|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | INTERNATIONAL TELECOMMUNICATION UNION  **TELECOMMUNICATION STANDARDIZATION SECTOR**  STUDY PERIOD 2022-2024 | | FG-AI4H-S-036 | |
| **ITU-T Focus Group on AI for Health** | |
| **Original: English** | |
| **WG(s):** | | Plenary | Geneva, 3-5 July 2023 | |
| **DOCUMENT** | | | | |
| **Source:** | | Ministry of Telecommunications and Technology | | |
| **Title:** | | Health Data Management and Governance for Trustworthy AI | | |
| **Contact:** | | Anas DAHABIAH, Ph.D. Director of IT Development  Ministry of Communications and Technology  Syrian Arab Republic | | Tel: +963931608608 Fax: +963112214578 Email: [planning@moct.gov.sy](mailto:planning@moct.gov.sy) |

|  |  |
| --- | --- |
| **Abstract:** | The rapid growth in digital transformation and the Internet of Things has led to the generation of a huge volume of data, which has become the fuel necessary to push the wheel of artificial intelligence forward, especially due to the current trend to adopt deep learning systems in many applications of the Fourth Industrial Revolution, particularly in the health and clinical field in which data (in its confidentiality and privacy) represents the backbone to take decisions.  From here, it was necessary to work diligently to devise the best ways to manage and govern this data properly, provided that this management includes: processes, people, and tools. Furthermore, it should be applied to all types of medical data (Master Data, Reference Data, and Metadata).  In this contribution, we propose a unified framework for health data management that is based on the essential pillars of management: Quality Management, Privacy and Security Management, Analytics Management, Technical Management, and above all Data Governance which is the core of Data Management connected to all the aforementioned knowledge area, according to DAMA (**Da**ta **M**anagement **A**ssociation) applied with respect to clinical data principles and ethics, indicated by the “European’s high-level expert group” and deliverable 1 [6] of our group. We also show that it is important to have this global vision of management and understand the relations and the dependencies among these sectors when planning, designing, and deploying the processes and models during the life-cycle of data.  FG-AI4H deliverable 5\_4 introduced data technical requirements specifications for datasets used in AI models. This contribution proposes other essential technical specifications in addition to management requirements specifications in master, reference, and meta health data  **Acronyms**: **AI**: Artificial Intelligence, **DAMA**: International Data Management Association, **DG**: Data Governance, **DQ**: Data Quality, **MDM**: Master Data Management, **RFM**: Reference Data Management, **ICD**: International Classification of Diseases, **CDE**: Critical Data Element, **CNN**: Convolutional Neural Network, **ETL**: Extract, Transform, Load. |

# Introduction:

It has become necessary and essential to know the art of “medical data management and governance”, because of data exponential increase in the health sector due to the:

* Riding the wave of the upcoming “digital transformation revolution”,
* The availability and cheapness of the tools that generate this data, like the various medical imaging devices, analysis equipment, and electronic health record.
* The availability of various types of affordable storage media, allows the preservation and archiving of these amounts.

Figure 1 depicts the main sources of big data in health.

A picture containing text, screenshot, circle, font

Description automatically generated

Fig 1. The sources of big data in health (source: NEJM Catalyst © Massachusetts Medical Society)

This data has the potential to achieve improvement and prosperity in the health domain, only if it is well-exploited, managed, and invested. Some interesting case studies are also introduced in Fig 2.

A picture containing text, screenshot, software, computer icon

Description automatically generated

Fig 2. Useful case studies of health big data exploitation (source: NEJM Catalyst © Massachusetts Medical Society)

Figure 3 presents the opportunities for big data in health care:

A picture containing text, screenshot, circle, font

Description automatically generated

Fig 3. Opportunities of Big Data @medudoc (based on Sanskruti Patel & Atul Patel in International Journal of Information Sciences and Techniques (IJIST) Vol.6, №1/2, March 2016)

Nevertheless, all the efforts to get these opportunities and benefits are useless and don’t pay off, unless robust management and governance of data are conducted to ensure data quality and data privacy. Data management should mainly include and conduct:

* **Data Quality**: Artificial intelligence and data-driven decision systems rely mainly on the data used in training and validation, and on the structure of the designed model. Accordingly, data should be pre-processed and cleansed in order to get good results and increase the trustworthiness of the systems. Data mining and knowledge discovery techniques can lead to disastrous rules and decisions because of poor quality data, according to the principle “garbage-in garbage-out”.
* **Data Privacy and Security**: Data privacy is the control of private/confidential / Personal Identifying Information (PII) through policy and compliance monitoring. Data security is the protection of data assets through controls for the availability, usability, integrity, consistency, auditability, and confidentiality of data. Good data management should also ensure regulatory compliance (i.e., USA: **HIPAA** [**H**ealth **I**nsurance **P**robability and **A**ccountability **A**ct], Europe: **GDPR** [**G**eneral **D**ata **P**rotection **R**egulation], etc.)
* **Data Analytics and Intelligence**: Data is considered nowadays the digital assets of any organization: clinics, hospitals, insurance companies, etc. Useful insights and actionable decisions could be extracted from it, to get information and knowledge out of the raw data. The success of this phase depends mainly on the quality of the critical data elements CDEs, and the robust management of the metadata used in the descriptive, predictive, prescriptive, or cognitive models.
* **Data Governance**: Data Governance (DG) is defined as the exercise of authority and control (planning, monitoring, and enforcement) over the management of data assets. The Data Governance function guides all other data management functions. The purpose of Data Governance is to ensure that data is managed properly, according to policies and best practices. While the driver of data management overall is to ensure that the organization gets value out of its data. Actually, data governance allows organizations to balance two needs: the need to collect and secure information (data privacy and security) while also getting value from that information (data analytics and intelligence).

In the following sectors, we will explain these sectors in more detail.

# Health Data Quality Management:

DQ Management is the planning, implementation, and control of activities that apply quality management techniques to data, in order to assure it is fit for consumption and meets the needs of data consumers, through the following steps:

1. Define High-Quality Data, the dimensions of quality, the domain rules, and the metrics and indicators. There are many data quality dimensions introduced in the literature but we will introduce in the next section the quantitative ones, proposed by the DAMA (**D**ata **M**anagement **A**ssociation), in order to be able to measure the performance.
2. Define the Scope of the Initial Assessment, based on data visuals and “Data Profiling” using specialized tools, like Microsoft Powe BI, Power Query, Power Pivot, Python (Pandas, Numpy, Matplotlib, Seaborn, etc.), the ggplot library in R language, etc. Although some data quality issues can be discovered during data profiling activity, the purpose of data profiling is to give insights for data quality assessment.
3. Define a “Data Quality Strategy” that accounts for the work which needs to be done and the way people will execute it. A robust framework should include methods to:

• Understand and prioritize the domain needs

• Identify the critical data elements and their associated quality dimensions to meet the needs.

• Assess data against expectations.

• Share findings and get feedback from stakeholders.

• Prioritize and manage issues.

• Identify and prioritize opportunities for improvement.

• Measure, monitor, and report on data quality.

• Manage Metadata produced through data quality processes.

• Integrate data quality controls into business and technical processes.

1. Prioritize Actions based on the Domain Impact

5. Develop Preventative and Corrective Actions

6. Confirm Planned Actions

7. Develop and Deploy Data Quality Operations as follows

Develop Data Quality Operational Procedures.

Correct Data Quality Defects.

Measure and Monitor Data Quality.

Report on Data Quality levels and findings.

Data Quality management in health has the same importance as data privacy and security because poor data in some cases can lead to bad decisions, incorrect medical rules and inferences, and even disastrous results.

**Data Quality Dimensions:**

A Data Quality dimension is a measurable feature or characteristic of data. Data quality dimensions can be used to define the results of initial data quality assessment as well as ongoing measurement. In order to measure the quality of data, an organization needs to establish characteristics that are both important to the domain processes (worth measuring) and measurable. Dimensions provide a basis for measurable rules, which themselves should be directly connected to potential risks in critical processes. All the dimensions cited in the literature focus on whether there is enough data (completeness), whether it is right (accuracy, validity), how well it fits together (consistency, integrity, uniqueness), whether it is up-to-date (timeliness), accessible, usable, and secure. The most applicable and practical measurable dimensions are accuracy, completeness, consistency, validity, uniqueness, and timeliness (Fig 4).

A diagram of data quality

Description automatically generated with medium confidence

Fig 4. Data Quality Dimensions.

Each quality dimension has a specific metric, which measures its performance. There are several data quality dimensions, which can be organized into 4 categories: intrinsic, contextual, accessibility, and representational. Two important categories (intrinsic and contextual) are illustrated in Fig. 5.

A diagram of data quality dimensions

Description automatically generated

Fig 5. The categories of DQ dimensions [4]

ISO 8000, the international standard for data quality, is being developed to enable the exchange of complex data in an application-neutral form. In the introduction to the standard, ISO asserts: “The ability to create, collect, store, maintain, transfer, process and present data to support business processes in a timely and cost-effective manner requires both an understanding of the characteristics of the data that determine its quality and an ability to measure, manage and report on data quality.

It would be appropriate here to get advantage of the unified DQ framework proposed in [4], and illustrated in Fig 6, where all the components cooperate, relying on the Data Quality Profile. It is initially created as a Data Profile and is progressively extended from the data collection phase to the analytics phase to capture important quality-related information. For example, it contains quality requirements, targeted data quality dimensions, quality scores, and quality rules.

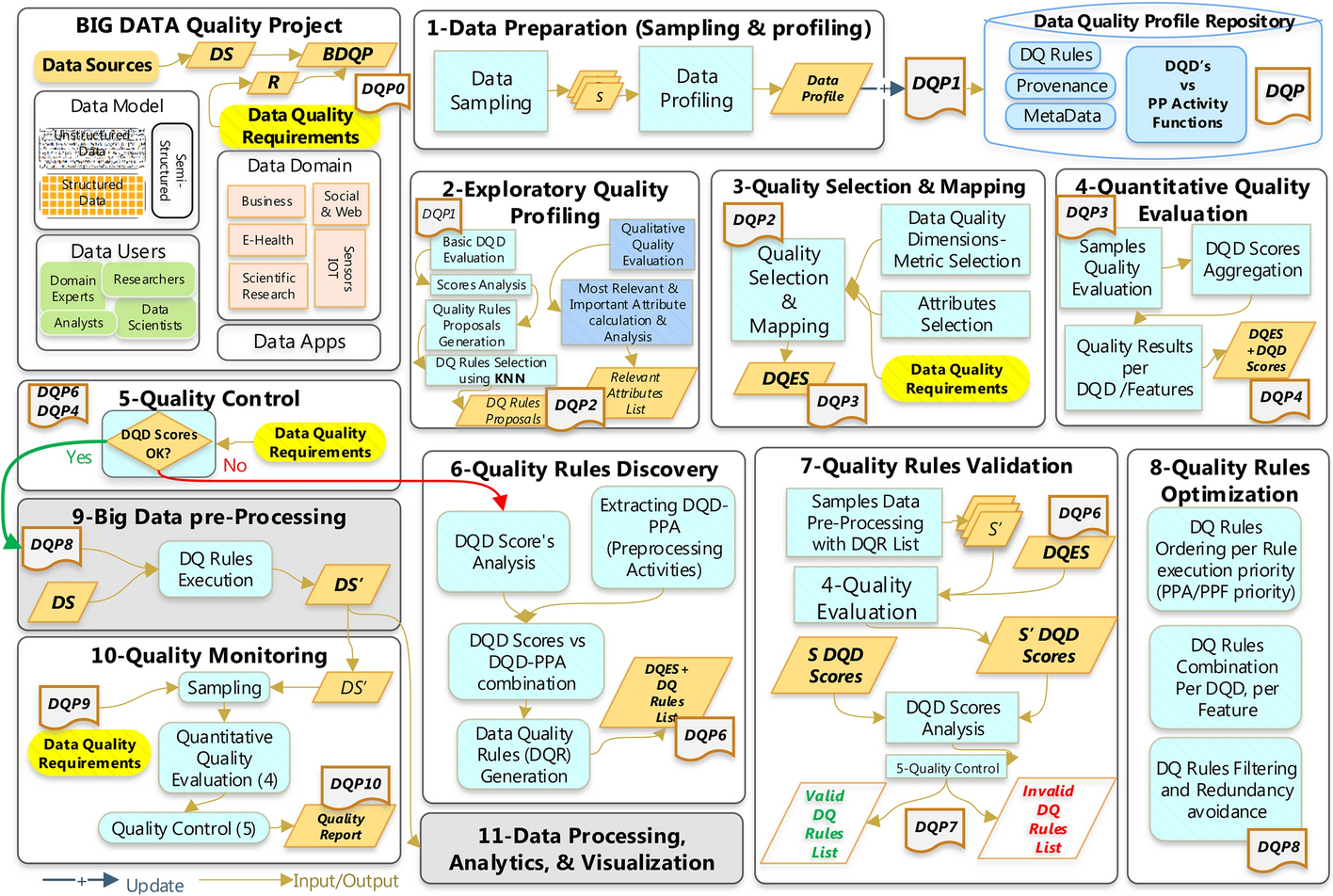


Fig5. General DQ Framework

3- **Clinical Data Security and Privacy**: Data Security includes the planning, development, and execution of security policies and procedures to provide proper authentication, authorization, access, and auditing of data and information assets. The specifics of data security (which data needs to be protected, for example) differ between industries and countries. Nevertheless, the goal of data security practices is the same: To protect information assets in alignment with privacy and confidentiality regulations, contractual agreements, and business requirements.

These requirements come from:

Stakeholders: Organizations must recognize the privacy and confidentiality needs of their stakeholders, including clients, patients, doctors, citizens, suppliers, or partners. Everyone in an organization must be a responsible trustee of data about stakeholders.

Government regulations: Government regulations are in place to protect the interests of some stakeholders. Regulations have different goals. Some restrict access to information, while others ensure openness, transparency, and accountability.

Proprietary domain concerns: Each organization has proprietary data to protect. An organization’s data provides insight into its customers and, when leveraged effectively, can provide a competitive advantage. If confidential data is stolen or breached, the trust among all the parties will be destructed.

Legitimate access needs: When securing data, organizations must also enable legitimate access. Business processes require individuals in certain roles be able to access, use, and maintain data.

Contractual obligations: Contractual and non-disclosure agreements also influence data security requirements.

Effective data security policies and procedures ensure that the right people can use and update data in the right way and that all inappropriate access and update is restricted.

Data Security Management activities focus on the following steps:

1. Identify Relevant Data Security Requirements

2. Define Data Security Policy

3. Define Data Security Standards

4. Assess Current Security Risks

5. Implement Controls and Procedures

**4- Clinical Data Governance**: Data governance is important for a healthcare organization because these organizations deal with large volumes of sensitive data, are faced with numerous complex regulations, and are typically rather siloed in operation. Data governance helps because healthcare organizations unlock the value of their data so they can use their data to make impactful decisions.

The scope and focus of a particular data governance program will depend on organizational needs, but most programs include:

• Strategy: Defining, communicating, and driving execution of Data Strategy and Data Governance

Strategy.

• Policy: Setting and enforcing policies related to data and Metadata management, access, usage, security, and quality

• Standards and Quality: Setting and enforcing Data Quality and Data Architecture standards

• Oversight: Providing hands-on observation, audit, and correction in key areas of quality, policy, and data management (often referred to as stewardship)

• Compliance: Ensuring the organization can meet data-related regulatory compliance requirements

• Issue management: Identifying, defining, escalating, and resolving issues related to data security, data access, data quality, regulatory compliance, data ownership, policy, standards, terminology, or data governance procedures

• Data management projects: Sponsoring efforts to improve data management practices

• Data asset valuation: Setting standards and processes to consistently define the business value of data assets

To accomplish these goals, a Data Governance program will develop policies and procedures, cultivate data stewardship practices at multiple levels within the organization, and engage in organizational change management efforts that actively communicate to the organization the benefits of improved data governance and the behaviors necessary to successfully manage data as an asset. Figure 6 simplifies the DG framework which is the core of all the other management domains. This Model aligns with “Gartner’s Golden Triangle” (2017) of ‘People, Process and Technology’ (with Data at the centre), which recognizes that effective data governance is an ongoing effort executed by people, enabled by repeatable processes, and supported by technology.

A diagram of a government model

Description automatically generated with low confidence

Fig 6. General Data Governance Model

Data governance enables shared responsibility for data-related decisions. Data governance activities cross-organizational and system boundaries in support of an integrated view of data. Successful data governance requires a clear understanding of what is being governed and who is being governed, as well as who is governing.

**5- Technical Management**: technical management is the framework concerned with data acquisition, data integration, data modeling, data architecture, data interoperability, and data warehousing. The majority of the big medical organizations are moving towards building “XaaS Private Clouding” where everything is provided as a service, in a secure environment supported by a CIRT (**C**yber **I**ncident **R**esponse **T**eam), using data buses capable to ensure trusted communication between the different ends, regardless of their programming languages, protocols, and datasets.

**6- Data Ethics**: Ethics are principles of behavior based on ideas of right and wrong. Ethical principles often focus on ideas such as fairness, respect, responsibility, integrity, quality, reliability, transparency, and trust. Data handling ethics are concerned with how to procure, store, manage, use, and dispose of data in ways that are aligned with ethical principles. Handling data in an ethical manner is necessary for the long-term success of any organization that wants to get value from its data. Unethical data handling can result in the loss of trustworthiness. All the aforementioned management pillars are nothing if they are not constructed on the ethics base and foundation, particularly in the medical domain [6].

The “European’s high-level expert group” defined the “trustworthy clinical AI” as a system where the users and the affected people can trust that the system, the processes, the data, and the people behind it are aligned with the foundational human values. They cited that the trustworthy AI should meet three objectives during its complete life cycle:

1. It should comply with applicable laws and regulations.
2. It should adhere to ethical principles and values.
3. It should be robust, both technically and from a social perspective.

The expert group set the four main ethical principles as follows:

1. Respect for human autonomy and freedom: AI systems should not hinder humans to make self-determined decisions. They should especially not deceive or manipulate humans. They will instead complement and empower their skills.
2. Prevention of harm: AI should only be used in safe and secure environments and the system should not be open to malicious use.
3. Fairness: there should be an equal distribution of costs and benefits between the different groups of society. It would also be ensured that there is no unfair bias discrimination against certain groups or individuals and no one should be deceived or impaired in their freedom of choice. Fairness also means fostering equal access to developed technologies.
4. Explicability: an explicable AI system can explain its decision processes. It is used in transparent processes and its purpose, capabilities, and limitations are openly communicated. This is especially the case in healthcare, where the AI system can be interacting with highly skilled professionals who want to understand how applicable the AI system’s decisions are. Moreover, the decisions of explicable AI systems are much easier to communicate to patients than the decisions of inexplicable ones. In addition to the AI’s decision process, it should also be documented what data the AI was trained on, to better identify the reasons why a given decision might be correct or erroneous. Furthermore, AI systems should always be identifiable as AIs, not pretend to be humans. It should be clear why they are used and what their capabilities and limitations are.

**7- Reference and Master Data Management**: Master data can be defined as the single source of basic medical data used across multiple systems, applications, and processes (i.e. patient and doctors data, medical drugs, injections, and products, or the locations of hospitals, health-care centres, pharmacies, blood analysis labs, etc.). Master data management (MDM) is the fact of managing shared data to meet organizational goals, reduce risks associated with data redundancy, ensure higher quality, and reduce the costs of data integration. It has three main phases:

1. Identify that records are potential matches.
2. Applying the medical rules to match and merge the records.
3. Creating the master record with trusted attributes.

On the contrary, Reference Data is data that is used solely to characterize other data in a hospital, a medical centre, a pharmacy, etc., or solely to relate data in a database to information beyond the boundaries of the organization. Reference Data Management (RDM) entails control over defined domain values and their definitions. The goal of RDM is to ensure the organization has access to a complete set of accurate and current values for each concept represented. The most basic Reference Data consists of codes and descriptions, but some Reference Data can be more complex and incorporate mappings and hierarchies. Reference Data Management entails control and maintenance of defined domain values, definitions, and the relationships within and across domain values. The goal of Reference Data Management is to ensure values are consistent and current across different functions and that the data is accessible to the organization. Depending on the granularity and complexity of what the Reference Data represents, it may be structured as a simple list, a cross-reference, or a taxonomy.

Industry Reference Data is a broad term to describe data sets that are created and maintained by industry associations or government bodies, rather than by individual organizations, in order to provide a common standard for codifying important concepts. This codification leads to a common way to understand data, and is a prerequisite for data sharing and interoperability. For example, the International Classification of Diseases (ICD) codes provide a common way to classify health conditions (diagnoses) and treatments (procedures) and thus to have a consistent approach to delivering health care and understanding outcomes. If every doctor and hospital create their own code set for diseases, it would be virtually impossible to understand trends and patterns.

**8- Meta-Data Management**: It involves managing data about other data, whereby this “other data” is generally referred to data models and structures, not the content. It includes managing information about data structures from different models and their mutual association (e.g. the medical terms in the glossary, attributes in the logical data model, or tables and columns in the database, as well as their associations). There are 3 main types of meta-data:

1. Technical Meta-Data: describes data elements from the technical user perspective and includes information like logical data models, source and target systems, tables and fields structures and attributes, as well as cross-model dependencies.
2. Operational Meta-Data: includes information about application runs: their frequency, record counts, component-by-component analysis and other statistics for auditing purposes.
3. The Domain Meta-Data: describes data elements from the medical uses perspective and includes the medical glossaries with terms and definitions, synonyms, acronyms, medical rules, and ownership. This plays a remarkable role in data mining, impact analysis, and data lineage (the traceable path for specific critical data elements from end-user reports upstream to the ultimate source including all the aggregated sources such as data warehouses and data marts, operational data stores, staging areas, and transactional systems.

Meta-Data Management Process consists of five phases:

1. Identify Critical Data Elements (CDEs): includes activities to analyze the business domain requirements, conduct interviews with the stakeholders and identify in-scope the critical data elements.
2. Collect the CDE’s Domain Meta-Data: for in-scope CDEs, define the domain term, synonyms, acronyms, definitions, and taxonomies. Also, define the domain rules and determine the ownership.
3. Collect the CDE’s Technical Meta-Data: for in-scope CDEs, identify CDE representations in data systems, and determine authoritative data source and data lineage. Also, determine associations with logical data models, if exists.
4. Create CDE Data Standards (known as 360 View): create associations between the CDE’s domain and technical metadata, and validate CDE data standards.
5. Enforce CDE Data Standard: enforce the data standard to ensure that every CDE is managed and used according to the defined standard (Data Governance).

**9-** **Added Values of our Contribution**: In the deliverable “DEL 5\_4”| FG-AI4H-R-068 [5] only general interesting technical data requirements specifications were proposed in each phase of the data life-cycle without distinguishing the types of data (mainly: master, meta, and reference data), although these specifications should be aligned with the management and governance specifications to lead to real fruitful trustworthy AI systems and solutions, and some specifications should be exploited in all the stages of life-cycle as it has been proved in DAMA Body of Knowledge [1] and should include process-people-tools according to Gartner (Golden Triangle). Concerning the points discussed in the aforementioned sections and looking closely at the following figure (Fig 7) which depicts the health data ecosystem and its different sources, one can realize that important requirements and specifications regarding vital issues in data were overlooked in the deliverable, particularly, data policy, data classification, data integration, data fusion and merging, data architecture, data interoperability, data assessment, etc. In the following, we will try to cover some of them.

A picture containing text, menu, font

Description automatically generated

Fig 7. Health data ecosystem

In Table 1 of deliverable 5 [5], we recommend adding the data augmentation specifications as it is widely used in medical deep-learning-based research which gets use of the convolutional neural networks in medical imagery classification, recognition, and categorization. In addition, data classification as confidential, private, public, etc. should be considered in this phase. Furthermore, the connections and relationships among the data and among the processes and the processes and the technology that supports these processes should be specified. Moreover, The design and architecture of systems and the data they produce and store should be indicated here. Above all, data policies and strategies used to advance organizational strategy should be clear in this stage. Planning for better data requires a strategic approach to architecture, modeling, and other design functions. It also depends on strategic collaboration between management and IT leadership. Robust data governance is a must here to assign the main roles (data chief officer, data owners, data steward, data custodians, etc.).

The plans, a compelling vision of data management, guiding principles, values, and perspectives, the mission and long-term directional goals of data management, proposed measures of data management success, roles and responsibilities should be specified in the data strategies.

Organizations, hospitals, medical research units, assurance companies, etc., require reliable Metadata to manage data as an asset and augment the data valuation. Metadata in this sense should be understood comprehensively. It includes not only the domain, technical, and operational Metadata described in the sections above, but also the Metadata embedded in Data Architecture, data models, data security requirements, data integration standards, and data operational processes. Metadata describes what data we have, what it represents, how it is classified, where it came from, how

it moves within the organization, how it evolves through use, who can and cannot use it, and whether it is of high quality, and above all the reliability of the systems that generate them. This last aspect is of high importance in the data fusion systems that depend on evidential and possibilistic approaches like Dempster-Shafer, Dezert-Smarandache, and Dubois-Prade approaches. Data is abstract, but definitions and other descriptions of context enable it to be understood. They make data, the data lifecycle, and the complex systems that contain data comprehensible. The challenge is that Metadata is a form of data and needs to be managed as such. Organizations that do not manage their data well generally do not manage their metadata at all. Metadata management often provides a starting point for improvements in data management overall.

Well-managed data is managed strategically, with a vision of how the organizations will use their data. A strategic organization will define not only its data content requirements but also its data management requirements. These include policies and expectations for use, quality, controls, and security; an approach to architecture and design; and a sustainable approach to both infrastructure and software development.

Data can be classified by type of data (e.g., transactional data, Reference Data, Master Data, Metadata; alternatively category data, resource data, event data, detailed transaction data) or by content (e.g., data domains, subject areas) or by format or by the level of protection the data requires. Data can also be classified by how and where it is stored or accessed. Because different types of data have different requirements, are associated with different risks, and play different roles within an organization, many of the tools of data management are focused on aspects of classification and control. For example, Master Data has different uses and consequently different management requirements than transactional data does.

“Issues Management Specifications” are also essential to succeed in identifying, defining, escalating, and resolving issues related to data acquisition, data security, data access, data quality, regulatory compliance, data ownership, policy, standards, terminology, or data governance procedures.

“Data Profiling Specifications” must also be a part of Table 1 to achieve a good assessment of our data quality from the beginning.

Understanding data content and structure is essential to the successful integration of data. Data profiling contributes to this end. Actual data structure and contents always differ from what is assumed. Sometimes differences are small; other times they are large enough to derail an integration effort. Profiling can help integration teams discover these differences and use that knowledge to make better decisions about sourcing and design. If data profiling is skipped, then the information that should influence design will not be discovered until testing or operations.

Basic profiling involves the analysis of:

• Data format as defined in the data structures and inferred from the actual data

• Data population, including the levels of null, blank, or defaulted data

• Data values and how closely they correspond to a defined set of valid values

• Patterns and relationships internal to the data set, such as related fields and cardinality rules

• Relationships to other data sets

More extensive profiling of the potential source and target data sets is required to understand how well the data meets the requirements of the particular data integration initiative. Profile both the sources and targets to understand how to transform the data to match requirements.

“Data Modeling and Design Specifications” should also be added including Data model management procedures, data modeling naming conventions, definition standards, standard domains, and standard abbreviations.

“Data Storage and Operations Specifications” should also be appended including tool standards, standards for database recovery and business continuity, database performance, data retention, and external data acquisition.

Another table dedicated to “Data Architecture Specifications” should be added containing: “Data Architecture outcomes”, such models, definitions, and data flows on various levels, usually referred to as Data Architecture artifacts and “Data Architecture activities”, to form, deploy and fulfill Data Architecture intentions.

Data Architecture is fundamental to data management. Because most organizations have more data than individual people can comprehend, it is necessary to represent organizational data at different levels of abstraction so that it can be understood and management can make decisions about it.

Data Architecture artifacts include specifications used to describe existing states, define data requirements, guide data integration, and control data assets as put forth in a data strategy. An organization’s Data Architecture is described by an integrated collection of master design documents at different levels of abstraction, including standards that govern how data is collected, stored, arranged, used, and removed. It is also classified by descriptions of all the containers and paths that data takes through an organization’s systems.

Data Flow and Data Lineage specification is an important pillar for relevant data management and should be added as well, as it depicts how data moves through business processes and systems. End-to-end data flows illustrate where the data originated, where it is stored and used, and how it is transformed as it moves inside and between diverse processes and systems. Data lineage analysis can help explain the state of data at a given point in the data flow.

Data flows map and document relationships between data and

• Applications within a process.

• Data stores or databases in an environment.

• Network segments (useful for security mapping).

• Business roles, depicting which roles have responsibility for creating, updating, using, and deleting data (CRUD).

• Locations where local differences occur Data

Data Integration and Interoperability (DII) requirements table are also necessary to describe processes related to the movement and consolidation of data within and between data stores, applications, and organizations. Integration consolidates data into consistent forms, either physical or virtual. Data Interoperability is the ability for multiple systems to communicate.

Data Integration and Interoperability are central to the emerging area of Big Data management. Big Data seeks to integrate various types of data, including data structured and stored in databases, unstructured text data in documents or files, and other types of unstructured data such as audio, video, and streaming data. This integrated data can be mined, used to develop predictive models, and deployed in operational intelligence activities.

Furthermore, “Data Achieving Specifications” should be also considered: An organization shall retain its information for an appropriate time, taking into account all operational, legal, regulatory, and fiscal requirements, and those of all relevant binding authorities. Any organization shall also provide secure and appropriate disposition of information in accordance with its policies, and, applicable laws, regulations, and other binding authorities. Data that is used infrequently or not actively used may be moved to an alternate data structure or storage solution that is less costly to the organization. ETL functions can be used to transport and possibly transform the archive data to the data structures in the archive environment. Use archives to store data from applications that are being retired, as well as data from production operational systems that have not been used for a long time, to improve operational efficiency. It is critical to monitor archive technology to ensure that the data is still accessible when technology changes. Having an archive in an older structure or format unreadable by newer technology can be a risk, especially for data that is still legally required [1].

“Data Migration Specifications” should be appended to Table 6 in deliverable 5 [5], because it represents an insisting necessity: Data needs to be moved when new applications are implemented or when applications are retired or merged. This process involves the transformation of data to the format of the receiving application. Almost all application development projects involve some data migration, even if all that is involved is the population of Reference Data. Migration is not quite a one-time process, as it needs to be executed for testing phases as well as final implementation.

Data migration projects are frequently under-estimated or under-designed because programmers are told to simply move the data; they do not engage in the analysis and design activities required for data integration. When data is migrated without proper analysis, it often looks different from the data that came in through normal processing. Or the migrated data may not work with the application as anticipated. Profiling data of core operational applications will usually highlight data that has been migrated from one or more generations of previous operational systems and does not meet the standards of the data that enters the data set through the current application code. In Table 7 in deliverable 5 [5] in REQ. ID 38, it is useful to add to its description the wavelet transformation features (Haar, Daubechies, Mexican hat coefficients, etc.)and the CNN features as well for their wide use in this domain. Moreover, it is necessary to add to Table 8 in [5] some specifications concerning the independencies and relation between the features and the attributes of the dataset itself and the exhaustivity and the scope of its representation of the studied domain and all the other specifications that may lead to overfitting in the learning algorithm to take into accounts weighing these features when training the AI proposed model.

A table for “Data Sharing Specification” could be added to deliverable 5 [5], it should specify anticipated use and access to the data, restrictions on use, as well as expected service levels, including required system up times and response times. Master Data can enable the simplification of data-sharing architecture to reduce costs and risks associated with a complex environment.

Data not only represents value, but it also represents risk. Low-quality data (inaccurate, incomplete, or out-of-date) obviously represents a risk because its information is not right. But data is also risky because it can be misunderstood and misused. Organizations get the most value from the highest quality data – available, relevant, complete, accurate, consistent, timely, usable, meaningful, and understood. Accordingly, in Table 3 depicted in deliverable 5 [5], the critical element with their associated quality measures, rules, measures, and metrics (presented in the above sections) should be considered because they actually represent the core of data quality requirements presented in the quality literature and scientific publications, where they have been highlighted more than the current specifications mentioned in Table 5 in deliverable 5. Furthermore, Table 3 REQ ID. 24 should be extended and constitute an independent table, because Data privacy, confidentiality, and security are vital aspects of health data. It should consider data threats and internal and external data risks specifications, and the “Risk Classification”: Risk classifications describe the sensitivity of the data and the likelihood that it might be sought after for malicious purposes. Classifications are used to determine who (i.e., people in which roles) can access the data. The highest security classification of any datum within a user entitlement determines the security classification of the entire aggregation. Example classifications according to DAMA [1] include Critical Risk Data (CRD), High-Risk Data (HRD), and Moderate Risk Data (MRD).

In addition to data descriptive techniques, data modes, and data distribution presented in Table 5 in deliverable 5\_4 [5] to describe data visualization point-of-view, data scaling specification and the relationships among the represented data should be explicitly defined. Charts and graphs can be used to present data in a misleading manner. For instance, changing the scale can make a trend line look better or worse. Leaving data points out, comparing two facts without clarifying their relationship, or ignoring accepted visual conventions (such as that the numbers in a pie chart representing percentages must add up to 100 and only 100), can also be used to trick people into interpreting visualizations in ways that are not supported by the data itself.

A considerable number of deceiving and misleading visuals have been spread on the internet at the beginning of corona pandemic. Alberto Cairo [8] and Edward Tufte [9] deeply studied data visualization requirements in their publications and research. Michigan University teaches their findings in its curriculums concerning data visualization and representation.

None of the aforementioned specifications should be neglected if we aim at building robust AI-based systems and solutions driven by high-quality data in order to realize transparency, sustainability, reliability, scalability, interoperability, and consistency.

The proposed contribution also covers the six key topic areas of regularity concepts (Documentation & Transparency, Risk Management, Intended Use, Data Quality, Data Privacy and Security, Engagement and collaboration) [10].

**10- Summary and Conclusion**: Big and sensitive health data requires solid management and governance in several overlapping, interrelated, and associated domains of knowledge while generating, integrating, storing, processing, sharing, and archiving data, taking into account the principles and ethics of data science and the legislation imposed by the country or within the organization. This management must take place within a unified framework that achieves homogeneity and non-conflict or redundancy in the completion of the required tasks while achieving a balance between all types of data that can be encountered in this field so that this data is usable by various parties (health-care centers, hospitals, insurance companies, pharmacies, data miners, etc.) and this data must be standardized to enable it to be shared, expanded, analyzed, and exploited in AI systems for good. Therefore, in this paper, we put the seed of a general platform for data management and governance (Fig. 8), but we must work hard and diligently to develop this framework to include open data and medical content as well.

We must also focus on expanding the explanation of managing processes, managing roles and people, and managing the tools in every field of management we conducted in this paper. In fact, data management has become an indispensable necessity to make correct decisions in machine learning that deeply depend on it. In our upcoming contributions, we will expand on these areas from the technical, administrative and knowledge points of view, and get acquainted with the best international practices to assess and exploit data.

In this contribution, we introduced the main technical and management data requirements and specifications supported by examples to build trustworthy AI-based systems that respect data ethics and principles, during the data lifecycle. The diversity of health data and its different sources, types, and methods of processing require special attention when conducting this data and using it in artificial intelligence systems and solutions.

A diagram of a health data management system

Description automatically generated with medium confidence

Fig 8. The proposed general platform

**11 - References**:

[1] DAMA International, “DAMA-DMBOK Guide - Data Management Body of Knowledge”, second edition, 2017, by Technics Publications – Jersey.

[2] High-Level Expert Group on Artificial Intelligence, “Ethical guidelines for trustworthy AI, “European Commission, Apr. 2019

[3] J. Whittlestone, R. Nyrup, “Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research, 2019

[4] I. Taleb, M. Serhani, “Big data quality framework: a holistic approach to continuous quality management”, Journal of Big Data, Springer, 2021

[5] DEL 5\_4: Training and test data specification, ITU-T FG-AI4H-R-068, 2022-2024

[6] ​DEL 1: Ethics and governance of artificial intelligence for health, ITU-T FG-AI4H-R-068, 2022

[7] Vayena E, Dzenowagis J, Langfeld M. Evolving health data ecosystem. Geneva: World Health Organization; 2016

[8] How Charts Lie - Getting Smarter about Visual Information, Alberto Cairo, W. Norton & Company, 2019

[9] The Visual Display of Quantitative Information, Edward Tufte, 2nd edition, Graphics Press, 2001

[10] DEL 2: Overview of Regulatory Concepts on Artificial Intelligence for Health, ITU-T FG-AI4H-, 2022

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_