2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC) | 979-8-3503-7696-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/COMPSAC61105.2024.00127

Intent-based Pseudonymization for Healthcare Workflows on Intra-Hospital Data Space Domain

Gabriele Morabito*, Armando Ruggeri†, Antonio Celesti†, Massimo Villari†, Maria Fazio†
*Department of Biomedical Sciences, Dental Sciences, and Morphological and Functional Imaging
University of Messina, Messina, Italy
†Department of Mathematics and Computer Sciences, Physical Sciences and Earth Sciences
University of Messina, Messina, Italy
{gamorabito,armruggeri,acelesti,mvillari,mfazio}@unime.it

***Abstract*—Hospitals suffer from implementing Data Spaces due to the risks related to data security aspects. To ensure patients' data privacy, healthcare organizations can incorporate pseudonymization strategies into their data management practices, promoting collaboration and information sharing among several hospital departments and healthcare professionals. In this paper, we defined and implemented the intent-based multilevel granular approach for HL7 FHIR JSON documents pseudonymization, by comparing it with non-granular encryption of the entire document. With this approach, we enhance patient confidentiality and facilitate efficient healthcare data sharing within the Intra-Hospital Data Space, facilitating enhanced flexibility and scalability in deploying and utilizing data management systems.**

***Index Terms*—Data Space, Health Data, HL7 FHIR, Workflow, Privacy, Intent-based Pseudonymization**

I. Introduction

The emergence of Data Space has heralded a paradigm shift in the management of data within hospital environments. It facilitates the seamless integration and oversight of a wide spectrum of clinical, administrative, and operational data originating generated within the hospital ecosystem. This transformative technology holds immense potential to revolutionize healthcare delivery by providing a unified platform for data aggregation and analysis. However, amidst this digital transformation, one of the foremost challenges that healthcare institutions face pertains to the preservation of data security and privacy. Securing the confidentiality and integrity of sensitive patient data within Data Spaces is crucial for maintaining trust and compliance with regulatory standards. As a result, healthcare organizations are increasingly acknowledging the significance of deploying robust data protection protocols, employing data de-identification to protect individuals' privacy. Thanks to data de-identification techniques, it is possible to mitigate the risk of identifying individuals by making it more difficult or even impossible to directly associate data with people. Data pseudonymization emerges as a critical strategy to achieve data de-identification. In this context, our research aims to tackle the pressing requirement for enhanced data security within hospital Data Spaces through the development and implementation of a novel pseudonymization approach. In general, depending on the purposes for which data are collected, shared, or used, data governance can extend beyond intra-organizational or inter-organizational boundaries [1]. Equivalently Hospital Data Spaces can be grouped into two main categories: Intra-Hospital Data Spaces, and Inter-Hospital Data Spaces. The first ones refers to the Data Space within a single hospital or medical institution. In this context, the data is managed within the hospital's own infrastructure and computer systems and is used to provide patient care, manage hospital resources, and conduct internal administrative and clinical activities. The second one refers to Data Space involving multiple hospitals or medical institutions. In this case, data may be shared among different healthcare facilities for purposes such as coordinated patient management across multiple hospitals, sharing medical information among specialists or research institutions, exchanging epidemiological data to monitor the spread of diseases, or other research purposes. In this paper, we focus on the intra-domain because securing privacy within the hospital is paramount as it represents the initial stride towards achieving a similar level of privacy protection in inter-hospital settings. By establishing robust privacy measures within the confines of a single hospital, we lay the groundwork necessary to extend these safeguards across multiple institutions. This approach ensures that the fundamental principles of privacy protection are firmly in place before expanding to more complex inter-hospital data-sharing scenarios. Specifically, we propose an intent-based multilevel granular method tailored for pseudonymizing HL7 FHIR® (Fast Healthcare Interoperability Resources) JSON documents. This approach offers a nuanced and context-aware approach to data protection, allowing for the customization of pseudonymization strategies based on the specific intent and sensitivity of the data being processed. Additionally, our study aims to assess the effectiveness of this detailed pseudonymization method by comparing it with traditional non-granular encryption techniques applied to entire HL7 documents. By conducting a comprehensive comparative analysis, we delineate the advantages and limitations of each approach in terms of data security, privacy preservation, and operational efficiency within the hospital Data Space. Our research endeavors to contribute to the advancement of data security practices in healthcare settings. By fostering a culture of data privacy and compliance, we aspire to empower healthcare organizations to harness the full potential of Data Spaces while

2836-3795/24/$31.00 ©2024 IEEE
DOI 10.1109/COMPSAC61105.2024.00127

upholding the highest standards of patient confidentiality and trust. In this paper, we follow a Privacy by Design approach for anonymization, pseudonymization, and security rules to adhere to legal requirements and best practices for sharing clinical data [2].

The remainder of this paper is organized as follows. Section II depicts the state of the art about data pseudonymization. An overview of Intra-Hospital Data Space and workflow applications is delineated in Section III. Section IV provides a detailed analysis of health data and their sensitivity. In Section V we describe a workflow application use case on the Intra-Hospital Data Space domain. In this context, we introduce the concept of intent-based pseudonymization in Section VI. The Design of our granular pseudonymization algorithm, tailored for intent-based pseudonymization, is depicted in Section VII. The experiments conducted and the results are discussed in Section VIII. Finally, in Section IX, we draw the conclusions and discuss possible future works.

II. State of the art

Pseudonymization is a fundamental practice in protecting the privacy of personal data. It is based on the idea of replacing direct identifying information with unique identifiers or pseudonyms, which do not allow individuals to be directly identified without the use of a relinking key. This process makes it nearly impossible for third parties to directly associate data with specific individuals without consent or authorized access to the relinking key.

The data itself is depersonalized through the use of pseudonyms, while the reconnection key is kept securely and accessible only to authorized individuals. This helps ensure that even if data is compromised or disclosed, individuals' privacy will not be compromised, as the data cannot be directly linked back to individuals without access to the key.

The idea of pseudonymization started at the beginning of the century when the first RESTful web applications required a new approach to exchange data in an unsecured channel [3], and gained increased importance after the introduction of GDPR (General Data Protection Regulation) and especially the exponential growth of big data [4], which led to the review of the legal and technological aspects, introducing the European Health Data Space [5].

Among various pseudonymization techniques, they generally consist of replacing the sensitive data with identifiers, and storing the original data elsewhere for the re-identification process [6]. Alternatively, encryption [7] and tokenization [8] are other techniques that do not need to store the original data as plaintext elsewhere.

With a particular focus on the healthcare scenario, the utilization of clinical data for research purposes should uphold data confidentiality, and patient privacy rights, and obtain patient consent. Clinical data inherently contains personal information such as name, age, and identification number, thus making it nominative. Consequently, clinical data must undergo de-identification before being transferred to research databases. Concerning the privacy aspects, it is fundamental to protect health records from unauthorized access, offering the patient, i.e. the data owner, the option to decide whom to disclose his/her health information to [9].

The emergence of data spaces has revolutionized the approach to data management, providing a virtual environment where information from diverse sources can be integrated and analyzed synergistically. Data integration is the primary focus within the realm of Data Space technology research. It serves as the cornerstone and central element of Data Space. Data Space itself is established through the process of data integration, and it is subsequently governed and sustained by ongoing data integration efforts [10].

In healthcare settings, Data Spaces enable comprehensive and personalized analysis of clinical data, allowing healthcare professionals to access crucial information to provide high-quality care. The integration of data from various sources, such as medical records and laboratory data, both structured and non-structured [11], facilitates early diagnosis and personalized treatment [12].

Despite the advantages, the use of Data Spaces in healthcare raises crucial privacy issues for patients [13]. The need to protect personal and sensitive information is fundamental to ensure public trust and compliance with privacy regulations.

Differently from the recent scientific initiatives mentioned, and progressing the research of previous works [14], [15], this paper proposes a granular approach for health data pseudonymization, facilitating intent-based data de-identification. This strategy enables the possibility of pseudonymizing data with precise granularity, allowing for varying levels of pseudonymization. Consequently, it becomes feasible the introduce of several rules for data re-identification, tailored to the specific intent behind the data usage.

III. Data Space Within Hospital

In this section, we provide an overview of the Data Space concept. Then we describe how the advent of Data Space influences the development of applications as workflows. Finally, we describe the scenario of the Intra-Hospital Data Space.

A. Data Space Overview

The concept of Data Space [16] refers to a virtual environment in which data is organized, managed, and made available for use by applications and end-users. It operates as a fundamental abstraction from the physical infrastructure of data storage that opens new perspectives and opportunities in the field of application development, by freeing developers from the technical and logistical limitations associated with direct data access. In other words, the Data Space is a conceptual or virtual representation of the space where data is stored, managed, and manipulated, regardless of its actual physical location [17]. This idea enables the decoupling of data management tasks from the intricate technical details of the underlying infrastructure, thereby facilitating enhanced flexibility and scalability in deploying and utilizing data management systems. This abstraction simplifies the development of distributed applications and services that require transparent

and efficient data handling, without the need to delve into the technical complexities of the underlying infrastructure. This means that applications can focus exclusively on business logic and offered functionalities, without having to dedicate significant resources to data infrastructure management. This allows for speeding up the application development cycle and reducing overall implementation and maintenance costs. The application of Data Space spans various fields such as the public sector, business, scientific research, social services [18], and healthcare [19], [20]. In each of these scenarios, the effective management of the Data Space is crucial to guarantee data accuracy, security, accessibility, and usability for their intended objectives. This could entail deploying data management systems, enforcing cybersecurity protocols, adhering to data storage and backup procedures, and ensuring compliance with privacy regulations and data protection laws [21].

B. Workflows in the Data Space Ecosystem

In the Data Space context, applications are not simple isolated tools, but rather complex workflows (Figure 1) that leverage the resources and services offered by the data space itself.

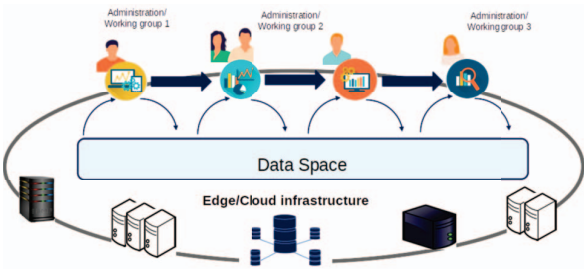


Figure 1. Data Space Workflow Application

Each step within a workflow does not require a direct connection between input and output for data transfer. Instead, interaction occurs through the unified interface of the Data Space, which acts as an intermediary between the application and the underlying data resources. This design enables applications to achieve a high degree of modularity and scalability. The individual microservices comprising the application can operate independently and communicate with the data space interface to retrieve essential data. This fosters adaptability and flexibility in creating and executing workflows, empowering developers to insert, delete, or adjust process steps without needing to reconfigure the entire system. Moreover, it's crucial to acknowledge that workflows can consist of steps of two types: those that are completely automated and those that require human interaction. This differentiation is vital as it impacts the design and implementation of workflows. While automated steps can be seamlessly integrated and executed within the system, steps requiring human interaction necessitate interfaces or mechanisms for human input and decision-making. Therefore, when designing workflows within the Data Space framework, developers must consider both automated and human-dependent steps, ensuring a cohesive and effective workflow execution. Furthermore, the use of the Data Space interface simplifies data management and ensures the consistency and integrity of the data itself. Applications can access data through a standardized interface, regardless of their location or origin, minimizing integration complexity and ensuring greater consistency in data access and usage.

C. Intra-Hospital Data Space

In the health sector, implementing a Data Space-based architecture within a hospital setting (Intra-Hospital Data Space) represents a cutting-edge approach to the management and organization of healthcare data within medical institutions. In this context, the Data Space serves as a central digital infrastructure that allows for the efficient integration and management of a wide range of clinical, administrative, and operational data generated within the hospital. By leveraging the capabilities of a Data Space, hospitals can effectively aggregate and consolidate data from a myriad of sources. These sources encompass electronic health record systems, medical devices, diagnostic equipment, patient monitoring systems, and administrative platforms. The breadth of data gathered within the Data Space encompasses vital patient information, comprehensive laboratory test results, intricate diagnostic images, detailed medical prescriptions, meticulous procedure records, appointment scheduling details, and much more. However, one of the foremost challenges in implementing Data Spaces within hospitals lies in ensuring the security and privacy of the data. Safeguarding sensitive patient information and regulating access to this data is paramount to maintaining patient confidentiality and adhering to privacy regulations. Therefore, robust security measures and access controls must be implemented to mitigate risks associated with data breaches or unauthorized access. Additionally, compliance with relevant data protection laws, such as the Health Insurance Portability and Accountability Act (HIPAA) in the United States or the General Data Protection Regulation (GDPR) in the European Union, is imperative to uphold patient privacy rights and prevent regulatory violations. The efficient management of the Data Space within hospitals is vital to guarantee data security, privacy, and accessibility and promote collaboration and information sharing among several hospital departments and healthcare professionals.

IV. Health Data Classification

Implementing Intra-Hospital Data Space necessitates a clear classification of healthcare data types and the degree of sensitivity associated with each type of data. Health data can be classified as follows:

- Patient's medical history. This encompasses details about a patient's previous and current medical conditions, past treatments, and allergic reactions.
- Results of laboratory tests. This refers to the data obtained from analyzing samples collected from patients, such as

blood, urine, tissue, or other bodily fluids. These data typically consist of numerical values, units of measurement, and reference ranges that indicate what is considered normal for a particular test.
- Results from diagnostic imaging tests. This comprises image data describing the internal structures of the human body for diagnostic or therapeutic purposes and resulting from the utilization of a broad range of medical techniques and technologies, such as X-rays, CT scans, MRIs, and ultrasound scans.
- Patient's vital parameters. This refers to measurements like the patient's blood pressure, heart rate, body temperature, and respiratory rate.
- Prescribed medications. This data usually includes information about the prescribed drugs and their respective dosages.
- Patient's genetic information. It is the patient's genetic profile, including information about the patient's genetics, such as DNA data.
- Demographic data. These are personal data that can involve a wide range of subjects within the healthcare sector, such as patients, physicians, nurses, laboratory technicians, administrators, administrative staff, physical therapy operators, and more.
- Administrative data. These data concern hospital administration, human resource management, logistics, medical equipment management, and other internal operations. These data are used to ensure efficient hospital functioning and optimize available resources.
- Financial data. This includes billing information, payments, operational costs, insurance reimbursements, and other financial transactions related to healthcare. They are fundamental for financial management and hospital administration.

As regards health data sensitivity, instead, four different sensitivity classes can be identified:

- Direct Identification. This class of sensitivity includes data that allow to directly identify a person, like name and surname.
- Indirect Identification. It includes data that, alone, may not directly reveal a person's identity but can be combined with other data to potentially identify an individual.
- Sensitive data. These are data that do not reveal the identity of someone, but, if disclosed or accessed without authorization together with identifying data, it could result in privacy violation.
- Ultra-sensitive data. This class of sensitivity includes particular data whose unauthorized disclosure could result in a more invasive privacy violation.

In Table I we map the different types of health data to the corresponding sensitivity class.

Table I
HEALTH DATA SENSITIVITY MAP

Sensitivity class	**Health data type**
Direct identification	Demographic Data - Name Demographic Data - Surname Demographic Data - Tax code
Indirect identification	Demographic Data - Date of Birth Demographic Data - Gender Demographic Data - Address Demographic Data - Phone Number Administrative Data Financial Data
Sensitive data	Results of Laboratory Tests Diagnostic Imaging Vital Parameters Prescribed Medications Medical History
Ultra-sensitive data	Genetic Information
Non sensitive data	All the other data

V. WORKFLOW USE CASE

In this section, we present a simple healthcare workflow use case within the context of applications built upon the Intra-Hospital Data Space, leveraging HL7 FHIR standards for enhanced interoperability and efficiency. Indeed HL7 FHIR, a standard for exchanging healthcare information electronically developed by Health Level Seven International, is an optimal choice for building Intra-Hospital Data Space based applications because it has the following features:

- Interoperability. It provides a standardized method for representing and exchanging healthcare data, irrespective of the systems or technologies employed. FHIR delineates a collection of resources, which are modular components representing different aspects of healthcare information, such as patients, providers, medications, observations, and so on. Additionally, they are represented through contemporary web standards like JSON and XML.
- Flexibility. It facilitates the structured and standardized representation of intricate clinical concepts and relationships. FHIR promotes a resource-oriented approach, where each resource signifies a discrete element of data or functionality. This modular design enables healthcare organizations to adopt FHIR incrementally, seamlessly integrating it into existing systems and workflows without the need for a complete overhaul.
- Scalability. HL7 FHIR is designed to support scalability, enabling the efficient exchange of healthcare data across large-scale systems and networks. It capitalizes on contemporary web technologies and adheres to RESTful principles, enabling lightweight, stateless communication between clients and servers. This architecture is particularly well-suited for distributed systems, capable of accommodating the escalating volume and complexity of healthcare data.
- Future-proofing. Through the adoption of HL7 FHIR standards, healthcare organizations can safeguard their systems and infrastructure for the future, guaranteeing compatibility with forthcoming technologies and evolving regulatory standards. Indeed, FHIR is designed to be extensible, allowing for the addition of new resources and capabilities to address the evolving needs of healthcare

over time.

The workflow (Figure 2) we present is composed of five steps:

1) Admission. The initial step in the workflow involves the admission of a patient to the healthcare system. Upon admission, relevant patient information such as demographics, insurance details, and initial complaints are collected and recorded. This step ensures the seamless integration of patient data into the system, facilitating continuity of care and interoperability across different healthcare entities.
2) Medical Examination. In this step, the patient undergoes a medical examination wherein various diagnostic tests, procedures, and observations are conducted. Data generated during the medical examination, such as vital signs or clinical notes, are immediately captured and stored in the Data Space in HL7 FHIR format since they are already available. Other data, such as laboratory results or imaging studies, may need more time before to be available. After the examination, the patient leaves the structure.
3) Reporting. In this phase of the workflow, an automated AI-based specialized software, which can automatically elaborate the data collected, produces the diagnostic reports containing all the results related to the conducted medical examination.
4) Validation. The generated clinical data undergoes rigorous validation by the doctor to ensure accuracy and completeness. Through the validation processes, discrepancies and errors within the data are identified and, eventually, rectified.
5) Visualization. The patient is notified that all the results are available and gain access to tools allowing them to visualize their own healthcare data, empowering them to better understand and engage with their health information.

During each one of the five workflow steps, seamless interaction with the Data Space is paramount. This interaction involves retrieving data required as input for the operations within each step and potentially saving the output data back into the Data Space.

VI. Intent-based Pseudonymization

Pseudonymization refers to a method of anonymizing data, where identifiable information in datasets is substituted with artificial identifiers or pseudonyms. This approach enables the de-identification of sensitive data while preserving its usefulness. In contrast to anonymization, which permanently obscures identifiable information, pseudonymization is a reversible process where the original data can be restored as needed. This characteristic adds an extra layer of flexibility and utility to pseudonymized data, as it allows for the restoration of identifiable information if required for specific purposes. Pseudonymization is indispensable to responsibly handle health data and comply with regulations governing data privacy and security. It serves as a fundamental tool for achieving compliance by reducing the risk of re-identification while still allowing for legitimate data use cases in healthcare settings. Therefore, it is imperative for healthcare organizations to incorporate pseudonymization into their data management practices to protect patient privacy rights and mitigate regulatory risks. In the context of Intra-Hospital Data Spaces, where applications are comprised of interconnected workflows interfacing with the Data Space for data input and output, the complexities of data pseudonymization become more intricate. Within these workflows, each consisting of multiple interconnected steps, users from different roles within the hospital organization can be involved. Unlike a binary approach where data pseudonymization is applied depending solely on the sensitivity of the data itself, pseudonymization within this context requires different levels of granularity. The decision of how to apply pseudonymization varies depending on strictly interconnected factors such as the user's role within the hospital and the intent behind the specific workflow step, which influences the required input and output data. For instance, let us consider the contrast between personnel working in admissions and medical professionals within a hospital setting. In the admissions department, staff members primarily require access to demographic information for patient registration purposes. They typically handle data such as names, addresses, and contact details but do not need access to detailed medical records. On the other hand, medical professionals, such as doctors, require access to a more comprehensive set of data to provide effective patient care. While they may need access to some partial demographic information, their primary focus lies in accessing relevant medical records for diagnosis and treatment planning. Moreover, some phases of a healthcare workflow could be automated through some software (for example, the *Reporting* phase of the workflow described in Figure 2), which does not require human intervention to retrieve or store data from and to the Data Space. In that case, the re-identification of data, during the data retrieval operation, strictly depends solely on the workflow step intent.By tailoring the level of pseudonymization to match the specific requirements of each workflow step intent itself and, when human interaction is required, of each user role within the specific workflow step, hospitals can achieve a delicate balance between data privacy and usability. This approach not only enhances patient confidentiality but also facilitates efficient healthcare delivery within the Intra-Hospital Data Space. As shown in Figure 2, different users or software interact with the system at different stages and need to access the corresponding data. Moreover, as already depicted above, privacy needs strictly depend on both the user who accesses the data and the intent of the specific operation. For this reason, all the data that are stored in the Data Space needs to be totally pseudonymized when saved and partially recovered when accessed. Therefore, although the HL7 documents must be fully obfuscated, the pseudonymization process must be granular and each field of the documents must be processed independently. This approach enables partial recovery of in-

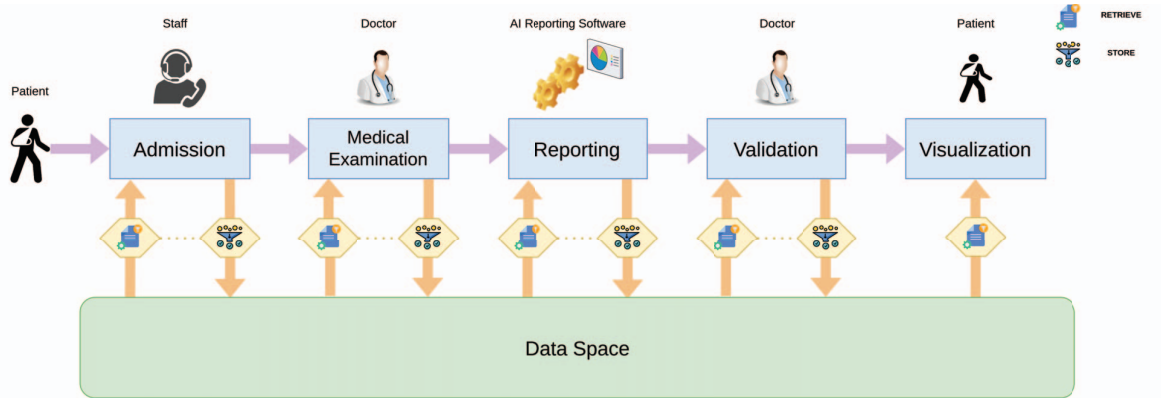


Figure 2. Health Workflow Use Case

formation based on the user accessing it and the intent of the corresponding operation. This approach requires a detailed analysis of the HL7 FHIR resources involved in each phase of the workflow. For example, concerning the workflow described in Section V, we identified which resources are retrieved, processed, eventually created, and stored at each workflow step. Moreover, we established the fields to be re-identified for each retrieved resource, implementing a strategy to ensure data protection both during data retrieval and storage. Table II summarizes the mapping of HL7 FHIR resources and fields across each workflow step. We distinguish between various instances of resources of the same type. To accomplish this, we utilized the *_old* notation to denote preexisting HL7 FHIR resources. On the contrary, the *_new* notation indicates the resources that are generated during the workflow execution. More detailed information about the HL7 FHIR resource types can be found on the HL7 FHIR website[1].

Table II
HL7 FHIR RESOURCE MAPPING

WORKFLOW STEP	RETRIEVE OPERATION		STORE OPERATION
	Resource	**Fields**	
Admission	Patient	All fields	Patient
			Encounter
Medical Examination	Patient	Name	Observations_new
		Birthdate	
		Gender	
	Observations_old	All fields	
	Conditions_old	All fields	
	DiagnosticReports_old	All fields	
	Practictioners	Name	
Reporting	Patient	Id	DiagnosticReport_new
	Observations_new	All Fields	
Validation	Patient	Name	Encounter
		Birthdate	
		Gender	Condition_new
	Conditions_old	All fields	
	DiagnosticReports_old	All fields	DiagnosticReport_new
Visualization	Conditions_old	All fields	None
	Conditions_new	All fields	
	DiagnosticReports_old	All fields	
	DiagnosticReports_new	All fields	

VII. ALGORITHM DESIGN

Once we mapped the data to the workflow steps, we developed two algorithms, for granular pseudonymization (Algorithm 1) and re-identification (Algorithm 2) respectively. These algorithms are devised to operate in a complementary fashion, ensuring both the pseudonymization of sensitive data and its subsequent re-identification as necessary. Algorithm 1 provides a method for pseudonymizing sensitive data while retaining its structure and granularity. This algorithm takes in input the plaintext data and a set of labels indicating the portions of the data to be pseudonymized. The algorithm operates recursively, traversing the data structure to pseudonymize individual elements based on their types. Moreover, it uses the generic *encrypt()* function to effectively pseudonymize the data. This function can be implemented by applying any encryption technique. The variable *p_data*, which will eventually contain the pseudonymized data, is initialized with the plaintext data (line 1). If the data is an object, it means that it is necessary to search for the labels to pseudonymize (lines 2-4). Instead, if the data is a list (line 14), it means that a label has already been found and the algorithm has already been called recursively, therefore all the elements need to be pseudonymized (line 15). In both cases, the following strategy is adopted to pseudonymize, depending on the data type:

- if the portion of data to which pseudonymization is applied is an object (line 5 and line 16), the pseudonymization algorithm is applied recursively to the object, and all the object keys are given in input to the algorithm as labels (line 6 and line 17);

[1]https://build.fhir.org/resourcelist.html

- if the portion of data to which pseudonymization is applied is a list (line 7 and line 18), the pseudonymization algorithm is applied recursively to the list, and the same labels are given in input to the algorithm as labels (line 8 and line 19);
- if the portion of data to which pseudonymization is applied is a string (line 9 and line 20), the data is encrypted.

Algorithm 1 P_FUNC(data, labels)

Require: $data$ {The plaintext data}
Require: $labels$ {The labels referring to the portions of data to pseudonymize}
1: $p\_data \leftarrow data$
2: **if** $data$ is object **then**
3: **for all** k in $data.keys$ **do**
4: **if** $k \in labels$ **then**
5: **if** $data[k]$ is object **then**
6: $p\_data[k] \leftarrow P\_FUNC(data[k], data.keys)$
7: **else if** $data[k]$ is list **then**
8: $p\_data[k] \leftarrow P\_FUNC(data[k], labels)$
9: **else if** $data[k]$ is string **then**
10: $p\_data[k] \leftarrow encrypt(data[k])$
11: **end if**
12: **end if**
13: **end for**
14: **else if** $data$ is list **then**
15: **for** i in $0, ..., data.length$ **do**
16: **if** $data[i]$ is object **then**
17: $p\_data[i] \leftarrow P\_FUNC(data[i], data[i].keys)$
18: **else if** $data[i]$ is list **then**
19: $p\_data[i] \leftarrow P\_FUNC(data[i], labels)$
20: **else if** $data[i]$ is string **then**
21: $p\_data[i] \leftarrow encrypt(data[i])$
22: **end if**
23: **end for**
24: **end if**
25: **return** p_data

Algorithm 2, instead, provides a method to re-identify portions of data. It takes in input the pseudonymized data and a set of labels indicating the portions of the data to be re-identified. It operates in the same way as Algorithm 2, but instead of encryption it applies a generic decryption function (line 10 and line 21) to re-identify the pseudonymized data.

VIII. Experiments

In this section, we detail the experiments conducted to assess the performance of our intent-based pseudonymization system. We begin by outlining the testbed setup, detailing the hardware and software configurations utilized. Subsequently, we present the results obtained from our experiments and evaluate their implications.

Algorithm 2 I_FUNC(p_data, labels)

Require: p_data {The pseudonymized data}
Require: $labels$ {The labels referring to the portions of data to re-identify}
1: $i\_data \leftarrow p\_data$
2: **if** p_data is object **then**
3: **for all** k in $p\_data.keys$ **do**
4: **if** $k \in labels$ **then**
5: **if** $p\_data[k]$ is object **then**
6: $i\_data[k] \leftarrow I\_FUNC(p\_data[k], p\_data.keys)$
7: **else if** $p\_data[k]$ is list **then**
8: $i\_data[k] \leftarrow I\_FUNC(p\_data[k], labels)$
9: **else if** $p\_data[k]$ is string **then**
10: $i\_data[k] \leftarrow decrypt(p\_data[k])$
11: **end if**
12: **end if**
13: **end for**
14: **else if** p_data is list **then**
15: **for** i in $0, ..., p\_data.length$ **do**
16: **if** $p\_data[i]$ is object **then**
17: $i\_data[i] \leftarrow I\_FUNC(p\_data[i], p\_data[i].keys)$
18: **else if** $p\_data[i]$ is list **then**
19: $i\_data[i] \leftarrow I\_FUNC(p\_data[i], labels)$
20: **else if** $p\_data[i]$ is string **then**
21: $i\_data[i] \leftarrow decrypt(p\_data[i])$
22: **end if**
23: **end for**
24: **end if**
25: **return** i_data

A. Testbed Setup

We start by detailing the hardware components utilized in our test environment. We adopted 9 Raspberry Pi 4 Model B to simulate the distributed infrastructure . Those devices are suitable for installation in a hospital setting for several reasons. From a practical and economic standpoint, these devices offer many advantages. In fact, their compact size makes them easy to deploy in a hospital setting, enabling flexibility in placement and installation. Furthermore, they are a cost-effective option, compared to other hardware solutions. At the same time, from the functionality point of view, Raspberry Pi can be equipped with community-supported open-source operating systems and software, being a flexible and customizable solution that allows it to meet specific needs within hospital environments. Each Raspberry Pi is equipped with 4GB LPDDR4-3200 SDRAM, a Broadcom BCM2711 Quad core Cortex-A72 (ARM v8) 64-bit SoC @ 1.8GHz, and Micro-SD card slot for loading operating system and data storage. In particular, for our test purposes, we used 64GB SD cards. As regards the software, instead, we flashed the SD cards with the Raspberry Pi OS. In our implementation, to achieve pseudonymization purposes, we chose to employ Secret Sharing [22] techniques. There are several reasons for our choice. First of all, Secret Sharing is a well-known approach to store

Table III
REST SERVER ENDPOINTS

Endpoint URI	HTTP Method(s)	Description	Parameters	Response
/split	POST	Pseudonymize an HL7 FHIR document by splitting it across the available nodes through Secret Sharing	JSON object (HL7 FHIR document and Secret Sharing configuration parameters)	Status code
/save	POST	Locally save a share	chunkRef (path)	Status code
/doc	GET	Retrieve a locally saved share	shareId	JSON object (share)
/retrieve	POST	Re-identify an HL7 FHIR document by retrieving and merging the corresponding distributed shares	JSON object (documentId and Secret Sharing configuration parameters)	JSON object (Re-identified HL7 FHIR document)

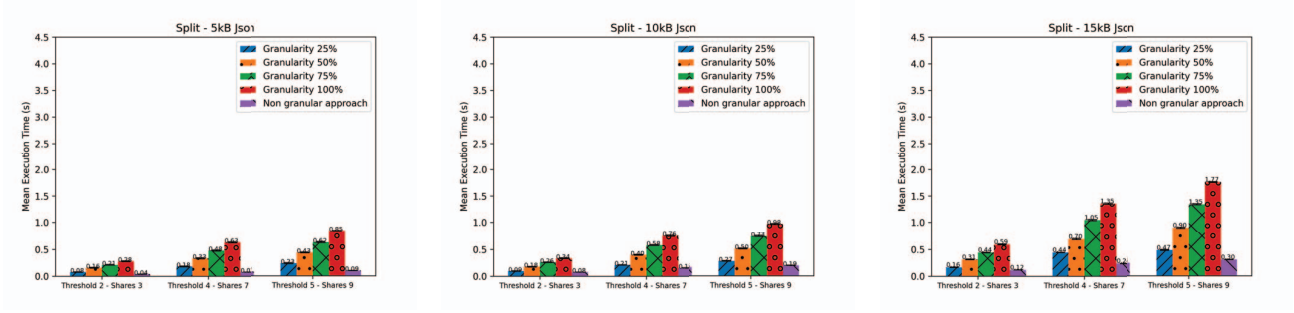
encrypted data across distributed systems increasing reliability in its management. It applies encryption to the data by dividing it into several units, known as *shares*. This operation is known as *split*. Thanks to its distributed nature, Secret Sharing is a good solution to implement pseudonymization in a distributed environment such as Data Space. To retrieve and decrypt the original data, which is called *secret*, a subset of shares is needed. This increases the availability of the data which can be retrieved even if some devices are not available for any reason. Furthermore, this feature is essential when handling health data, which should always be available since people's lives could depend on it. The cardinality of the subset of shares needed to retrieve the secret is known as *threshold*, while the data retrieval operation is called *merge*. For our experiments, we implemented a REST server written using the FastAPI[2] Python web framework. Once our server has been deployed on all Raspberry Pis within the same local network (LAN), we could effectively simulate communication in a hospital environment. Each Raspberry Pi acts as a node in the distributed system, allowing actors within the hospital to access and utilize the resources provided by the REST server through specific endpoints. In Table III, we provide a detailed overview of the endpoints exposed by our server, representing the functionalities and resources accessible to devices within the hospital.

B. Performance Evaluation

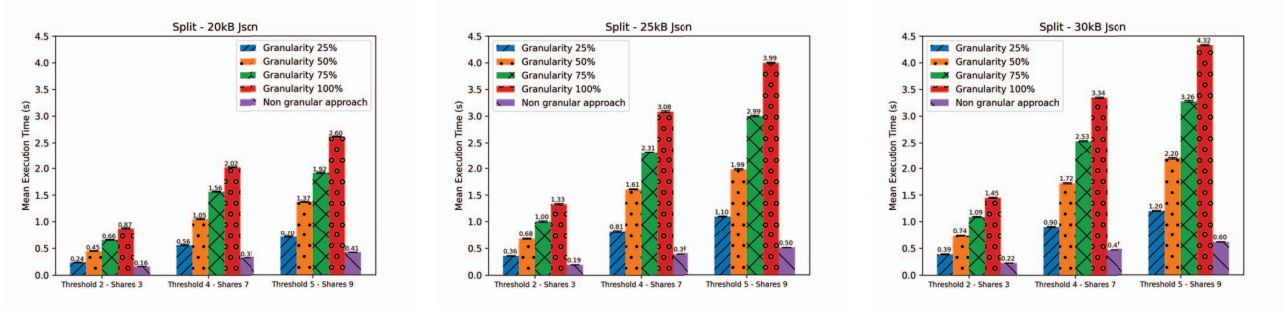
We tested the performance of our granular approach for intent-based pseudonymization of HL7 FHIR JSON documents, by comparing it with non-granular encryption of the entire document. More specifically we evaluated 4 different levels of granularity (25%, 50%, 75%, and 100%), in 3 distinct Secret Sharing configurations (Threshold 2 - Shares 3, Threshold 4 - Shares 7, and Threshold 5 - Shares 9). We sampled the execution times of the Secret Sharing split and merge operations on several HL7 FHIR document sizes (5kB, 10kB, 15kB, 20kB, 25kB, and 30kB). In particular, we collected 100 samples for both the split and merge operations performed on documents of each stated size. Moreover, we did it for each of the above-mentioned levels of granularity, and each Secret Sharing configuration. After that, we also calculated the mean value along with the 95% confidence interval of each group of 100 samples by employing the t-student distribution function. Figure 3 resumes the results obtained for the split operation, meanwhile, the results obtained for the merge operation are detailed in Figure 4.

The average execution times of both the split and merge operations increase linearly with the level of granularity. If we consider the several Secret Sharing configurations, instead, the increment of the execution time is always linear but depends on different parameters for each operation. On the one hand, the mean execution time of the split operation increases linearly with the number of shares. On the other hand, the average execution time of the merge operation increases linearly with the threshold. As regards the approach, in the case of the split operation, the mean execution time of the non-granular approach is always lower than the one of the granular approach for each one of the evaluated levels of granularity. On the contrary, in the case of the merge operation, the average execution time of the non-granular approach is lower than the ones of the 50%, 75%, and 100% levels of granularity. Instead, considering the Threshold 4 - Shares 7 and the Threshold 5 - Shares 9 configurations, it is greater than the one of the 25% level of granularity. Meanwhile, concerning the Threshold 2 - Shares 3 configuration, the execution times are almost equal to the one of the 25% level of granularity. In general, the non-granular approach, which encrypts the whole document, is faster than the granular one at 100% of granularity. However, when applying the non-granular approach, there is the need to decrypt the entire document to retrieve some information contained in it. Thus, this approach cannot be employed when handling health data and there is the need to be compliant with privacy regulations and, therefore, to access only the required portions of data. The granular approach, instead, enabling intent-based pseudonymization, allows decrypting only the required portion of the data in acceptable times. Furthermore, when the size of the portion of the HL7 document that needs to be retrieved is less or equal to the 25% of the document's total size, the granular approach performance is equal, or even better, than the non-granular one.

[2]https://fastapi.tiangolo.com/

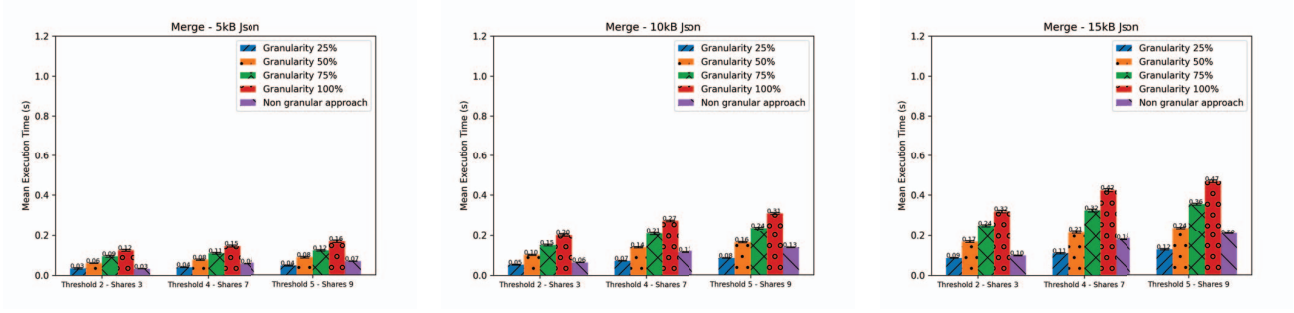


(a) (b) (c)

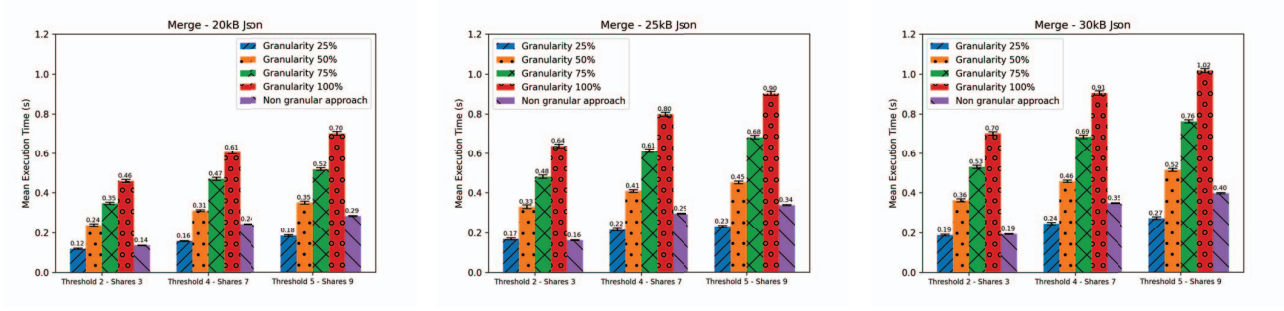


(d) (e) (f)

Figure 3. Split Average Execution Time (s)



(a) (b) (c)



(d) (e) (f)

Figure 4. Merge Average Execution Time (s)

IX. Conclusion

In this paper, we have addressed the critical need for enhanced data security in Intra-Hospital Data Spaces by proposing and implementing a novel intent-based multilevel

granular pseudonymization approach tailored for HL7 FHIR JSON documents. Our proposal offers a fine-grained and contextualized approach to data protection, allowing customization of pseudonymization techniques based, not only on data sensitivity but also on specific intentions. We presented a use case regarding the implementation of a workflow within the context of Intra-Hospital Data Spaces. This case has tangibly demonstrated the need for an intent-based approach in a real operational context. Through a careful comparative analysis with traditional non-granular encryption techniques, we have highlighted the advantages of our approach in terms of data security, privacy protection, and operational efficiency within hospital Data Spaces. Looking to the future, further research could explore the integration of key and access index obfuscation techniques into HL7 FHIR documents. Additionally, it is important to work towards standardization and interoperability of pseudonymization techniques across different healthcare institutions to promote data sharing and the effectiveness of data security measures in the healthcare sector.

ACKNOWLEDGMENT

This work has been partially supported by the European Union (NextGeneration EU), through the MUR-PNRR project SAMOTHRACE (ECS00000022), the Italian Ministry of Health, Piano Operativo Salute (POS) trajectory 2 "eHealth, diagnostica avanzata, medical device e mini invasività" through the project "Rete eHealth: AI e strumenti ICT Innovativi orientati alla Diagnostica Digitale (RAIDD)"(CUP J43C22000380001), and the Italian Ministry of Health, Piano Operativo Salute (POS) trajectory 4 "Biotechnology, bioinformatics and pharmaceutical development", through the Pharma-HUB Project "Hub for the repositioning of drugs in rare diseases of the nervous system in children" (CUP J43C22000500006).

REFERENCES

[1] D. Paparova, M. Aanestad, P. Vassilakopoulou, and M. K. Bahus, "Data governance spaces: The case of a national digital service for personal health data," *Information and Organization*, vol. 33, no. 1, p. 100451, 2023.

[2] D. N. Jutla, P. Bodorik, and S. Ali, "Engineering privacy for big data apps with the unified modeling language," *Proceedings - 2013 IEEE International Congress on Big Data, BigData 2013*, pp. 38 – 45, 1 2013.

[3] R. Noumeir, A. Lemay, and J. M. Lina, "Pseudonymization of radiology data for research purposes," *Journal of Digital Imaging*, vol. 20, pp. 284 – 295, 9 2007.

[4] L. Bolognini and C. Bistolfi, "Pseudonymization and impacts of big (personal/anonymous) data processing in the transition from the directive 95/46/ec to the new eu general data protection regulation," *Computer Law and Security Review*, vol. 33, pp. 171 – 181, 4 2017.

[5] A. C. Machado and D. F. Polónia, "Legal and technological aspects for the creation of a european health data space," *Iberian Conference on Information Systems and Technologies, CISTI*, vol. 2022-June, 2022.

[6] "Recommendations on shaping technology according to gdpr provisions - an overview on data pseudonymisation — enisa." https://www.enisa.europa.eu/publications/recommendations-on-shaping-technology-according-to-gdpr-provisions. (Accessed on 10/12/2023).

[7] W. Kim and J. Seok, "Privacy-preserving collaborative machine learning in biomedical applications," in *2022 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pp. 179–183, 2022.

[8] C. Tachepun and S. Thammaboosadee, "A data masking guideline for optimizing insights and privacy under gdpr compliance," in *Proceedings of the 11th International Conference on Advances in Information Technology*, IAIT2020, (New York, NY, USA), Association for Computing Machinery, 2020.

[9] T. Neubauer and J. Heurix, "A methodology for the pseudonymization of medical data," *International Journal of Medical Informatics*, vol. 80, pp. 190 – 204, 3 2011.

[10] Y. Liu, D. Niu, S. Geng, J. Sun, and H. Zhang, "Application of data integration in dataspace in multi-value chain collaboration of electric power manufacturing industry," in *2022 IEEE 25th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pp. 292–298, 2022.

[11] M. Parciak, M. Suhr, C. Schmidt, C. Bönisch, B. Löhnhardt, D. Kesztyüs, and T. Kesztyüs, "Fairness through automation: development of an automated medical data integration infrastructure for fair health data in a maximum care university hospital," *BMC Medical Informatics and Decision Making*, vol. 23, p. 94, 12 2023.

[12] W. Li and P. Quinn, "The european health data space: An expanded right to data portability?," *Computer Law & Security Review*, vol. 52, p. 105913, 2024.

[13] M. A. Sahi, H. Abbas, K. Saleem, X. Yang, A. Derhab, M. A. Orgun, W. Iqbal, I. Rashid, and A. Yaseen, "Privacy preservation in e-healthcare environments: State of the art and future directions," *IEEE Access*, vol. 6, pp. 464 – 478, 10 2017.

[14] G. Morabito, V. Lukaj, A. Ruggeri, M. Fazio, M. A. Astone, and M. Villari, "Docflow: Supervised multi-method document anonymization engine," in *2023 IEEE Symposium on Computers and Communications (ISCC)*, pp. 1–6, IEEE, 2023.

[15] G. Morabito, C. Sicari, A. Ruggeri, A. Celesti, and L. Carnevale, "Secure-by-design serverless workflows on the edge–cloud continuum through the osmotic computing paradigm," *Internet of Things*, vol. 22, p. 100737, 2023.

[16] M. Franklin, A. Halevy, and D. Maier, "From databases to dataspaces: a new abstraction for information management," *SIGMOD Rec.*, vol. 34, p. 27–33, dec 2005.

[17] E. Curry, S. Scerri, and T. Tuikka, *Data Spaces: Design, Deployment, and Future Directions*, pp. 1–17. Cham: Springer International Publishing, 2022.

[18] T. Coenen, N. Walraevens, G. Terseglav, A. Lampe, I. Lakaniemi, U. Ahle, L. Raes, B. Lutz, N. Reisel, W. V. D. Bosch, and G. V. Delannoy, "Gaia-x: Sustainable cloud computing," 2021.

[19] Gaia-X Association, "Gaia-x: Federated data infrastructure for the european data economy," 2022.

[20] K. Hoeyer, S. Green, A. Martani, A. Middleton, and C. Pinel, "Health in data space: Formative and experiential dimensions of cross-border health data sharing," *Big Data & Society*, vol. 11, no. 1, p. 20539517231224258, 2024.

[21] J. Hernandez, L. McKenna, and R. Brennan, "Tikd: A trusted integrated knowledge dataspace for sensitive healthcare data sharing," in *2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC)*, pp. 1855–1860, 2021.

[22] A. Galletta, J. Taheri, and M. Villari, "On the applicability of secret share algorithms for saving data on iot, edge and cloud devices," in *2019 International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*, pp. 14–21, 2019.