purpose

These metrics try to reflect the balance between privacy and information loss in each process for instance [66]. For a dataset, denoted by *s*, the IGPL(s) metric is defined as follows:

$$IGPL(s) = \frac{IG(s)}{PL(s) + 1}$$

Where IG(s) is the information gain, while PL(s) denotes the privacy loss.

4.4.4 KL-Divergence

The Kullback-Leibler Divergence, also known as KL-Divergence, is a useful metric when the identifiers are correlated and their replacement will also affect the other identifiers, entropy maximization could be used which can assume a uniform distribution of more information [67]. The original data set T and the anonymity T* is considered as probability distributions

$$KL = \sum p_1(t) log \frac{p_1(t)}{p_2(t)}$$

This metric quantifies the distance between the original dataset and the probability distribution reconstructed from the anonymized data.

4.5 Data Sharing

Data sharing is a simple task, but sometimes quite complex issues may arise that must be carefully considered. Personal data, as defined by the GDPR, means any data linked to other data sources (online or otherwise) could enable the identification of individuals. Anonymization should be studied from the beginning and considered an "integral part" of the project. The anonymization process should be completed along the way and not left for the end. Like all materials on the internet, shared research data is covered by a copyright license, which details what others can do with the data. This license affects how we can use and share the data as part of our work. Data is always covered by a copyright license, even when the license information is not included with the data. We can still use the data in our research if the data does not include a license. However, we can only copy, share, or publish this data once we contact the owner and obtain permission. When reusing data created by others, ensure we understand the copyright license that accompanies the data.

4.5.1 Entities

A typical scenario of file sharing and anonymization includes the following entities: a) the data owner/editor, b) the record holders, c) the data recipient, and d) the adversary. We will briefly examine each of these.

- Data Holder/Publisher: It refers to the person who publishes the data set, taking into account security and privacy considerations. The data holder and publisher may be a single person. However, in case the data holder cannot ensure the privacy of sensitive information, due to missing expertise or resources, the data holder and data publisher are distinguished.
- Record Owners: Anyone participating in the available shared dataset containing single or multiple records.
- Data Recipient: It refers to anyone that has legitimate access to the published data.

- Adversary: A hostile attacker is an intrusive recipient who wishes to gain additional knowledge about the participants in the dataset.

4.5.2 Publication release

Various data recipients may have different guidelines on how to publish their data, there are three different scenarios, namely a) publishing a single release, b) publishing in parallel, and c) publishing in sequence.

Publishing in a single release: The primary and most typical case assumes that the data publisher, based on the privacy assurances it wants, releases the anonymised dataset only once. The initial table T, or any subset thereof, has never been published before, and no further release of T, or any subset thereof, is to be published after this anonymous dataset T∗ [68].

Publishing in parallel: In the parallel publishing hypothesis, the initial dataset T is published in several anonymous datasets T∗ i. According to the different requirements of the data recipients, each Ti∗ could consist of a subset of the initial features. The objective of parallel publications is to decrease the information loss of the anonymous dataset caused by the dimension curse. The data publisher should consider that data recipients may conspire and combine all available Ti∗ to obtain more information about the record owner.

Publishing in sequence: Sequential release refers to the incremental release of anonymised datasets [69]. For instance, consider that a company periodically publishes anonymised data about its customers. The data in T is likely to change over time, typically by adding new records and modifying or deleting existing records. The Data Publisher should not ignore already published datasets, as a record holder's anonymity is at risk when an adversary cross-posts multiple publications.

4.5.3 Centralised and decentralised publishing

In centralized data publishing, the responsibility of holding the complete data set lies with a single entity known as the data publisher. This entity collects data from multiple sources, such as multiple hospitals in a country, which can then send their datasets to the central entity. The central entity, in this case, the department of health, performs the task of anonymizing the collected data set, eliminating the need for individual data holders to perform this task themselves. In decentralized data publishing, the data is split among multiple data holders who don't have trust in each other. Nevertheless, the recipients of the data desire to analyse the relationship between their datasets. However, because of legal or business restrictions, the data holders do not want to grant access to the original data to any other party. In this context, data can be shared between different parties in various ways [70].

- Horizontal partitioning (HP): data holders have separate sets of individuals but with the same attributes.
- Vertical partitioning (VP): data holders have separate feature sets but with the same individuals.
- Arbitrary partitioning (AP): a hybrid between VP and HP. This is the most feasible scheme in which datasets may contain an undefined number of coincidental attributes and individuals.

In the scenario of several data holders who want to release their data to a common anonymised table T∗ there are two different approaches, namely anonymise-and-aggregate and aggregate-and-anonymise.

As shown in Figure 12, in the anonymise-and-aggregate approach, each data controller shall anonymise the data separately, then aggregate all its anonymised tables into one and deliver it to the data recipients [71]. However, for this approach, the cost is relatively manageable, and the solution is very fast, it has the disadvantage of introducing unnecessary data degradation.

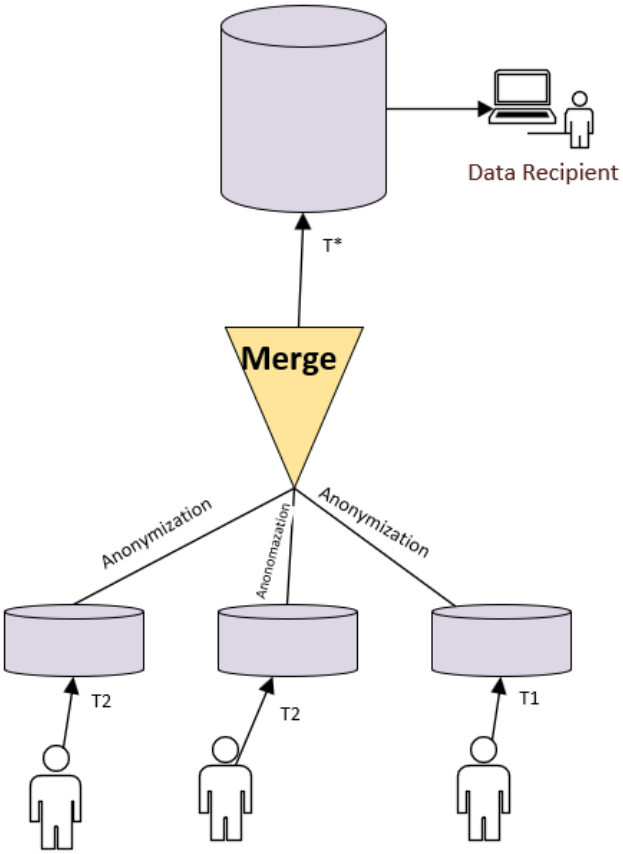

Figure 12: Anonymise-and-aggregate approach

On the other hand, the aggregate-and-anonymise approach (Figure 13) can result in less data distortion for the same privacy guarantees as the first. The process begins with data holders collecting all their original data and then undergoing the anonymization process. Legal restrictions may prevent the sharing of straightforward data with others, so alternative solutions such as involving a semi-trusted third party or utilizing secure multi-party computation protocols may be employed [72], [73], [74]. Nevertheless, this requires a considerable computational expense to ensure leakage-free encrypted communication.

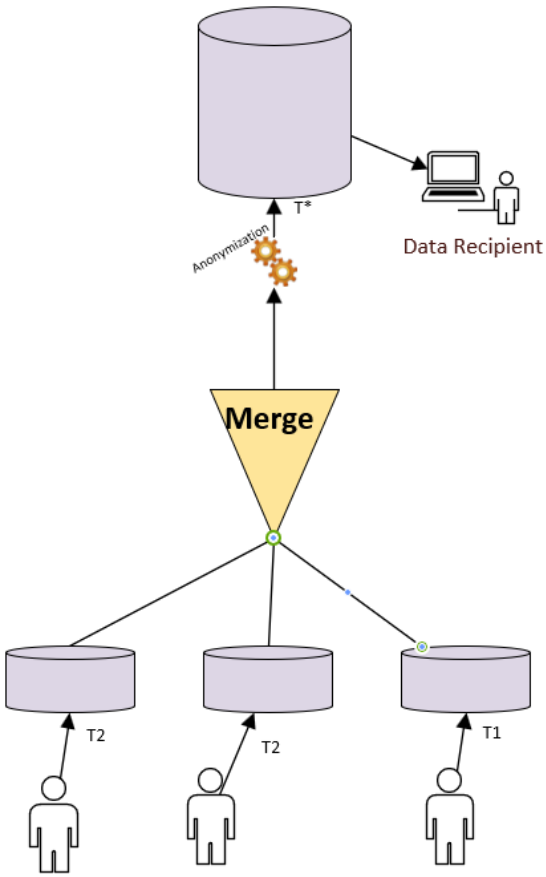

Figure 13: Aggregate and anonymise approach

In the area of anonymization, different attack scenarios involve opponents with different strengths and objectives. An opponent's goals vary from complete identification of record holders to simply learning whether its targets' record is included in a published dataset. The various scenarios can be classified into the following categories:

- Record Linkage: The objective of record linking is to successfully link an anonymous database record to an individual.
- Attribute Linkage: This concerns the case in which the opponent cannot link a particular record to its target but can link a particular value.
- Table Linkage Attack: We suppose that the opponent already knows that his target record is in the table of the released anonymous dataset.

An anonymous data publication hypothesis assumes that the original dataset has too many dimensions for an opponent to expect an attacker to have full knowledge of the target's file. For example, such a dataset is collected by a country's tax audit. The dataset has dozens of databases containing the financial information of individuals. The attackers may have partial knowledge of the victim's financial situation. For example, Table 9 contains income data for individuals. This multi-dimensional dataset with sparse data can be seen as a collection of datasets. Each figure in the table represents a different source of income, such as income from agriculture, donations, capital gains, and others. Typically, each record holder has income from a number of different subsets of all possible sources of income. It is a realistic assumption for an attacker to have knowledge of certain types of

income rather than the complete additional information of their target. In our example, George may know that Dimitris annual salary is 11,000 and further that his capital gains range from 11,000 to 22,000. When the unique identifiers in the anonymous publication of Table 9 are removed, George can use her partial knowledge to identify Dimitri's entry in the dataset and then his remaining income. That is, attackers can combine multiple datasets to identify the subject of interest.

Therefore, where the goal was anonymization, to protect the subjects it uses predefined data generalisation hierarchies. Each record is an item set, and each item is a value from a value-field set. The anonymization of such datasets is done with km-anonymity, which guarantees that any attackers who know up to *m* items of a target record cannot use this knowledge to identify fewer than k individuals in the dataset being parsed [75].

Table 9: Original Tax data

Name	Various income sources (annual)
George	{11000, 11000, 20000, 40000, 40000}
Grigoris	{11000, 30500, 40000}
Christos	{11000, 11000, 40000, 40000}
Dimitris	{11000}
Nikos	{20000}

Our goal will be to provide an anonymity guarantee to prevent identity leakage attacks without a predefined data hierarchy to reduce the loss of information of the original model. To accomplish this, we will need a generalization approach to overall re-coding that preserves utility by generalizing the minimum number of values required for each combination of m values occurring in at least *k* records in the dataset. The attacker is assumed to have no negative knowledge, which is logical for sparse multidimensional data and to know up to m values of a target record.

Let *T* be a sparse multidimensional matrix with continuous features $Q_1, Q_2, \ldots, Q_n$ of the same domain. Let *D* as the itemset representation of *T*, where each entry is a set of non-zero values from its corresponding entry in *T.* In this scenario, the attackers have limited knowledge of no more than *m* values from a target entry *t.* To recover entries in *D,* the attackers use unique or rare combinations of *m* values. If the attack is successful, the attackers can gain further knowledge about the target, including the remaining values.

There may be many different anonymizations of a dataset that meet the anonymity *km* for a given knowledge threshold *m* of the attacker. The worst-case scenario would involve fully anonymizing all values to their fullest extent. Although this option is technically feasible, it results in a significant loss of information and makes the released data nearly worthless. The challenge in determining the optimal km-anonymization is to find the most appropriate set of generalizations that meet the criteria for km-anonymization while minimizing the loss of information.

5. EVIDENT Directions for Data Documentation Items

The EVIDENT project aims to estimate the causal impact of behavioural biases on household energy consumption and conservation. Through five large-scale use cases, the project will collect and analyse data regarding consumers' perspectives on energy consumption and will explore how behavioural biases affect consumers' decision-making. Through different methodologies, such as randomized control trials, quasi-experiments, serious games and big data analytics, the EVIDENT project aspires to explain consumers' decisions regarding energy consumption patterns through socio-economic factors such as financial and environmental literacy.

Use cases 1 and 2 estimate the importance of consumer feedback and peer comparison feedback in household energy consumption. The main goals of these two use cases are to raise consumers' awareness regarding energy consumption, inform them on how they can save energy through personalized energy consumption tips and suggestions and explore the effectiveness nudges in energy conservation. Use case three collects and analyses big data to explore whether artificial intelligence and machine learning could provide answers to questions that previously could not be easily interpreted. In use case 3, two analytical frameworks are developed to provide. The first framework builds on hourly consumption data and tries to forecast energy consumption at a household level. In contrast, the second framework identifies the characteristics that drive consumers' performance in the context of natural field experiments. Use case four, through a serious game designed by the EVIDENT consortium, seeks to explore the impact of socio-demographic factors, environmental literacy, and financial literacy on consumer willingness to pay for the repair of home appliances. Use case 5 tries to identify the elements of the factor, such as consumers' demographics and energy-related financial and environmental literacy levers, which affect their willingness to pay for more expensive but more efficient household appliances.

To design and implement the previous use cases, mainly use cases 1, 2 and 3, the EVIDENT consortium leverages existing consumption and demographic data provided by the two energy companies, CW and PPC. Both companies provide consumption data regarding their clients along with a set of demographic data. For use cases 4 and 5, the EVIDENT consortium leverages the EVIDENT platform to design and implement three different e-lab experiments. That way, the consortium members collect and analyse participants' data. In all use cases, additional data (e.g., weather data) are collected from 3rd party services to be used as additional control variables in the analyses that take place in the EVIDENT use cases. Thus, there are three main sources of data leveraged in the context of the EVIDENT project, (a) the two energy companies, (b) the EVIDENT platform developed to cover the data collection needs of the project and (c) third-party services.

Starting from the data provided by the two energy companies. The first dataset provided by CW refers to prosumers' (consumers who also produce electricity through photovoltaics) energy consumption and production measurements and a set of demographic data for the previous prosumers. On the other hand, PCC provided energy consumption measurements and a set of demographic data from the newly introduced platform "myEnergyCoach" built to consult customers about consumption within their household and provide feedback, including personalized tips and suggestions for energy savings. The data provided by the two energy companies are accessible only to consortium partners responsible for the design and implementation of use cases 1, 2 and 3. The dataset contains anonymised information about the two companies' clients; thus, it is impossible to get public. The

corresponding data has been extensively described in D3.2 ‘Implementation of preparatory actions for RCT, surveys and serious game’ and D4.2 ‘Econometric analysis and robustness tests’.

In addition, the EVIDENT consortium developed the EVIDENT platform, a platform built to cover the data collection needs for the EVIDENT project but also provide an advanced ecosystem for the design and the implementation of e-lab experiments combining both questionnaires and serious games. The consortium partners will implement two e-labs experiments through the EVIDENT platform for use cases 4 and 5, respectively. The experiment regarding use case four consists of a serious game designed by the consortium to identify the socio-demographic factors that affect consumers’ willingness to pay for the repair of home appliances. The second experiment referring to use case five consists of a discrete choice experiment that explores the impact of energy-related financial literacy, demographic factors and environmental literacy on the discount rate and willingness to pay for efficient household appliances. The data collected from the previous experiments are collected through the platform and securely stored in the web service database. The platform owner, the partner responsible for implementing the e-labs, can access and download the collected data. However, access to the data presupposes that the data have been anonymised, and there is no way to identify the participant's identity. In addition, no personal information (e.g., name, address, IP address) is stored. The EVIDENT platform integrates the Zenodo Application Programming Interface (API) to enhance data publicity. Thus, the platform users can create depositions for their data and share them directly with Zenodo. Technical information about the integration of the Zenodo API will be described in D6.4 ‘Datahub Services of the EVIDENT platform’.

The EVIDENT consortium also collects data from third-party services to enhance the analytical frameworks developed by the partners. An example of this data is the weather conditions in the regions of Sweden where CW’s clients are located. The weather data are used as extra covariances for the analyses of use cases 1, 2, and 3.

6. Conclusion

This deliverable presents the data and how to manage them, their documentation and the additional fields that may need to be created to enable their handling. Likewise, the research data and the use of platforms for open science were highlighted. Furthermore, it analysed how to share research data concerning privacy. The above we are considering adopting for both EVIDENT research data, for use cases 4 and 5, and allowing researchers to publish their experiments directly on platforms like Zenodo, ensuring data privacy.

In this way, transparency is provided without compromising anyone. Participants can instantly see the results of the research and, at the same time, contribute to policy development. We will consider for third parties using the platform that the records will not remain on the platform but will be erased and stored in planforms like Zenodo. Naturally, these will all be customizable from the platform. An updated version will come on D5.4 'Updated Data Documentation', where we will report further on the actions that have been implemented.

References

[1] B. Heinrich, D. Hristova, M. Klier, A. Schiller, and M. Szubartowicz, "Requirements for Data Quality Metrics," *Journal of Data and Information Quality*, vol. 9, no. 2, pp. 1–32, Jun. 2017, doi: 10.1145/3148238.

[2] J. Byabazaire, G. O'Hare, and D. Delaney, "Data Quality and Trust: Review of Challenges and Opportunities for Data Sharing in IoT," *Electronics*, vol. 9, no. 12, p. 2083, Dec. 2020, doi: 10.3390/electronics9122083.

[3] T. Nkonyana and B. Twala, "Impact of Poor Data Quality in Remotely Sensed Data," 2018, pp. 79–86.

[4] L. Cai and Y. Zhu, "The Challenges of Data Quality and Data Quality Assessment in the Big Data Era," *Data Science Journal*, vol. 14, p. 2, May 2015, doi: 10.5334/dsj-2015-002.

[5] E. Gyulgyulyan, F. Ravat, H. Astsatryan, and J. Aligon, "Data Quality Impact in Business Inteligence," in *2018 Ivannikov Memorial Workshop (IVMEM)*, May 2018, pp. 47–51, doi: 10.1109/IVMEM.2018.00016.

[6] N. L. Eisner, A. L. Murray, M. Eisner, and D. Ribeaud, "A practical guide to the analysis of non-response and attrition in longitudinal research using a real data example," *International Journal of Behavioral Development*, vol. 43, no. 1, pp. 24–34, Jan. 2019, doi: 10.1177/0165025418797004.

[7] K. Olson, "Do non-response follow-ups improve or reduce data quality?: a review of the existing literature," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 176, no. 1, pp. 129–145, Jan. 2013, doi: 10.1111/j.1467-985X.2012.01042.x.

[8] P. Sedgwick, "Non-response bias versus response bias," *BMJ*, vol. 348, no. apr09 1, pp. g2573–g2573, Apr. 2014, doi: 10.1136/bmj.g2573.

[9] T. Mostafa and R. D. Wiggins, "The impact of attrition and non-response in birth cohort studies: a need to incorporate missingness strategies," *Longitudinal and Life Course Studies*, vol. 6, no. 2, Apr. 2015, doi: 10.14301/llcs.v6i2.312.

[10] R. Nishimura, J. Wagner, and M. Elliott, "Alternative Indicators for the Risk of Non-response Bias: A Simulation Study," *International Statistical Review*, vol. 84, no. 1, pp. 43–62, Apr. 2016, doi: 10.1111/insr.12100.

[11] F. J. Fowler, *Survey research methods*, 5th ed. Thousand Oaks: SAGE Publications, Inc, 2013.

[12] I. Outes-Leon and S. Dercon, "Survey attrition and attrition bias in Young Lives," Young Lives, 2009.

[13] C. Cichy and S. Rass, "An Overview of Data Quality Frameworks," *IEEE Access*, vol. 7, pp. 24634–24648, 2019, doi: 10.1109/ACCESS.2019.2899751.

[14] R. Y. Wang, "A product perspective on total data quality management," *Communications of the ACM*, vol. 41, no. 2, pp. 58–65, Feb. 1998, doi: 10.1145/269012.269022.

[15] L. P. English, *Improving data warehouse and business information quality: methods for reducing costs and increasing profits*. New York, NY, USA: Wiley, 1999.

[16] D. Loshin, *Enterprise Knowledge Management: The Data Quality Approach*. San Francisco, CA,

USA: Morgan Kaufmann, 2001.

[17] Y. W. Lee, D. M. Strong, B. K. Kahn, and R. Y. Wang, "AIMQ: a methodology for information quality assessment," *Information & Management*, vol. 40, no. 2, pp. 133–146, Dec. 2002, doi: 10.1016/S0378-7206(02)00043-5.

[18] L. L. Pipino, Y. W. Lee, and R. Y. Wang, "Data quality assessment," *Communications of the ACM*, vol. 45, no. 4, pp. 211–218, Apr. 2002, doi: 10.1145/505248.506010.

[19] C. Batini, F. Cabitza, C. Cappiello, C. Francalanci, and P. di Milano, "A Comprehensive Data Quality Methodology for Web and Structured Data," in *2006 1st International Conference on Digital Information Management*, Apr. 2007, pp. 448–456, doi: 10.1109/ICDIM.2007.369236.

[20] C. Cappiello, P. Ficiaro, and B. Pernici, "HIQM: A Methodology for Information Quality Monitoring, Measurement, and Improvement," 2006, pp. 339–351.

[21] M. Del Pilar Angeles and F. Garcia-Ugalde, "Assessing quality of derived non atomic data by considering conflict resolution function," *International Journal on Advances in Software*, vol. 2, no. 2, pp. 259–274, 2009.

[22] B. Carlo, B. Daniele, C. Federico, and G. Simone, "A Data Quality Methodology for Heterogeneous Data," *International Journal of Database Management Systems*, vol. 3, no. 1, pp. 60–79, Feb. 2011, doi: 10.5121/ijdms.2011.3105.

[23] L. Sebastian-Coleman, *Measuring data quality for ongoing improvement: a data quality assessment framework*. Waltham, MA: Morgan Kaufmann, 2013.

[24] R. Vaziri, M. Mohsenzadeh, and J. Habibi, "TBDQ: A Pragmatic Task-Based Method to Data Quality Assessment and Improvement," *PLOS ONE*, vol. 11, no. 5, p. e0154508, May 2016, doi: 10.1371/journal.pone.0154508.

[25] A. Sundararaman and S. K. Venkatesa, "Data Quality Improvement Through OODA Methodology," 2017.

[26] M. C. Paulk, C. V. Weber, B. Curtis, and M. B. Chrissis, *The Capability Maturity Model: Guidelines for Improving the Software Process*, vol. 441. Reading, MA, USA: Addison-Wesley, 1995.

[27] Dutch National Expertise Centre, "File Formats," 2022. [Online]. Available: https://dans.knaw.nl/en/file-formats/.

[28] J. Shen and J. Han, *Automated Taxonomy Discovery and Exploration*. Springer International Publishing, 2022.

[29] D. Lis and B. Otto, "Towards a Taxonomy of Ecosystem Data Governance," in *54th Hawaii International Conference on System Sciences*, 2021, pp. 6067–6077, doi: 10.24251/HICSS.2021.733.

[30] J. P. Reiter, "Inference for partially synthetic, public use microdata sets," *Survey Methodology*, vol. 29, no. 2, pp. 181–188, 2003.

[31] B. Murtha, *Introduction to Metadata*, 2nd ed. Los Angeles, CA: Getty Research Institute, 2008.

[32] J. Leipzig, D. Nüst, C. T. Hoyt, K. Ram, and J. Greenberg, "The role of metadata in reproducible computational research," *Patterns*, vol. 2, no. 9, p. 100322, Sep. 2021, doi: 10.1016/j.patter.2021.100322.

[33] National Information Standards Organization Framework Working Group, "A Framework of

Guidance for Building Good Digital Collections," 2007. [Online]. Available: https://www.niso.org/sites/default/files/2017-08/framework3.pdf.

[34] L. Perrier *et al.*, "Research data management in academic institutions: A scoping review," *PLOS ONE*, vol. 12, no. 5, p. e0178261, May 2017, doi: 10.1371/journal.pone.0178261.

[35] K. Briney, H. Coates, and A. Goben, "Foundational Practices of Research Data Management," *Research Ideas and Outcomes*, vol. 6, Jul. 2020, doi: 10.3897/rio.6.e56508.

[36] KU Leuven, "RDM Policy at KU Leuven," 2022. [Online]. Available: https://www.kuleuven.be/rdm/en/policy/policy.

[37] E. Mackey, "A Best Practice Approach to Anonymization," in *Handbook of Research Ethics and Scientific Integrity*, Cham: Springer International Publishing, 2020, pp. 323–343.

[38] OpenAIRE, "How to comply with Horizon Europe mandate for Research Data Management," 2022. [Online]. Available: https://www.openaire.eu/how-to-comply-with-horizon-europe-mandate-for-rdm.

[39] A. M. Cox and S. Pinfield, "Research data management and libraries: Current activities and future priorities," *Journal of Librarianship and Information Science*, vol. 46, no. 4, pp. 299–316, Dec. 2014, doi: 10.1177/0961000613492542.

[40] European Open Science Cloud, "EOSC | EOSC Portal." [Online]. Available: https://eosc-portal.eu/about/eosc.

[41] "OpenAIRE." https://www.openaire.eu/.

[42] European Commission Directorate-General for Research & Innovation H2020 Programme, "Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020," 2017. [Online]. Available: https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf.

[43] P. Manghi *et al.*, "The OpenAIRE Research Graph Data Model," 2019. doi: 10.5281/zenodo.2643199.

[44] S. Ribaric, A. Ariyaeeinia, and N. Pavesic, "De-identification for privacy protection in multimedia content: A survey," *Signal Processing: Image Communication*, vol. 47, pp. 131–151, Sep. 2016, doi: 10.1016/j.image.2016.05.020.

[45] S. Löbner, F. Tronnier, S. Pape, and K. Rannenberg, "Comparison of De-Identification Techniques for Privacy Preserving Data Analysis in Vehicular Data Sharing," in *Computer Science in Cars Symposium*, Nov. 2021, pp. 1–11, doi: 10.1145/3488904.3493380.

[46] A. Majeed and S. Lee, "Anonymization Techniques for Privacy Preserving Data Publishing: A Comprehensive Survey," *IEEE Access*, vol. 9, pp. 8512–8545, 2021, doi: 10.1109/ACCESS.2020.3045700.

[47] P. Samarati, "Protecting respondents identities in microdata release," *IEEE Transactions on Knowledge and Data Engineering*, vol. 13, no. 6, pp. 1010–1027, 2001, doi: 10.1109/69.971193.

[48] L. Sweeney, "Datafly: a system for providing anonymity in medical data," 1998, pp. 356–381.

[49] L. Sweeney, "k-Anonymity: A Mopdel for Protecting Privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 557–570, Oct. 2002,

doi: 10.1142/S0218488502001648.

[50] S. Murthy, A. Abu Bakar, F. Abdul Rahim, and R. Ramli, "A Comparative Study of Data Anonymization Techniques," in *2019 IEEE 5th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing, (HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS)*, May 2019, pp. 306–309, doi: 10.1109/BigDataSecurity-HPSC-IDS.2019.00063.

[51] J. Jayapradha, M. Prakash, Y. Alotaibi, O. I. Khalaf, and S. A. Alghamdi, "Heap Bucketization Anonymity—An Efficient Privacy-Preserving Data Publishing Model for Multiple Sensitive Attributes," *IEEE Access*, vol. 10, pp. 28773–28791, 2022, doi: 10.1109/ACCESS.2022.3158312.

[52] R. Indhumathi and S. Sathiya Devi, "Anonymization Based on Improved Bucketization (AIB): A Privacy-Preserving Data Publishing Technique for Improving Data Utility in Healthcare Data," *Journal of Medical Imaging and Health Informatics*, vol. 11, no. 12, pp. 3164–3173, Dec. 2021, doi: 10.1166/jmihi.2021.3901.

[53] D. Li, X. He, L. Cao, and H. Chen, "Permutation anonymization," *Journal of Intelligent Information Systems*, vol. 47, no. 3, pp. 427–445, Dec. 2016, doi: 10.1007/s10844-015-0373-4.

[54] J. Domingo-Ferrer and K. Muralidhar, "New directions in anonymization: Permutation paradigm, verifiability by subjects and intruders, transparency to users," *Information Sciences*, vol. 337–338, pp. 11–24, Apr. 2016, doi: 10.1016/j.ins.2015.12.014.

[55] A. Kiran and N. Shirisha, "K-Anonymization approach for privacy preservation using data perturbation techniques in data mining," *Materials Today: Proceedings*, vol. 64, pp. 578–584, 2022, doi: 10.1016/j.matpr.2022.05.117.

[56] C. Eyupoglu, M. Aydin, A. Zaim, and A. Sertbas, "An Efficient Big Data Anonymization Algorithm Based on Chaos and Perturbation Techniques," *Entropy*, vol. 20, no. 5, p. 373, May 2018, doi: 10.3390/e20050373.

[57] J. Domingo-Ferrer and Ú. González-Nicolás, "Hybrid microdata using microaggregation," *Information Sciences*, vol. 180, no. 15, pp. 2834–2844, Aug. 2010, doi: 10.1016/j.ins.2010.04.005.

[58] J. Yoon, L. N. Drumright, and M. van der Schaar, "Anonymization Through Data Synthesis Using Generative Adversarial Networks (ADS-GAN)," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 8, pp. 2378–2388, Aug. 2020, doi: 10.1109/JBHI.2020.2980262.

[59] J. Domingo-Ferrer and J. M. Mateo-Sanz, "Practical data-oriented microaggregation for statistical disclosure control," *IEEE Transactions on Knowledge and Data Engineering*, vol. 14, no. 1, pp. 189–201, 2002, doi: 10.1109/69.979982.

[60] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "L-diversity: Privacy beyond k-anonymity," *ACM Transactions on Knowledge Discovery from Data*, vol. 1, no. 1, p. 3, Mar. 2007, doi: 10.1145/1217299.1217302.

[61] L. Yao, X. Wang, X. Wang, H. Hu, and G. Wu, "Publishing Sensitive Trajectory Data Under Enhanced l-Diversity Model," in *2019 20th IEEE International Conference on Mobile Data Management (MDM)*, Jun. 2019, pp. 160–169, doi: 10.1109/MDM.2019.00-61.

[62] N. Li, T. Li, and S. Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity," in *2007 IEEE 23rd International Conference on Data Engineering*, Apr. 2007, pp. 106–115, doi: 10.1109/ICDE.2007.367856.

[63] G. Hao and X. Ya-Bin, “Research on privacy preserving method based on T-closeness model,” in *2017 3rd IEEE International Conference on Computer and Communications (ICCC)*, Dec. 2017, pp. 1455–1459, doi: 10.1109/CompComm.2017.8322783.

[64] C. Dwork, F. McSherry, K. Nissim, and A. Smith, “Calibrating Noise to Sensitivity in Private Data Analysis,” in *Theory of Cryptography*, 2006, pp. 265–284.

[65] A. Wood *et al.*, “Differential Privacy: A Primer for a Non-Technical Audience,” *Vanderbilt Journal of Entertainment and Technology Law*, vol. 21, no. 1, p. 209, 2020.

[66] B. C. M. Fung, Ke Wang, and P. S. Yu, “Top-Down Specialization for Information and Privacy Preservation,” in *21st International Conference on Data Engineering (ICDE’05)*, pp. 205–216, doi: 10.1109/ICDE.2005.143.

[67] J. M. Joyce, “Kullback-Leibler Divergence,” in *International Encyclopedia of Statistical Science*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 720–722.

[68] C. Yao, X. S. Wang, and S. Jajodia, “Checking for K-Anonymity Violation by Views,” in *Proceedings of the 31st International Conference on Very Large Data Bases*, 2005, pp. 910–921.

[69] K. Wang and B. C. M. Fung, “Anonymizing sequential releases,” in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, Aug. 2006, pp. 414–423, doi: 10.1145/1150402.1150449.

[70] H. Polat and W. Du, “Privacy-Preserving Collaborative Filtering on Vertically Partitioned Data,” 2005, pp. 651–658.

[71] A. Solanas, A. Martinez-Balleste, and J. M. Mateo-Sanz, “Distributed Architecture With Double-Phase Microaggregation for the Private Sharing of Biomedical Data in Mobile Health,” *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 6, pp. 901–910, Jun. 2013, doi: 10.1109/TIFS.2013.2248728.

[72] C. Zhao *et al.*, “Secure Multi-Party Computation: Theory, practice and applications,” *Information Sciences*, vol. 476, pp. 357–372, Feb. 2019, doi: 10.1016/j.ins.2018.10.024.

[73] J. Zhou, Y. Feng, Z. Wang, and D. Guo, “Using Secure Multi-Party Computation to Protect Privacy on a Permissioned Blockchain,” *Sensors*, vol. 21, no. 4, p. 1540, Feb. 2021, doi: 10.3390/s21041540.

[74] D. Karapiperis and V. S. Verykios, “An LSH-Based Blocking Approach with a Homomorphic Matching Technique for Privacy-Preserving Record Linkage,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 4, pp. 909–921, Apr. 2015, doi: 10.1109/TKDE.2014.2349916.

[75] M. Terrovitis, N. Mamoulis, and P. Kalnis, “Privacy-preserving anonymization of set-valued data,” *Proceedings of the VLDB Endowment*, vol. 1, no. 1, pp. 115–125, Aug. 2008, doi: 10.14778/1453856.1453874.