HealthMesh: An Architectural Framework for Federated Healthcare Data Management

Aniol Bisquert[1,*], Achraf Hmimou[1], Josep Ll. Berral[1,2], Alberto Gutierrez-Torre[2] and Oscar Romero[1]

[1]*Universitat Politècnica de Catalunya, UPC-BarcelonaTech*
[2]*Barcelona Supercomputing Center*

Abstract

Recently, significant milestones have been achieved in the field of healthcare data analysis. However, alongside these accomplishments, substantial data-related challenges have emerged in the domain of big data management. Modern healthcare projects are no more dealing with a single data repository but many heterogeneous ones and must overcome data variety, privacy and governance issues. Yet, current solutions face a privacy-decentralization trade-off. To address this dual challenge, we introduce HealthMesh, a novel layered architectural framework based on the Data Mesh principles, providing a domain-decentralised paradigm. In addition, the framework incorporates a Semantic Data Model which establishes robust governance, enables interoperability and guarantees policy compliance for all the data assets. To demonstrate the capabilities of the proposed approach, we provide an illustrative example inspired by the use case of the INCISIVE project for breast cancer analytics. Overall, this work makes a significant contribution on collecting key challenges, identifying actors and providing a set of components and guidelines for establishing a holistic framework for the complex field of healthcare data management.

Keywords

Architectural framework, Healthcare, Data Mesh, Data Governance, Data Management, Federated data

1. Introduction

In recent years, the healthcare landscape underwent a remarkable transformation driven by the digitalisation of a wealth of health-related data and the advent of Big Data analytics (e.g., machine learning -ML- technologies). These developments have opened the path for novel data-driven techniques where the incorporation of ML tools has enabled a shift from subjective interpretations to a more objective and accurate approach in diagnostics and treatment [1]. A representative example of this approach is the INCISIVE project [2], a major European initiative [3] that aims to create an interoperable federated pan-European data repository with secure data sharing and distributed analytical capabilities related to the diagnosis, prediction, and monitoring of cancer[1].

However, the rapid advancement of such initiatives in the healthcare domain often outpaces the concurrent development of the data management infrastructure. This imbalance between the progress in data analytics and data management is due to several multifaceted challenges, which collectively compromise the efficient development of (big) data-driven solutions in healthcare [4, 5].

Healthcare data management, like any other data management system, requires means to ingest, store, process and analyze data. However, the specificities of this domain have been traditionally ignored in the general field of data management, but they naturally raise new challenges when tackling projects such as INCISIVE, which we summarize as follows:

Federated data management for healthcare is a must since these projects require minimizing centralized approaches. Indeed, it is not acceptable to build a single (even if distributed) management system since **data governance** would then be centralized. Instead, data assets (i) are often distributed across various providers (typically in different medical centres or even distributed in the same medical centre), making it difficult to access and share critical information, and (ii) come with a strong sense of **data ownership** from health institutions [6] that want to decide what piece of data is shared with the federation. For this reason, distribution alone is not a solution and, instead, a federated data governance protocol should be defined to provide clear guidelines to enable healthcare organizations to harness data effectively [4].

Data privacy and security compliance present huge obstacles for researchers in this field. [7] reviews privacy preservation methods used in healthcare, including encryption and anonymization, pointing their limitations.

DOLAP 2024: 26th International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data, co-located with EDBT/ICDT 2024, March 25, 2024, Paestum, Italy

[*] Corresponding author.

aniol.bisquert@upc.edu (A. Bisquert); achraf.hmimou@upc.edu (A. Hmimou); josep.ll.berral@upc.edu (J. Ll. Berral); alberto.gutierrez@bsc.es (A. Gutierrez-Torre); oscar.romero@upc.edu (O. Romero)
0009-0001-8246-3767 (A. Bisquert); 0009-0004-0448-3406 (A. Hmimou); 0000-0003-3037-3580 (J. Ll. Berral); 0000-0002-5548-3359 (A. Gutierrez-Torre); 0000-0001-6350-8328 (O. Romero)

CEUR Workshop Proceedings (CEUR-WS.org)

[1] https://incisive-project.eu/

In addition, it is typically ignored that privacy and security principles also demand that data computations must be executed locally and data cannot be moved from where it resides. In this context, **Federated Learning** (FL) is a promising solution for this problem [8]. However, within the healthcare domain, a lack of consensus on global privacy policies for enhancing data sharing has a detrimental impact on research studies [9].

Data variety is a major challenge in healthcare data management since this domain encompasses data related to diagnosis, testing, monitoring, treatment, and health data stored in heterogeneous storage systems, often in varying **standards** and formats [10]. Variety may refer to (i) standards and format-related issues since healthcare data is typically produced following standards. [11] reviews predominant standards including openEHR, ISO13606, HL7, DICOM, etc. that are serialized following specific formats (e.g., JSON, CSV). Also, it may refer to (ii) hardware-specific issues, since a fair portion of health data is generated by medical equipment (e.g., CT scan, X-ray, ventilators, etc.). The use of diverse models of equipment can introduce bias to measurements, stemming from variations in manufacturing origins, the utilization of specific methods for image production, and differences in scans, such as varying backgrounds (e.g., black vs. white) or alterations in image contrast. In addition, addressing variety in healthcare also requires precise domain interpretation which implies that data must be interoperable at the **semantic level** [11]. Otherwise, it may compromise the quality of care provided to patients and waste resources [12].

Despite the relevance of these challenges, current state-of-the-art architectural solutions suffer from a **privacy-decentralization trade-off**. Solutions collecting data centrally fall short of privacy and security needs whereas current distributed solutions do not provide means for governing a federation and, therefore, lacking a federated governance model and interoperability framework.

To cover the above-mentioned challenges, we present HealthMesh, an innovative architectural framework for federated healthcare data management. HealthMesh is grounded on the principles of **Data Mesh** [13], advocating for the decentralization of heterogeneous data assets into autonomous and independent units referred to as "**data products**" that can execute code locally and share results with the federation. Simultaneously, this approach fosters data ownership and accessibility due to a domain-decentralized organization: i.e., data products are categorized into domains and associated with consensus-driven policies (constraints of use) and available analytical services. HealthMesh also incorporates a data governance layer responsible for managing these data products towards the establishment of a dynamic yet robust federated big data ecosystem.

The HealthMesh framework comprises three layers that all together provide answers to the challenges above identified: The Data Product Layer, the Federated Computational Governance Layer and the Data Platform Layer. The Data Product Layer defines all the data products (i.e., the data assets) federated into the system. The Federated Computational Governance Layer contains the artefacts to manage and govern all the data products. Finally, the Data Platform Layer acts as a gateway to utilize all the data management processes and workflows provided by the system, such as registering a new data source or performing an analytical study over selected data products.

At the core of the Federated Computational Governance Layer lies a **Semantic Data Model**, which captures the relationships between data products and the prescriptive guidelines established by the **consortium** (i.e., the governing body of the resulting federated system) together with semantic metadata. These guidelines have a computational nature, serving to validate the integrity and compliance of data products with the consortium agreements. The model portrays the system's complexity and facilitates the automation, integration, discovery and governance of data products while guaranteeing compliance with the policies defined. Further, analytical pipelines that need to adhere to specific policies can also be represented in the semantic model to govern analytical studies as well.

Contributions. HealthMesh is a novel architectural framework for federated healthcare data management with the following contributions: (i) it introduces a domain decentralized paradigm, grounded on the data mesh concept, granting autonomy and data ownership of federated data products within healthcare institutions. At its core, (ii) it incorporates a Semantic Data Management model, which governs the federated data products and guarantees their compliance with privacy and security policies set by the consortium. And (iii), this layer facilitates the discoverability of relevant data products, facilitates their integration (overcoming data variety) and triggers federated learning by means of robust governance mechanisms.

2. Related Work

Current solutions for healthcare data management fall short of covering the gaps discussed in the motivation. Specifically, they either provide a centralized approach not meeting the privacy and security requirements or they support distribution but not the creation and management of a federation. Further, the challenges introduced by data variety are not properly addressed.

Many current healthcare solutions are based on Data Warehousing architectures that prevent the unleashing of the potential of health data. [14] highlights their limitations in scalability, interoperability and privacy. As an evolution, Data Lakes rely on Cloud infrastructures [15]. However, these solutions suffer from several limitations, especially privacy concerns [16] but also the fact that they do not create a federation but a distributed data management system. Current solutions trying to address privacy concerns (e.g., [17], which discusses the Blockchain benefits and limitations in healthcare) fall into scalability and interoperability problems. Approaches discussed above are either centralized, falling short with privacy concerns (Data warehouse, Data Lake), or decentralized falling short with governance and interoperability.

In response to this dilemma, innovative architectures supported by semantic-based solutions were raised, which tackle the lack of data governance in other architectures. Specifically, data governance may be defined as *to what decisions must be made to ensure effective data management and data usage and who makes the decision (locus of accountability for data assets)* [18]. In this category, we focus on two: Data Fabric and Data Mesh.

Data Fabric [19] is defined as a collection of architectural principles as specific modules. Based on a knowledge graph (data catalog), the architecture enables working with data at the logical level instead of at the physical level through data virtualization, providing robust data governance and interoperability. However, defining and managing data by a central organization, as discussed by the authors, make it fall into privacy and security issues, following the same pattern observed in centralized approaches. Indeed, this solution, like other semantic-based solutions, does not allow the creation of a data federation.

Data Mesh [13] is a decentralized architecture built upon four fundamental principles. Firstly, "*Decentralized domain data ownership*" advocates for ensuring that those closest to the data take control. Secondly, "*Data as a product*" emphasizes the integration of data, metadata, and code as a logical unit for sharing. Thirdly, the concept of a "*self-serve data platform*"empowers data owners to manage the entire life cycle of their data products. Lastly, "*Federated Computational Governance*" establishes a model that strikes a balance between domain autonomy, global conformance, interoperability, and security within the mesh. Data Mesh advocates for the decentralization of data assets, emphasizing data ownership and team autonomy, ultimately enhancing data quality and unlocking the full potential of analytical insights [20].

The analysis of the existing literature reveals a gap in the current architectural solutions, particularly in the absence of a robust decentralized framework able to provide federated governance and privacy measures. The theoretical concept of Data Mesh is a promising alternative to properly manage all the factors previously mentioned. However, this paradigm sits at a high level of abstraction lacking concrete descriptions and definitions, which does not allow to operationalize their principles in a given project. Further, the data variety aspect, specifically in healthcare, is not considered. However, there is yet no available federated data platform covering all the problems previously discussed. For this reason, we propose HealthMesh, a novel architectural framework addressing the privacy-decentralization trade-off effectively and operationalizing it for the healthcare domain, while providing means to tackle data variety in this domain.

3. The HealthMesh Framework

HealthMesh is composed of a set of defined requirements and an architecture design which includes descriptions of the components, roles and workflows. We pay special attention to the Federated Computational Governance Layer, which sits at the core of HealthMesh.

3.1. Requirements

HealthMesh must cover the whole data life cycle following the challenges previously defined.

Functional Requirements: (i) *Data registration*: Incorporate new data assets into the big data system. (ii) *Data discovery*: Search and filter capabilities of the ingested data using metadata parameters. (iii) *Data analysis*: Ability to perform different types of analytical studies using data assets of interest.

Non-Functional Requirements: (i) *Domain-decentralization*: Data assets should be domain-decentralized meaning that they should be organized and aligned with the federation policies and analytical requirements. (ii) *Compliance*: A contract is established between a consortium and the owners of a data asset. Any federated data assets should be compliant with the contract and therefore respond to the expected behaviour agreed. (iii) *Privacy and Security*: Data must remain where it resides. Only processed results, in the form of aggregates, can be retrieved, using Federated Learning techniques. Individual data pieces should never be compromised. (iv) *Interoperability*: The infrastructure must facilitate the integration and usage of new heterogeneous data assets, regardless of the standard, format or hardware-specific issues. We refer to this as semantic interoperability among data assets.

3.2. Running example: Breast Cancer Analytics within the INCISIVE project

In this paper we will use the INCISIVE project, briefly introduced in the introduction, as a running example. One of the most crucial application areas of INCISIVE is that of **Breast Cancer Analytics**. This use case is based on [21], which introduces a comprehensive machine learning solution for mammography classification

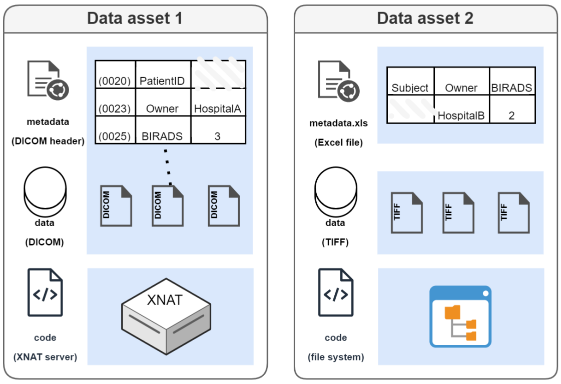

Figure 1: Running example: the Breast Cancer Analytics use case within INCISIVE. It represents two different hospitals with different file formats as starting points.

using **BIRADS** score, which is a quality control system that refers to the mammography assessment categories.

Example. Figure 1 describes *data asset 1*, owned by *Hospital A*, a XNAT server [22] with mammography images in ***DICOM*** format alongside its metadata (PatientID, owner, *BIRADS*, etc) annotated in the same file headers. Analogously, *data asset 2* (owned by *Hospital B*), is a file system with *TIFF* mammography images also with similar metadata but stored separately in an ad-hoc Excel file. In this example, in both data sets, patient identifiers have been anonymized. Also, within the same data types, there may be differences which should be treated, e.g. the difference of contrast in images due to different scanner machines (different brands, models...).

In the following, we will show how to manage and facilitate the integration of these heterogeneous data assets to enable researchers to perform a federated study to obtain a single BIRADS score classification model by using HealthMesh.

3.3. Architecture Design

HealthMesh (see Figure 2) includes three layers: the Federated Computational Governance Layer, the Data Product Layer and the Data Platform Layer. In our approach, data assets are registered and represented as data products. Data products *DP* are decentralized self-contained entities encompassing comprehensive elements, including data, metadata, and accompanying code responsible for their maintenance. Every data product must have a designated **data owner** responsible for its accessibility and maintenance. In this section we introduce the components of each layer and explain their functionalities but, due to space constraints, we focus on the most relevant ones that show the feasibility of the overall approach.

A. Federated Computational Governance Layer

The goal of this layer is to manage and govern data products. This layer is maintained by a ***Federated Team*** that provides the guidelines for all data products to be discovered, integrated and consumed. This is a multidisciplinary team consisting of domain experts. Platform, legal and analytical experts create the guidelines (constraints to guarantee when federating a data product) and features (i.e., specific analytical services) for data products in a consensus-driven way by means of the global definitions, policies and analytical services components. Healthcare institutions negotiate and establish a contract with the federated team when registering their data assets.

Global Definitions. Global Definitions *GD* provide means to enable governance and interoperability, and include the set of **Domains** *D*, **Common Data Models** *CMD* and the reference **Ontology** *ONTO*. Domains *D* are a set of healthcare disciplines given their analytical requirements, providers, etc. The federated team is responsible for the definition and evolution of domains. Consequently, every data product must be associated with at least one specific domain. Every domain has a Common Data Model *CDM* that functions as a data standard essential to enable interoperability. It sets a structure and content for the data assets. *ONTO* are vocabularies (i.e., the day-by-day terminology used by end users), typically in the form of ontology, that enable precise interpretation of data and, therefore, remove ambiguity when interpreting the data meaning.

Example. Data assets 1 and 2 are assigned to the "**Breast Cancer Analytics**" domain d_1. Moreover, the federated team agrees to use **DICOM** as Common Data Model (CDM_{DICOM}), a widely used standard for imaging data, and **SNOMED CT**[2] vocabulary ($ONTO_{SNOMED}$), one of the largest and most widely used collections of OWL vocabularies that enable sharing medical records, clinical trials, and other healthcare data [11].

Computational Catalogues. The **computational catalogues** *CC* store the Policy Checkers *C* that implements the agreed Policies *P*. All the procedures stored in the computational catalogues are defined over the *GD* previously defined to allow interoperability across heterogeneous *DP*.

Policies and Policy Compliance Checkers. **Policies** $\rho \in P$ are defined by the federated team and embody the different guidelines that data products must be compliant with. Regulatory experts within the domains

[2] https://www.snomed.org/

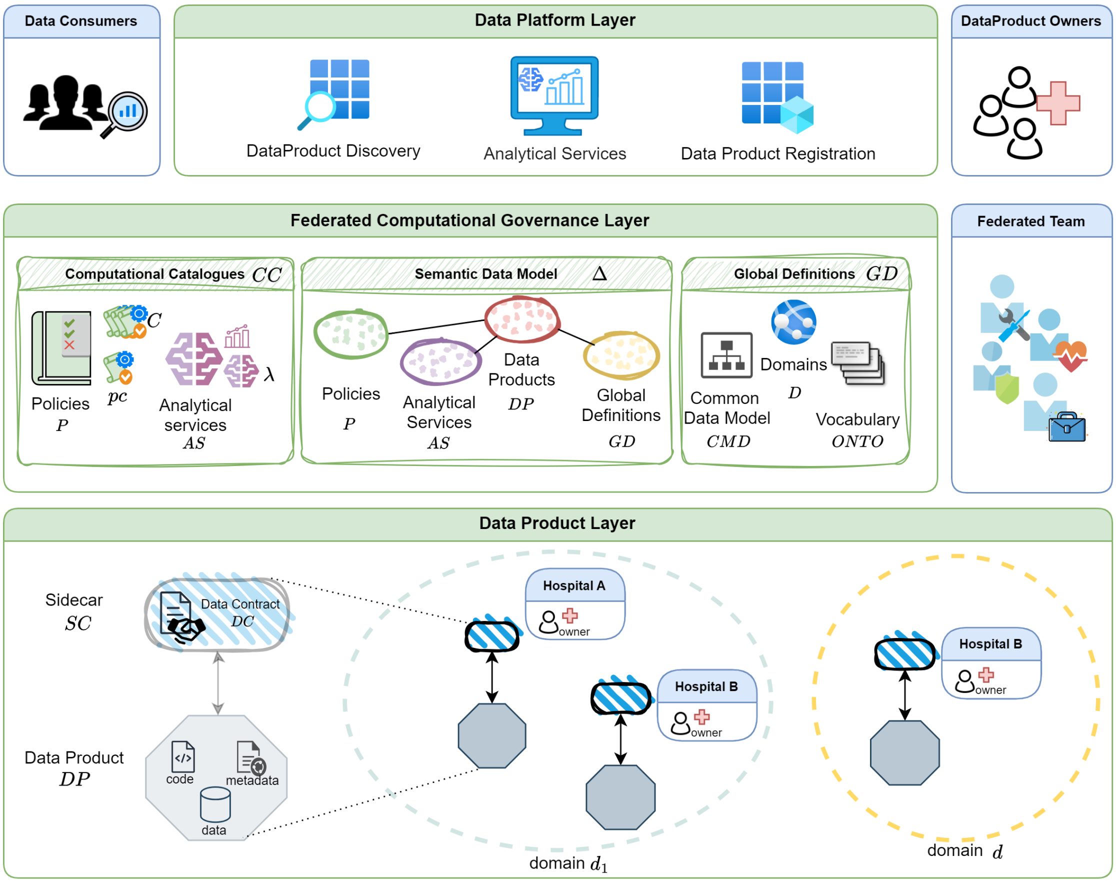

Figure 2: Overview of HealthMesh architecture components

come together and agree that all the related data must be compliant with relevant laws, regulations, and industry standards related to the handling, processing, and storage of data. Similarly, domain-specific policies are defined to validate data integrity within its context.

Policy compliance checkers $pc \in C$ are computational resources implementing Policies per domain. In HealthMesh we implement them as test functions to be executed on data products. Thus, pc is shipped and executed on each data product and, if a data product fails to meet a specific agreed policy for a given domain, that data product is not available for exploitation.

Analytical Services. Analytical experts in the federated team establish and develop a series of analytical services (AS), designed to operate on the data products within the system, generating aggregated results and comprehensive reports. An analytical service λ is related to a specific domain d and is tailored to the specific typology of data and the set of policies that the data products agreed to adhere (ρ) to fulfil the analytical requirements.

Example. In our running example, BIRADS mammogram classification λ_1 is defined by the analytical team within the domain d_1. Legal representatives in the federated team define that data consumed by λ_1 should be compliant with ρ_1, which states that personal data must be collected, processed, and stored in compliance with privacy regulations such as GDPR, CCPA, HIPAA, etc. In this context, policy checker pc_1 is the computational function that addresses ρ_1 ensuring compliance with a specific typology of data (e.g., there should not be any personal name or identifiable data). Similarly, policy ρ_2 which has been specifically defined for this service, states that all mammogram image data should be annotated with BIRADS score using DICOM headers.

Semantic Data Model. Data governance is an essential requirement for the proposed architecture. This component orchestrates all the components previously in-

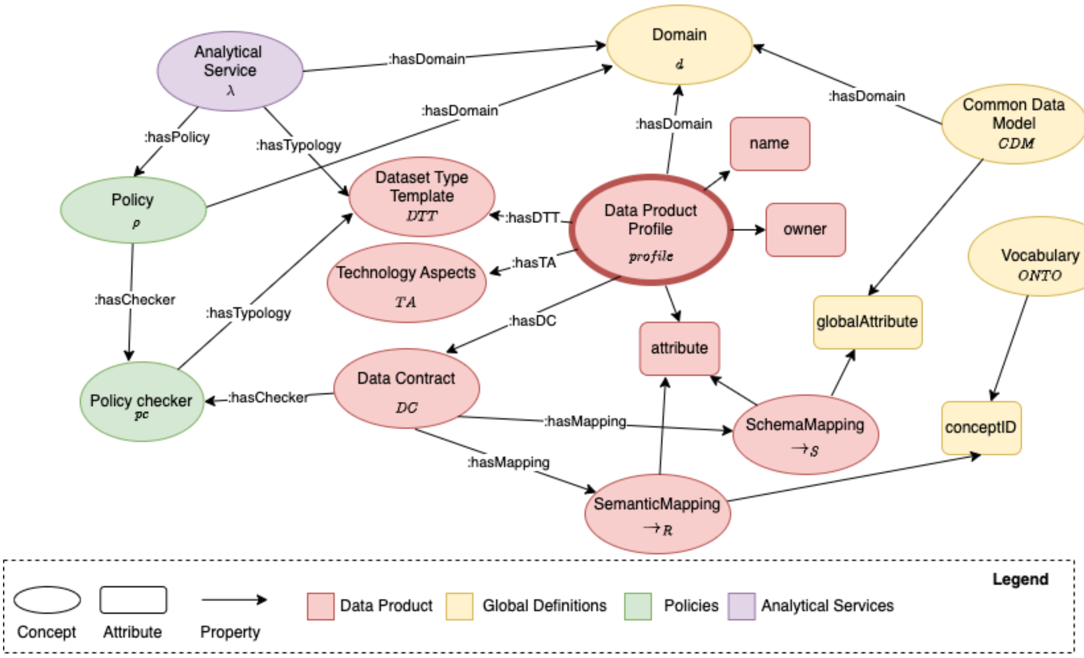

Figure 3: Semantic Data Model: Represented as a RDFS TBOX (terminology) model

troduced and defines the metadata needed to describe and govern the data assets. Thus, *DP* in each domain must be mapped to its specific data standard (i.e., *CMD*) and the vocabulary (i.e., *ONTO*). Similarly, the data product should adhere to its policies agreed (*P*) and could be eligible for the analytical services (*AS*) defined for that domain. All this metadata is described utilizing a knowledge graph.

Further, the semantic data model Δ (Figure 3) establishes the relation between the Data Product metadata provided by their owners and the guidelines defined by the federated team to enhance the integration, governance, and discoverability of data products. The goal of this model is to provide a consistent semantic model across the entire framework. Δ is a **knowledge graph**, leveraging its capacity to offer a holistic and interconnected perspective of data. Knowledge graphs are a good choice because they are flexible, heterogenous, intuitive, formal and scalable [19]. Further, several previous works have discussed the relevance of knowledge graphs to tackle governance in Big Data scenarios (e.g., [23]). From a logical perspective, A data product (*DP*) can be represented as a set of $< profile, DTT, TA, DC >$ containing a **Profile** (*profile*), **Dataset Type Template** (*DTT*), **Technology Aspects** (*TA*) and a **Data Contract** (*DC*). All this metadata is defined at the time of the data product registration process. *profile* includes all the metadata related to the data asset, including its schema (i.e., list of attributes), owner, version, etc. This is unique to each data asset. *DTT* specifies the typology of the data asset to properly categorize the data product. *DTT* contains information about the data product format (e.g., text, annotated images, etc.). The same *DTT* can be used in various domains. *TA* contains all information to grant authorization and access to data from a technological point of view. *TA* includes the data access layer credentials and data repository (e.g., access URL) metadata. *DC* acts as an agreement between data providers and the federated team. It maps the data profile schema to the Federated Computational Governance layer to facilitate data integration. The *DC* definition:

$$\begin{aligned} DC = \langle C_{DC} &= \{pc_1, pc_2, ..., pc_n\}, \\ S_{DC} &= \{\rightarrow_{S_1}, \rightarrow_{S_2}, ..., \rightarrow_{S_n}\}, \\ R_{DC} &= \{\rightarrow_{R_1}, \rightarrow_{R_2}, ..., \rightarrow_{R_n}\} \rangle \end{aligned} \tag{1}$$

contains the Data Product Schema (S_{DC}) and the semantic attribute mappings (R_{DC}) to *CDM* and *ONTO*, respectively. It also contains the set of policy checkers *C* to guarantee its compliance with the domain policies *P* and compatibility with analytical services *AS*. From a data integration perspective, the *DC* maps the local data source schema (i.e., S_{DC}) to the integration schema (i.e., the *CDM* and *ONTO*). This is a direct application of the knowledge graph data federation approach presented in [23], which enables querying the data sources (i.e., the data products) via the integration schema. Without a valid data contract, a data product cannot be part of the federation.

The semantic data model Δ is the key component to guarantee that heterogeneous medical data assets can be effectively integrated, categorized, accessed, and maintained through the utilization of the resources previously defined and, from a semantic point of view, acts as an **orchestrator**. Furthermore, leveraging ontology languages such as OWL or DL-Lite family [24], the semantic data model can benefit from reasoning to validate the resulting Δ and infer additional information [25, 24].

Algorithm 1 Data product registration

```
Require: $domain, profile, TA$
  $DTT \leftarrow \Delta$.recTemplate($profile$)
  $S, R \leftarrow \Delta$.generateMappings($profile$, $GD$)
  $P' \leftarrow \Delta$.getPolicies($domain, S, R$)
  for $\rho$ in $P'$ do
    $pc \leftarrow \Delta$.getPolicyChecker($\rho$, $DTT$)
    $C'$.AddPolicychecker($pc$)
  end for
  $DC$ = <$C', S, R$>
  $\Delta$.addDPMetadata($profile$, $DTT$, $TA$, $DC$)
```

Following Algorithm 1, $profile$ and TA are provided by the data product owners and the domain assigned by the federated team. With $profile$, the semantic data model Δ determines the most suitable DTT. Based on that, and using as input the global definitions GD and profile $profile$ it semi-automatically generates the mappings S and R to CDM and $ONTO$, respectively. The policies to be followed P' are obtained using the domain D and mappings following the approach in [26]. Moreover, Δ infers the respective C' based on P' and DTT. To complete the process, all metadata that constitutes DP is integrated into Δ.

Example. In our example, both data assets are registered using Algorithm 1 into domain d_1. Therefore, as input, the data product owner must provide the profile, which for simplicity, let us consider only contains the attribute "Subject" (which stands as a patient identifier). First, HealthMesh would assign as DTT "annotated images". Then, with the help of the data owner, who must supervise the process, the system generates the mappings to GD (in this example, we defined CDM_{DICOM} as CDM and $ONTO_{SNOMED}$ as $ONTO$). Thus, DC_2 mappings: $\rightarrow_{S_2}$ ($Subject_{profile_2}, PatientID\_id_{DICOM}$) $\in S_{DC_2}$ and $\rightarrow_{R_2}$ ($Subject_{profile_2}$, SCTID:116154003$_{SNOMED}$) $\in R_{DC_2}$. In addition, policy checkers pc_1 and pc_2 are determined to apply for d_1 (by checking the policies related to that domain via Δ) and added to their respective DC.

B. Data Product Layer

Data products (DP) are self-contained entities encompassing data, metadata and code. Therefore, physical data assets are stored and maintained by participating institutions/providers. This approach promotes data ownership and autonomy and is strongly favoured by hospitals and data owners [6].

Data Product owners are responsible for the life cycle of the data product and its maintenance. Data owners are the ones closest to the data and they can understand how it should be interpreted within each domain.

Sidecar An adjunct component in the form of a sidecar (SC) facilitates seamless integration with the broader mesh ecosystem. The sidecar is installed inside the institution/provider infrastructure but it is maintained by the platform representatives of the federated team. SC can retrieve the data of a data product through the data access layer specified in TA. Each SC contains a Data Contract DC that is retrieved from Δ.

Algorithm 2 illustrates the process of consuming a DP for a specific λ. Each time a data product is consumed, SC validates its DC to verify that data adheres to the mappings and policies specified. If data products are not interoperable or compliant, comprehensive reports are given to the data product owner specifying the errors obtained during validation. This way, the integrity and compliance of the data product are always validated in run-time guaranteeing that it conforms with its most recent contract. If none of the reports has failed, λ can be executed through the Sidecar SC to process the validated DP. The sidecar returns results in the form of aggregations. Therefore, individual data is never compromised. This approach creates a robust security measure while still allowing for analytical tasks to be performed in the context of Federated Analytics.

Algorithm 2 Data product consumption

```
Require: $DC, SC, DP, \lambda$
  $mappingsReport \leftarrow$SC.validateRS($DP$, $DC.S$, $DC.R$)
  $policyReport \leftarrow$SC.validateC($DP$, $DC.C$, $\lambda$)
  if $mappingReport$ and $PolicyReport$ are valid then
    $aggResult \leftarrow SC$.executeAS($\lambda$)
  else
    return Failed reports
  end if
```

Data products configuration strives to adhere to the **FAIR** principles of data management [27]. It is characterized by a concerted emphasis on fostering data ownership and the enhancement of data quality within the domain of healthcare data.

Example. Within our ongoing case study, DP_1 and DP_2 are candidates to be consumed for analytical service

λ_1. Following 2, pc_1 would scrutinize both DP_1 and DP_2 through their *SC* for any identifiable data. Given that both data assets are anonymized, pc_1 is expected to return successful results, confirming compliance with privacy standards.

However, the report obtained through *validateRS* on DC_2 would inform a format issue indicating that DP_2 is not available in DICOM format. The report would be sent to HospitalB to state that data should transformed to DICOM to adhere to *CDM*.

Considering that DP_2 owner applies the necessary processes over the data to be compliant with its *DC*, both DP_1 and DP_2 would be technically and semantically interoperable in terms of DICOM standard and SNOMED-CT vocabulary. Moreover, the data would be anonymized and annotated with BIRADS in DICOM format. Therefore, λ_1 could be operated in both Data Products.

C. Data Platform Layer

The data platform layer functions as an **interface** encompassing various tools/services to enable data product workflows such as (i) data product registration, (ii) discovery and (iii) execution of federated analytical tasks.

Data Consumers and **Data Product owners** use the platform to perform analytical studies and manage the Data Products, respectively.

Data product registration is a process to incorporate new data assets into the system. The process is semi-automated with the supervision of a data product owner.

Data discovery requires a query (Q) provided by data consumers containing keywords and/or filters in terms of *GD*. The function leverages Δ to effectively identify the most appropriate data products.

This architectural framework is specifically designed to enable and enhance secure **analytical tasks** in the realm of Federated Analytics, including Federated Learning. Upon selection of desired data products by data consumers, a λ can be performed over the interoperable versions acquired through Algorithm 2 to generate results.

Example. In our ongoing use case, *Data asset 1* and *Data asset 2* are registered by *Hospital A* and *Hospital B* data assets owners as DP_1 and DP_2, respectively. Data consumers can use the Data Discovery interface to post a query Q_1 containing the keywords ”Breast Cancer” to list all data products related to domain d_1 such as DP_1 and DP_2 through Δ. In this context, data consumers can select λ_1 in the Analytical Service Interface to be executed over the previously discovered data products. Consumption of both $DP1$ and $DP2$ using λ_1 would provide a local classification model. The local models can be aggregated into a global model for BIRADS breast cancer classification. Notice that this process could be done iteratively until the global model converges. Furthermore, a similar procedure can be used to perform federated exploratory data analysis to better understand the underlying data.

4. Conclusions and Future Work

We presented HealthMesh, an architectural framework designed for the healthcare domain. Building upon data mesh principles, we present a design encompassing multiple layers, components and workflows that we illustrated employing a real ongoing example. HealthMesh adopts a federated approach, ensuring that data remains within healthcare institutions to uphold security and privacy. The framework strategically employs a Semantic Data Model in conjunction with computational resources to achieve data interoperability and governance. HealthMesh is a novel architectural framework in the field that has been built upon the requirements identified collaboratively with experts from INICISVE. Our work has certain limitations that we plan to address in the near future. For example, there is an absence of in-depth technical considerations due to space constraints and an experimental evaluation with real data in real scenarios. Currently, HealthMesh is a relevant step in the right direction collecting concepts of relevance, their relationships, and the identification of key actors, which is a key contribution in the complex and limited field of federated data management for healthcare.

The development of HealthMesh opens the door for future work. For example, to study how blockchain can be integrated into the framework, the potential of Graph Neural Networks leveraging the Semantic Data Model, etc. Last, but not least, we also plan to explore the feasibility of generalizing this solution to other domains requiring a data federation (e.g., Data Spaces).

Acknowledgments

The work reported here was supported by the EU-H2020 programme under GA.952179 INCISIVE, Horizon Europe Programme under GA.101095717 (SECURED) and GA.101135513 (CYCLOPS), the Spanish Research State Agency (MICINN/AEI, ERDF/FEDER) under GA.MCIN/AEI/10.13039/501100011033/FEDER (UE DALEST), the Spanish Ministerio de Ciencia e Innovación under project PID2020-117191RB-I00 / AEI/10.13039/501100011033 (DOGO4ML), and the Generalitat de Catalunya (AGAUR) 2021-SGR-00478 (CROMAI).

References

[1] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Q. Dong, H. Shen, Y. Wang, Artificial intelligence in healthcare: past, present and future, Stroke and Vascular Neurology 2 (2017) 230–243. doi:10.1136/svn-2017-000101.

[2] I. Lazic, F. Agullo, S. Ausso, B. Alves, C. Barelle, J. L. Berral, P. Bizopoulos, O. Bunduc, I. Chouvarda, D. Dominguez, D. Filos, A. Gutierrez-Torre, I. Hesso, N. Jakovljević, R. Kayyali, M. Kogut-Czarkowska, A. Kosvyra, A. Lalas, M. Lavdaniti, T. Loncar-Turukalo, S. Martinez-Alabart, N. Michas, S. Nabhani-Gebara, A. Raptopoulos, Y. Roussakis, E. Stalika, C. Symvoulidis, O. Tsave, K. Votis, A. Charalambous, The holistic perspective of the INCISIVE project—artificial intelligence in screening mammography, Applied Sciences 2022, Vol. 12, Page 8755 12 (2022) 8755. doi:10.3390/APP12178755.

[3] H. Kondylakis, V. Kalokyri, S. Sfakianakis, K. Marias, M. Tsiknakis, A. Jimenez-Pastor, E. Camacho-Ramos, I. Blanquer, J. D. Segrelles, S. López-Huguet, C. Barelle, M. Kogut-Czarkowska, G. Tsakou, N. Siopis, Z. Sakellariou, P. Bizopoulos, V. Drossou, A. Lalas, K. Votis, P. Mallol, L. Marti-Bonmati, L. C. Alberich, K. Seymour, S. Boucher, E. Ciarrocchi, L. Fromont, J. Rambla, A. Harms, A. Gutierrez, M. P. Starmans, F. Prior, J. L. Gelpi, K. Lekadir, Data infrastructures for ai in medical imaging: a report on the experiences of five EU projects, European radiology experimental 7 (2023). doi:10.1186/S41747-023-00336-X.

[4] X. Wang, C. Williams, Z. H. Liu, J. Croghan, Big data management challenges in health research—a literature review, Briefings in Bioinformatics 20 (2019) 156–167. doi:10.1093/BIB/BBX086.

[5] S. Dash, S. K. Shakyawar, M. Sharma, S. Kaushik, Big data in healthcare: management, analysis and future prospects, Journal of Big Data 6 (2019) 1–25. doi:10.1186/S40537-019-0217-0/FIGURES/6.

[6] T. Hulsen, S. S. Jamuar, A. R. Moody, J. H. Karnes, O. Varga, S. Hedensted, R. Spreafico, D. A. Hafler, E. F. McKinney, From big data to precision medicine, Frontiers in Medicine (Lausanne) 6 (2019) 34. doi:10.3389/fmed.2019.00034.

[7] K. Abouelmehdi, A. Beni-Hessane, H. Khaloufi, Big healthcare data: preserving security and privacy, Journal of Big Data 5 (2018) 1–18. doi:10.1186/S40537-017-0110-7/TABLES/5.

[8] N. Rieke, J. Hancox, W. Li, F. Milletarì, H. R. Roth, S. Albarqouni, S. Bakas, M. N. Galtier, B. A. Landman, K. Maier-Hein, S. Ourselin, M. Sheller, R. M. Summers, A. Trask, D. Xu, M. Baust, M. J. Cardoso, The future of digital health with federated learning, npj Digital Medicine 3 (2020). doi:10.1038/s41746-020-00323-1.

[9] Y. Flaumenhaft, O. Ben-Assuli, Personal health records, global policy and regulation review, Health Policy 122 (2018) 815–826. doi:https://doi.org/10.1016/j.healthpol.2018.05.002.

[10] A. Stylianou, M. A. Talias, Big data in healthcare: a discussion on the big challenges, Health and Technology 7 (2017) 97–107. doi:10.1007/S12553-016-0152-4/FIGURES/2.

[11] B. H. de Mello, S. J. Rigo, C. A. da Costa, R. da Rosa Righi, B. Donida, M. R. Bez, L. C. Schunke, Semantic interoperability in health records standards: a systematic literature review, Health and technology 12 (2022) 255–272. doi:10.1007/S12553-022-00639-W.

[12] A. Torab-Miandoab, T. Samad-Soltani, A. Jodati, P. Rezaei-Hachesu, Interoperability of heterogeneous health information systems: a systematic literature review, BMC Medical Informatics and Decision Making 23 (2023) 18. doi:10.1186/s12911-023-02115-5.

[13] Z. Dehghani, M. Fowler, Data Mesh: Delivering Data-driven Value at Scale, O'Reilly Media, 2022. URL: https://books.google.es/books?id=M5J5zgEACAAJ.

[14] R. D. Thantilage, N.-A. Le-Khac, M.-T. Kechadi, Healthcare data security and privacy in data warehouse architectures, Informatics in Medicine Unlocked 39 (2023) 101270. doi:https://doi.org/10.1016/j.imu.2023.101270.

[15] R. Hai, C. Koutras, C. Quix, M. Jarke, Data lakes: A survey of functions and systems, IEEE Transactions on Knowledge and Data Engineering 35 (2023) 12571–12590. doi:10.1109/tkde.2023.3270101.

[16] L. Rajabion, A. A. Shaltooki, M. Taghikhah, A. Ghasemi, A. Badfar, Healthcare big data processing mechanisms: The role of cloud computing, International Journal of Information Management 49 (2019) 271–289. doi:10.1016/J.IJINFOMGT.2019.05.017.

[17] I. Yaqoob, K. Salah, R. Jayaraman, Y. Al-Hammadi, Blockchain for healthcare data management: Opportunities, challenges, and future recommendations, Neural Computing and Applications 34 (2022). doi:10.1007/s00521-020-05519-w.

[18] S. Nadal, P. Jovanovic, B. Bilalli, O. Romero, Operationalizing and automating data governance, J. Big Data 9 (2022) 117. doi:10.1186/S40537-022-00673-5.

[19] C. C. Michael DeBellis, Livia Pinera, Interoperability Frameworks, volume 3, CRC Press, 2023. doi:10.1201/9781003310785-13.

[20] I. A. Machado, C. Costa, M. Y. Santos, Data mesh: Concepts and principles of a paradigm shift in

data architectures, Procedia Computer Science 196 (2022) 263–271. doi:10.1016/J.PROCS.2021.12.013.

[21] I. Tzortzis, S. Sykiotis, I. Rallis, N. Doulamis, An integrated framework for classifying mammograms according to birads scale and breast tissue density., in: Proceedings of the 16th International Conference on PErvasive Technologies Related to Assistive Environments, PETRA '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 728–731. doi:10.1145/3594806.3596577.

[22] D. S. Marcus, T. R. Olsen, M. Ramaratnam, R. L. Buckner, The extensible neuroimaging archive toolkit: an informatics platform for managing, exploring, and sharing neuroimaging data, Neuroinformatics 5 (2007) 11–34. doi:10.1385/ni:5:1:11, pMID: 17426351.

[23] S. Nadal, A. Abelló, O. Romero, S. Vansummeren, P. Vassiliadis, Graph-driven federated data management, IEEE Trans. Knowl. Data Eng. 35 (2023) 509–520. doi:10.1109/TKDE.2021.3077044.

[24] D. Calvanese, G. De Giacomo, D. Lembo, M. Lenzerini, A. Poggi, M. Rodriguez-Muro, R. Rosati, Ontologies and Databases: The DL-Lite Approach, Springer Berlin Heidelberg, Berlin, Heidelberg, 2009, pp. 255–356. doi:10.1007/978-3-642-03754-2_7.

[25] X. Chen, S. Jia, Y. Xiang, A review: Knowledge reasoning over knowledge graph, Expert Systems with Applications 141 (2020) 112948. doi:https://doi.org/10.1016/j.eswa.2019.112948.

[26] J. Flores, K. Rabbani, S. Nadal, C. Gómez, O. Romero, E. Jamin, S. Dasiopoulou, Incremental schema integration for data wrangling via knowledge graphs, Semantic Web – Interoperability, Usability, Applicability accepted, tbp (2023). URL: https://www.semantic-web-journal.net/content/incremental-schema-integration-data-wrangling-knowledge-graphs-0.

[27] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J. W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. t Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S. A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. V. D. Lei, E. V. Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, The FAIR guiding principles for scientific data management and stewardship, Scientific Data 2016 3:1 3 (2016) 1–9. doi:10.1038/sdata.2016.18.