Securing and sharing clinical data

Pedro Miguel Baptista dos Reis

Thesis to obtain the Master of Science Degree in

Information Systems and Computer Engineering

Examination Committee

Chairperson: Prof. José Carlos Alves Pereira Monteiro
Supervisor: Prof. Ana Teresa Correia de Freitas
Co-Supervisor: Prof. Alexandre Paulo Lourenço Francisco
Member of the Committee: Prof. José Luís Brinquete Borbinha

November 2013
Acknowledgements

First and foremost, I must thank my co-advisor, professor Alexandre Francisco, for his availability in the revision of this thesis. Also, his sharp and objective comments promptly set me right on track, whenever I felt bewildered. His undeniable talent serves as inspiration to self-improvement and I hope to continue working with him in the forthcoming years.

I also thank my advisor, professor Ana Teresa Freitas, for the trust deposited in me and for providing me with the needed liberty to achieve my academic goals while investing in demanding entrepreneurship endeavours.

Lastly, I want to thank those who shared precious time with me in the KDBIO group at INESC-ID. I already miss the richness and diversity of our conversations, ranging from deep philosophical subjects to recreational suggestions.

Thank you all.
Abstract

The low success rate of clinical trials for rare diseases is a growing concern. The cause for this failure lies in the segregation of clinical data, promoted by informed consents and international data protection laws. Only on the second half of 2012 medical and research communities have started the discussions and debates to address this problem. Some of the proposals are to put the majority of patient data openly available on the internet. However, studies on personal data exposure and usurpation reveal worrying consequences in cases of identity theft, as well as high financial impacts for the state and insurance companies due to resulting lawsuits.

The purpose of this thesis is to synthesize and formalize the legal requirements to deal with clinical data, resorting to an analysis targeting specification beyond the sole technological perspective of the problem. Moreover, it properly frames the problem in the context of the already known data access control models. Additionally, it uses recent proposals in distributed authentication protocols to simplify the implementation process on a global scale. Finally, it demonstrates the applicability of Linked Data technologies to cope with the required heterogeneity of clinical data, as well as to promote the integration of data across multiple healthcare institutions.
A baixa taxa de sucesso dos estudos clínicos de doenças raras é uma crescente preocupação. A causa deste insucesso reside na segregação de dados clínicos, fomentada pelos consentimentos informados e pelas leis internacionais de protecção de dados. Apenas na segunda metade de 2012 começaram os debates e discussões no seio da comunidade médica e de investigação para abordar este problema. Algumas das propostas visam a total disponibilização dos dados dos doentes na internet. Contudo, estudos sobre exposição e usurpação de dados pessoais relevam consequências preocupantes em casos de furto de identidade, bem como, elevados impactos financeiros para o estado e companhias de seguros, derivados de processos judiciais.

Esta tese tem como objectivos a síntese e formalização dos requisitos legais para lidar com dados clínicos, recorrendo a uma análise que visa a especificação para além da perspectiva tecnológica do problema. Além disto, enquadrada devidamente o problema à luz dos já conhecidos modelos de controlo de acesso a dados. Adicionalmente, recorre a propostas recentes de protocolos de autenticação distribuídos para facilitar o processo de implementação a uma escala global. Finalmente, demonsstra a aplicabilidade das tecnologias de Linked Data para satisfazer o requisito de heterogeneidade dos dados clínicos, bem como, para promover a integração dos dados entre várias instituições de saúde.
Keywords

Clinical Data
Security
Privacy
Linked Data
Systems Architecture

Palavras-Chave

Dados Clinicos
Segurança
Privacidade
Dados Interligados
Arquitectura de Sistemas
Contents

1 Introduction 
   1.1 Context and Motivation ........................................... 1
   1.2 Contributions ..................................................... 2
   1.3 Document Structure .............................................. 2

2 Background 
   2.1 Privacy Laws and Principles .................................... 5
   2.2 Information Security ............................................... 6
     2.2.1 Identification ................................................. 6
     2.2.2 Authentication .................................................. 6
     2.2.3 Authorization .................................................... 7
     2.2.4 Confidentiality ............................................... 7
     2.2.5 Accountability ............................................... 8
   2.3 Access Control .................................................... 8
     2.3.1 Access Control Policies ................................... 9
     2.3.2 Access Control Models ................................... 12
   2.4 Linked Data ..................................................... 19
     2.4.1 Serialization Formats .................................... 20
   2.5 Private Health Data ............................................. 21
   2.6 Ongoing Research ............................................... 22
     2.6.1 Open Access .................................................. 22
     2.6.2 Streamlined Access ....................................... 23
     2.6.3 Research Commons ....................................... 23
     2.6.4 Data Analysis Servers .................................... 24
3 Global Healthcare System

3.1 Analysis Framework .................................................. 25

3.2 Requirements ......................................................... 28

3.2.1 Multi-Institutional and Global Scales ................................ 31

3.3 Identification and Authentication .................................... 31

3.3.1 Practitioners and Researchers .................................... 32

3.3.2 Patients .............................................................. 33

3.4 Anonymization ........................................................ 35

3.4.1 k-Anonymity ......................................................... 37

3.4.2 l-Diversity .......................................................... 38

3.4.3 t-Closeness ......................................................... 39

3.5 Summary .............................................................. 39

4 Proposed System ....................................................... 41

4.1 Data Heterogeneity ................................................... 41

4.2 Case Study ............................................................. 42

4.3 Data Structure ........................................................ 42

4.3.1 Evolution and Integration ......................................... 45

4.4 Access Control ......................................................... 45

4.4.1 Practical Example .................................................. 46

5 Conclusions and Future Work ....................................... 49

A Decomposing Management ........................................ 57

A.1 Management as a Natural Phenomenon .......................... 57

A.2 Management and Leadership ....................................... 57

B Virtual Environment .................................................. 59

C Demo Setup ........................................................... 61

D Architecture .......................................................... 63
List of Figures

2.1 Policy enforcement architecture. ................................................. 8
2.2 Structures in MAC rules. ...................................................... 12
2.3 The RBAC model .................................................................. 12
2.4 Phases of Access Control Design ........................................... 13
2.5 Organization of the Chinese Wall model. .............................. 16
2.6 The RBAC family of models. .................................................. 16
2.7 The RBAC\(_0\) Model. ............................................................. 17
2.8 The RBAC\(_1\) Model. ............................................................. 17
2.9 The RBAC\(_2\) Model. ............................................................. 18
2.10 An RDF triple. .................................................................. 20

3.1 Access Levels and Transitions. ............................................... 27
3.2 Processes Activities and Tasks. .............................................. 28
3.3 Architecture for requirements 1 to 7. ................................... 29
3.4 Architecture for requirements 1 to 11. ................................... 30
3.5 Example of WebID utilization. ................................................. 33
3.6 Generating an UHID. ............................................................ 35
3.7 Re-identification by linkage. ................................................ 37
3.8 Generalized dataset where \(k = 2\). ......................................... 38
3.9 Homogeneity issue ................................................................. 38

4.1 Healthcare data structure ....................................................... 43
4.2 Data structure for clinical studies ........................................... 44
4.3 Results from select all SPARQL query with access control ....... 47
List of Tables

2.1 Access Control Matrix. ................................................. 10
2.2 Authorization table. .................................................. 10
2.3 Access Control List for file1. ...................................... 10
2.4 Capabilities List for Alice. ........................................ 11

3.1 The Six Dimensions Framework ................................... 25
3.2 System Stakeholders .................................................. 28
3.3 HIPAA identification attributes (excluding 2 and 3) .......... 36
3.4 Requirements analysis summary .................................. 39

4.1 List of roles and respective privileges ......................... 46
4.2 Example dataset ...................................................... 47

A.1 Differences between management and leadership ........... 58
Introduction

We spend our time searching for security and hate it when we get it. – John Steinbeck

This dissertation addresses the topic of securely sharing clinical data, which is still being discussed on international research groups and governments. It reflects the recent awareness, in society, for what is known as the Privacy and Progress issue. Some of the supporting references are still under technical and political debate, hence, the cited scenarios and techniques described herein may be subjected to changes in a near future.

1.1 Context and Motivation

Data breach incidents are a global growing trend, not only because personal records have become digital, but also because sensitive information is largely handled in personal and mobile devices. The healthcare industry has the highest cost per exposed record, and also, it is the sector responsible for the largest amount of data breaches worldwide [37, 47]. Identity theft is the top reason for illegally obtaining private patient data. Stolen identities are generally used to obtain healthcare services and pharmaceuticals. A smaller percentage of stolen data is manipulated, records may be changed to show different kinds of drug allergies or blood types, and therefore, increasing the risk of lethal treatment being applied to any victim of identity theft [11, 13, 36].

On the other hand, research on cancer therapy and other rare genetic diseases is suffering a downturn in its progress, mainly due to information silos derived by the doctor-patient agreements and data protection laws [21]. Studies show that roughly 80% of the patients treated for cancer die because of drug inadequacy [14]. This incompatibility between drugs and patients is caused by insufficient information that can be mitigated if researchers have access to private clinical data at a larger scale. Current requirements for progress in healthcare science are data integration and access to whole genome sequence and to clinical data [39].

As a result, access to private data is required for scientific progress, and yet, addressing privacy concerns is fundamental to protect individuals from fraudulent schemes and social discrimination. Therefore, the architectural design of a possible solution must take into account requirements like large scale data integration, security policies enforcement and the heterogeneous nature of clinical data.

Considering the heterogeneity of clinical data and the need for large scale data integration, the technologies to support this requirements are expected to be the same as those supporting the Web, or those derived from World Wide Web Consortium (W3C) standards. The most relevant standards are those related with the Semantic Web, which provide a rich set of tools to support exactly the described requirements. Because the W3C vision for the Semantic Web is based on Linked Data, it is relevant to evaluate its potential for enabling
large scale heterogeneous data integration. This, however, has already been showcased by services like the Linked Life Data [35]. By integrating tens of public databases it enables the definition of faster and more precise drug discovery processes due to simultaneous analysis of medical, biological and chemical databases. It also demonstrates how Semantic Web technologies can be used to provide new insights based on already existing data, given the fact that inferred data is nearly 35% of the whole dataset.

The purpose of this thesis is, firstly, to provide a simplified and concise description of the legal requirements in healthcare data. Also, it aims to discuss the access control models and protocols applicable to implementing a system supporting the already described requirements. Moreover, it identifies open standards and tools that enable the development of large scale systems featuring the security enforcement needed to ensure the appropriate access to private clinical data, while providing integration and data heterogeneity support.

1.2 Contributions

This thesis provides the following contributions:

1. The definition of a six dimensional framework for the analysis of healthcare systems.

2. A concise definition of the legal requirements and representation of the resulting architectural designs.

3. A data structure to enable role-based access control to healthcare data, also being ontology agnostic in order to enable the adoption of already existing medical ontologies.

4. A case-study of authentication standards applicability for Linked Data [19].

5. Accepted paper “Securing and sharing clinical data” on NIPS 2013 Workshop on Machine Learning for Clinical Data Analysis and Healthcare.

6. A prototype providing access control authorization using Linked Data.

1.3 Document Structure

This thesis is organized as follows:

- Chapter 2 - Background - Covers the legal directives of data protection laws and related security principles. It discusses access control models and their applicability. Finally, it describes how patient data is stored in the healthcare sector and what are the ongoing research approaches to implement controlled access on healthcare data at a global scale.

- Chapter 3 - Global Healthcare System - Defines an analysis framework and enumerates the concise requirements resulting from analysis. Presents architectural diagrams to help envisioning a global healthcare system. Discusses the alternatives for implementing the fundamental security principles.
Chapter 4 - Proposed System - Discusses the essential functional requirements. It describes the advantages and limitations on the first developed efforts to implement a system using Linked Data and providing access control. It proposes a data structure appropriate for Linked Data, while considering the natural evolution of the system in terms of ontology design and data integration. Finally, it provides an example of how to implement access control on healthcare data using Linked Data.

Chapter 5 - Conclusions and Future Work - Presents the conclusive remarks resulting from this work and describes some remaining concerns towards the implementation of a global healthcare system.
In this chapter we introduce the definitions of privacy and describe its legal principles, focusing on the healthcare industry. Section 2.2 describes the fundamental information security principles involved, and how they map onto the legal principles. Section 2.3 outlines the policies and models that constitute the foundation of access control, and which security principles they target or support. Section 2.4 defines Linked Data focusing on its importance for the representation of healthcare data, and Section 2.5 discusses the nature of private data in the healthcare sector and its architectural implications. Finally, Section 2.6 presents related work from other research groups.

### 2.1 Privacy Laws and Principles

Privacy, in its classical sense, is a concept related with the control of physical exposure, thus not being directly associated with information. The information age has changed this notion, redefining it on both popular and legal perspectives. Nowadays, the popular (and previously physical) notion of privacy is considered to be intimacy, while privacy *per se*, is perceived as the capacity of an individual to decide when, to whom and how information about himself may disclosed.

In its turn, data protection laws consider privacy to be a direct consequence of the inability to access personal data. Data is considered personal when it allows anyone to identify a person either directly or indirectly [17]. Direct identification is usually enabled by nationwide number-person association (e.g. Social Security Number). Indirect identification occurs when someone narrows down a search result to a single person based only on their characteristics and other related data, for example, age, gender, height, state, country, etc. Data protection laws also aim to prevent identification resulting from data integration and processing, meaning that even in the cases when an entity (person or institution) cannot identify an individual with the data they have, they still must comply with the legal directives. The European Community has defined seven principles within the private data law, described as follows:

- **Notice**: if data is collected, the subject must be informed.
- **Purpose**: collected data should only be used for the specified purpose and no other purposes.
- **Consent**: subjects are required to provide consent before data is shared with third parties.
- **Security**: private data should be kept secure from loss, manipulation or theft.
• **Disclosure**: subjects should be informed of who is collecting their data.

• **Access**: subjects should have access to their personal data and be able to correct it if needed.

• **Accountability**: entities holding personal data are accountable to subjects in respect to compliance with these principles.

The informed consent is the legal instrument that complies with the data protection laws, thus enabling clinical studies. It takes the form of a document that describes the agreement made between a patient and either a doctor, a medical team or research group. The ‘purpose’ principle is the one responsible for information segregation, since informed consents explicitly state the purpose of collected data, it prevents data reuse on other closely related studies. One way to overcome this limitation is to perform a re-consent, which may not be always feasible if patients are inaccessible or deceased. An informed consent is, nonetheless, a promise of conduct from the data collectors, that enables trust from the general public and subjects in a clinical study. In order to maintain this trust relationship, the public expectations are transparency, compliance and consequence if agreements are not followed.

### 2.2 Information Security

Information security is defined as “protecting information and information systems from unauthorized access, use, disclosure, disruption, modification, or destruction” [52]. Like data protection laws, it is composed of several principles that help in defining the policies required to secure an information system. Considering the presented legal principles, it is required to understand how they will be supported in the technological realm. From all the existing security principles, only those required for legal compliance are described in the following subsections.

#### 2.2.1 Identification

Identification is the fundamental pillar in information security, its the principle on which all other depend. The unique identification of an individual or group of individuals, takes the form of an identifier that may be either a number or a string. By associating the identifier with an individual, identification takes place.

Some systems require identifiers to be reusable, specially when technical limitations on the memory space or value representation implies that the system runs out of available identifiers. The Internet Protocol (IP) addresses are an example of such reusable identifiers.

#### 2.2.2 Authentication

To authenticate is to prove an identity. In other words, when someone claims to be a user of a specific system (e.g. by introducing a user-name), the system requests a challenge that the person is required to know in order to prove its identity.
There are three types of authentication methods, known as authentication factors:

1. **Knowledge**: Based on something the user knows, for example, a Personal Identification Number (PIN), a password, a code, etc.
2. **Possession**: The user is required to possess something, such as, a key, a smart card, etc.
3. **Inherence**: Based on the recognition of biometric characteristics, for example, fingerprint, voice, retina, etc.

Systems may require multi-factor authentication, being the two-factor authentication the most widely adopted. For example, two-factor authentication is used in Automated Teller Machines (ATM), requiring the ATM card (possession) and the PIN (knowledge).

### 2.2.3 Authorization

Authorization consists in either granting or denying access to certain resources. The collection of grants a user has on the resources is called user privileges. Each item in this collection is usually defined as a triple of the form \((\text{user}, \text{resource}, \text{operation})\) where the resource represents an object of the system (e.g. file, document, image) and the operation stands for the actions allowed (e.g. read, copy, delete). An authorization example is when a user has read access to a specific document, but has no permission to modify it.

Policy definition and enforcement are the requirements for authentication to take place. The policies that govern authorization impose a specific model to control operations on resources. These models are known as access control models. If an organization defines the level of privileges by directly matching a user with a role, then a role-based access control model should be used. Likewise, if an organization has per user definition of privileges, then a discretionary access control model should be used.

The three essential principles on security are identification, authentication and authorization, having to occur in this exact order. Having authorization, any system will be able to comply with the legal principle of security, in the sense that it will control intentional loss, manipulation and theft. There may be, however, the case of unintentional loss, which is solved with regular data backup. Differentiating between unintentional manipulation or theft its only possible outside the computer realm, as described ahead in the accountability subsection 2.2.5.

### 2.2.4 Confidentiality

Confidentiality is somewhat related with privacy in the healthcare sector, nonetheless, they are not the same thing. As already stated, privacy is the right that an individual has of deciding when, to whom and how information about himself may disclosed. Confidentiality, on the other hand, requires an agreement, between two or more parties, consisting in the non-disclosure of shared information. Typically, confidentiality is used to keep certain trade secrets unknown to competition.
In the healthcare sector confidentiality results from doctor-patient agreements, whereby caregivers agree not to disclose any information that otherwise would be completely private. According to data protection laws, disclosure of private information can only happen after the patient has consented it.

The security implications of confidentiality extend beyond the restrictions that any information system can enforce. This happens whenever a caregiver, even if legally authorized to access a patient record, discloses private information with people not authorized by the patient. This usually happens in informal conversations outside the professional environment, thus making it nearly impossible to detect these occurrences. Confidentiality, in an information system, can still be ensured in cases of unauthorized access if data is encrypted before storage.

### 2.2.5 Accountability

Accountability is the extension of responsibility to include accounting (explanations or reasons) for the performed actions. Hence, when someone is responsible for a specific task, it only means that is possible to identify who executed that task. When someone is accountable for a specific task, it means that explanations, as to the reasons substantiating the task execution, must be given to other stakeholders. As defined in the legal principles, accountability is expected from any element of a medical staff (included in an informed consent or agreement) in relation to a patient. Moreover, accountability is likely to become liability if patients decide to take the matter to any court, whenever considerable damage takes place. Liability, is therefore the extension of accountability to include the possibility of sanctions or penalties deemed in a court of law.

The only way to determine if loss, manipulation and theft are intentional or unintentional, is thought accountability. However, in order to hold someone accountable, it is required to determine who is responsible for every action. The tools that ascertain responsibility are logs or audit trails.

### 2.3 Access Control

Access control is what enables authorization by mediating user access to resources and it is composed of policies, models and mechanisms. Usually the mechanisms involved in access control follow the architectural pattern shown in Figure 2.1.

![Policy enforcement architecture](image-url)
Resource accesses in this architecture start with a user request to a resource (arrow 1) to perform a specific operation. The request is intercepted by a Policy Enforcement Point (PEP) that provides the request details (arrow 2) to the Policy Decision Point (PDP) and awaits the evaluation result. The PDP accesses the Policy Information Point (PIP) containing all the policies regarding resource access (arrow 3). The PDP determines if the user may (or may not) access the resource for the requested operation and informs the PEP (arrow 5). The PEP enforces policy decision by either allowing access to the resource, or denying it.

There are three known principles that should be considered as guidance during policy definition, which are:

- **Least privilege**: users should not be assigned more privileges than those strictly needed to perform the task at hand.

- **Separation of duties**: the same user within one organization should not be assigned responsibilities that lead to conflicting interests. For example, creating a budget plan and authorizing it.

- **Need to know**: users should only access information that allows them to perform their job. For example, it is not required for a systems administrator to know the passwords of users in order to create user accounts.

### 2.3.1 Access Control Policies

A policy is a rule governing the behaviour of a system. Therefore, access control policies are rules that govern user access to resources in a system. These policies are typically determined in management activities and are generally defined in human readable rules. Examples of such policies describe how a user account is created, by whom, its privileges, its validity and revocation. Policy complexity is directly proportional to organization size, resource heterogeneity, technological diversity and human specialization variety.

When by default all users have access to all resources, it is called an open policy. The opposite happens in closed policies, where users have denied access by default and rules only grant access. Policies are also divided into three classes: Discretionary Access Control, Mandatory Access Control and Role-based Access Control.

#### 2.3.1.1 Discretionary Access Control

Discretionary Access Control (DAC) policies are defined by the resource owner (or creator), who applies them individually to other users. The most widely used DAC model is the Access Control Matrix (ACM) initially proposed by Lampson [29], it was refined one year later by Graham and Denning [23] and it was formalized four years later by Harrison, Ruzzo, and Ullmann (HRU model) [25]. Table 2.1 shows an example of a Access Control Matrix.
The typical implementations of ACM are Authorization tables, Access Control Lists (ACL) and Capability Lists (CL). These three implementations are specific views on ACMs to allow space optimization. As shown in Table 2.1, there may be empty cells taking up space unnecessarily. Table 2.2 shows an example of an authorization table, where is observable the absence of empty cells. Authorization tables are generally used in Database Management Systems (DBMS).

<table>
<thead>
<tr>
<th>User</th>
<th>Resource</th>
<th>Operation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>file1</td>
<td>read</td>
</tr>
<tr>
<td>Alice</td>
<td>file1</td>
<td>write</td>
</tr>
<tr>
<td>Alice</td>
<td>file2</td>
<td>read</td>
</tr>
<tr>
<td>Bob</td>
<td>file1</td>
<td>read</td>
</tr>
<tr>
<td>Bob</td>
<td>file3</td>
<td>read</td>
</tr>
<tr>
<td>Bob</td>
<td>file3</td>
<td>write</td>
</tr>
<tr>
<td>Charlie</td>
<td>file2</td>
<td>read</td>
</tr>
</tbody>
</table>

Access Control Lists are the column projection of an ACM. They are associated with the resources and are widely used to manage file security in operating systems. Even in situations where the policy is set centrally, thus not at the user discretion, ACLs are suited if protection is resource oriented. A way of simplifying ACL management for a large number of users is to define groups of users. A disadvantage of ACLs stems from the need determine privileges by user, requiring all resource ACLs to be searched. Table 2.3 shows an example of an ACL.

<table>
<thead>
<tr>
<th>User</th>
<th>Operation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>read ; write</td>
</tr>
<tr>
<td>Bob</td>
<td>read</td>
</tr>
</tbody>
</table>

On the other hand, CLs are the row projection of ACMs and are associated with users. It is understood that ACLs and CLs have inverse advantages and disadvantages. For example, changing the privileges on a file would be time consuming due to the need to traverse all CLs of all users. However, a delegation task is trivially performed using CLs. Considering Alice wants to delegate her write privilege on file1 to Bob, she would just be required to create a signed certificate containing her write privilege and deliver it to Bob. Table 2.4 shows the capabilities list for Alice.

<table>
<thead>
<tr>
<th>User</th>
<th>Operation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>read ; write</td>
</tr>
<tr>
<td>Bob</td>
<td>read</td>
</tr>
</tbody>
</table>
DAC has three important disadvantages:

1. it depends on user ability to set the security restraints, when in most cases users are not prepared to assume that responsibility;

2. vulnerability to trojan horses, since users are the owners of the resources they create, some programs may exploit those privileges;

3. no control over flow of information. Information may be copied to other resources and the original owner has no longer control over his information.

### 2.3.1.2 Mandatory Access Control

Mandatory Access Control (MAC) means that user privileges are defined by a central authority instead of being at the discretion of each resource owner. This type of policy is typically used in military organizations and uses an intermediary classification (or labelling) for users and resources. Classifications on users are called categories and on resources are clearance levels. Clearance levels are hierarchical and reflect the potential damage resulting from information disclosure. Hence, MAC policies consider flow of information and not only resource access like in DAC policies. Examples of military clearance levels are Top Secret (TS), Secret (S) and Confidential (C).

Considering again the triple \((user, resource, operation)\) typical in access control, it is possible to say that users map to categories, resources map to clearance levels and operations (e.g. read) obey to rules like “no read up” [28].

Clearance levels are in fact partial order relations and they typically form lattices when graphically represented together with categories [42]. When considering a single category, a ring diagram is commonly used. Figure 2.2 shows a lattice (on the left) depicting the military clearance levels applied to two categories (Army and Navy), and on the right, a ring diagram illustrating the “no read up” rule applied only to clearance levels.

The main disadvantage of MAC is that it requires software to be trusted and compliant with the defined policies. This leads to considerable investments in software validation and certification.

### 2.3.1.3 Role-based Access Control

Role-based Access Control (RBAC) policies are based on user roles instead of user identification. It is considered to be more flexible than DAC and MAC because it aligns with commercial and civilian government security requirements [18]. It follows the notion that in most organizations resources are owned by the organization and not their employees, who are users within the system. Also, users have certain roles that require specific
privileges. These privileges are less likely to change than the users who require them, due to employee turnover or promotions. Roles are modelled as an indirection between users and privileges, providing a more wide and flexible usage than groups in DAC models. While groups are only a set of users, roles can form collections of users, privileges and other roles [41]. Figure 2.3 shows the RBAC model, where the relationship between users and roles is called User Assignment (UA), and the one between roles and privileges is called Permission Assignment (PA).

RBAC also enables users to activate specific allowed roles to perform their tasks. For example, an accountant may temporarily require access to sales information in order to perform the period-end closing activities. Administration of access control policies can also be included in RBAC, thus providing a major advantage in management over DAC or MAC.

The drawbacks of RBAC are not directly associated with it, but with conjectural issues surrounding RBAC implementation and utilization. One of the drawbacks lies in the intrinsic complexity of administration tasks; companies with complex access control policies require a high level of expertise from system administrators. Another complaint about RBAC, based in a study published by the National Institute of Standards and Technology (NIST), is the lack of standardization adopted by different vendors where RBAC is built into their products [20]. This ends-up constituting scalability problems in systems supported by different vendors.

### 2.3.2 Access Control Models

Access control models are the formalization of access control policies. Figure 2.4 details the multi-phase design of access control. Each of the following models targets a specific principle (e.g. confidentiality or integrity) and either focus on resource access or information flow.
2.3. ACCESS CONTROL

![Phases of Access Control Design](image)

Figure 2.4: Phases of Access Control Design

**2.3.2.1 Bell-LaPadula**

The Bell-LaPadula (BLP) model targets confidentiality and was based in the military notions of clearance levels. It was the first mathematical model of a multilevel access control policy and influenced the development of other models. BLP controls information flow through the simple security property and the confinement property (*-property). These are known as the “no read up” and “no write down” principles, respectively.

The simple security property follows the reasoning of clearance levels, in which a user with a clearance level for secret resources cannot access the top secret ones. However, this will not suffice to guarantee confidentiality, since anyone with top secret clearance could declassify information by writing it in a resource with lower clearance level. The confinement property precludes information declassification by stating that a user cannot write resources with lower classification than himself.

The main drawback on BLP is not enforcing integrity on the higher levels of clearance, allowing a user with low clearance level to promote invalid information to clearance levels above his own. Also, BLP does not model how high clearance levels communicate with the lower ones, neither it includes management or enables delegation.

**2.3.2.2 Biba**

The Biba model targets integrity and its principles are inverse to those of BLP. As previously stated, the BLP model has an integrity drawback allowing lower clearance levels to promote invalid information to higher levels. The Biba model solves the BLP integrity drawback by stating the “no write up” and “no read down” principles.

It defines three levels of integrity (or classification), Crucial (C), Very Important (VI) and Important(I), where $C > VI > I$ is the hierarchical relationship between these levels. Like the BLP model it supports a set of categories that are non-hierarchical (e.g. Army and Navy). It also proposes five mandatory policies: strict integrity policy, low-watermark policy for subjects, low-watermark policy for objects, low-watermark integrity audit policy and ring policy. Moreover, it proposes three discretionary policies: access control lists, object hierarchy and ring. Note that subjects in this model refers to users or processes, and objects are resources.
The strict integrity policy is composed of the “no write up”, “no read down” and the invocation property. This last property is applied to processes, stating that process $P_1$ can only invoke process $P_2$ if and only if, the integrity level of $P_2$ is lower than the integrity level of $P_1$. This is the most restrictive of the proposed policies, but the most widely adopted.

The low-watermark policies are composed by two of the strict integrity policy principles and by relaxing one of the three principles. More precisely, the low-watermark policy for subjects, is composed of the “no write up” and invocation property, but relaxes the “no read down” principle by stating that if a user reads a resource with lower integrity level, then the integrity level of the user will become the level of the resource. Formally it is defined as: a Subject $s$ can read any object $o$. However, the integrity level of the Subject $s$, $il(s)$, will become the lowest of both integrity levels, $il'(s) = \min(il(s), il(o))$. Likewise, the low-watermark policy for objects maintains the “no read down” principle and invocation property, but relaxes the “no write up” principle by lowering the integrity level of a resource to that of the user who as written it.

The low-watermark integrity audit policy is based on the same relaxed “no write up” principle of the low-watermark for objects, but instead of lowering the integrity level of the resource, it records the occurrence through audit log.

The ring policy discards the “no read down” principle but maintains the “no write up” and invocation property. However, this policy leads to integrity issues by recurrent hierarchical writing of resources to the same level. Simply put, considering that the integrity levels are 1, 2 and 3 where $1 < 2 < 3$ a user $u_1$ having integrity level $il(u_1) = 2$ can read resource $r_1$ such that $il(r_1) = 1$ and write the contents to $r_2$ having $il(r_2) = 2$. Later a user $u_2$ may read $r_2$ and write its contents to resource $r_3$ with integrity level $il(r_3) = 3$. Hence, breaking the integrity principle of the Biba model.

As expected the main drawback of the Biba model is not enforcing confidentiality. Moreover, it does not include management or delegation like BLP. There are also shortcomings to the low-watermark policies, in which is observable that by lowering the integrity level of a user, if later on that user legitimately needs to read resources of its initial integrity level, it will see his request denied, thus, invalidating the work the user is assigned to perform. On the other hand, by relaxing the “no write up” principle, resources are easily declassified and it defeats the purpose of the Biba model, because every resource may lose its integrity.

### 2.3.2.3 Clark-Wilson

The Clark-Wilson model was the first model to target separation of duties. The model defines integrity as “No user of the system, even if authorised, may be permitted to modify data items in such a way that assets or accounting records of the company are lost or corrupted” [9]. To ensure integrity, it introduces the well-formed transaction and separation of duty mechanisms.

Well-formed transactions prevent users from arbitrarily changing the resources, instead there are programs authorized to do so. Moreover, it follows the rationale of double entry bookkeeping, where actions registered in the credit book must match debits on other book.
Separation of duties requires users to access a specific set of programs, which handle resources that are not sufficient to perform a whole business process, but only some tasks within the process.

The model defines two classifications for resources: Constrained Data Items (CDI) and Unconstrained Data Items (UDI). The CDIs require Integrity Verification Procedures (IVP) to ensure their conformity. Another procedure, called Transformation Procedure (TP), is defined to allow well-formed transactions. The TPs enable CDIs to change from one valid state to another.

To guarantee the formal correctness of the model, certification and enforcement rules are orderly defined to support system initialization and operation. Certification rules are done by a security officer and enforcement rules are performed by the system. There rules are denoted, for example, as $C_1$ for the first certification rule and $E_2$ for the second enforcement rule.

$C_1$: All IVPs must ensure that the state of all CDIs is valid at the time an IVP is run.

$C_2$: All TPs must behave correctly, meaning that, given a valid CDI as input they must produce a valid CDI as output. All CDIs a TP can handle must be part of the certification, defining the relation $(TP_i, (CDI_a, CDI_b, ...))$.

$E_1$: All CDIs must only be handled by TPs belonging in the relations defined in $C_2$.

$E_2$: The system should maintain a list of the form $(Users, TP, CDIs)$ and must only allow executions of that list to be performed by users.

$C_3$: The list defined in $E_2$ must be certified to meet separation of duties.

$E_3$: The system must authenticate any user attempting to execute a TP.

$C_4$: All TP must be certified to write to an append-only CDI (the log resource).

$C_5$: All TP upon taking a UDI as input, must be certified to either transform the UDI to a CDI or reject the UDI.

$E_4$: Only the user certified for a TP, may change the list associated with that TP. However, it may not execute that TP.

The Clark-Wilson suffers from scalability problems because it does not define or presents a way of handling concurrent (parallel) execution of the same resources, implying mutual exclusive (or sequential) manipulation.

### Chinese Wall

The Chinese Wall model targets conflict of interest. It was introduced to balance commercial discretion by preventing information flows causing conflict of interest for consultants. Knowing that a consultant may provide services for several companies, it is desirable that the same consultant does not hold information about two competing companies.

The model states three entities, the basic objects, company datasets and conflict of interest classes. The basic objects are the system resources, the company datasets are collections of resources belonging to the same company and conflict of interest classes are groups of companies that are competitors.

Additionally, two rules are defined to avoid conflict of interest: (1) The simple security rule states that a user may only access resources of the same company, or it may access resources from more than one company
if they do not belong to the same conflict of interest class. (2) The *-property states that write access is only permitted if access is allowed by the simple security rule, and, that no object can be read if it belongs to another company or if it contains non-sanitized information. Figure 2.5 helps to illustrate these properties, considering a user has access to Resource 1 from Company A, then by the simple security rule, it cannot access resources from Company B. However, it can access all the resources from either Company C or D. The *-property follows exactly the same principle as the simple security rule, with the exception that even within the same company a user may not have read access to an untrusted resource if it is writing another resource.

![Figure 2.5: Organization of the Chinese Wall model.](image)

The Bell-LaPadula, Biba and Chinese Wall models, already described, are usually considered as lattice based models [42].

### 2.3.2.5 Role-based Access Control

Although the concept of RBAC dates back to the ’70s, as already described in 2.3.1.3, this concept was merely the recognition of corporate needs for defining role-based policies. Here, we describe the models resulting from the formalization step of RBAC polices presented twenty-five years later [43].

There are four RBAC models targeting different organizational needs. The first model (RBAC$_0$) is the simplest and is the basis for all other models in the family. RBAC$_1$ introduces role hierarchies and RBAC$_2$ provides constraints. Both these models extend the RBAC$_0$. Lastly, the RBAC$_3$ is the combination of RBAC$_1$ and RBAC$_2$ providing all the available features. Figure 2.6 illustrates the relationships between the four RBAC models.

![Figure 2.6: The RBAC family of models.](image)

The RBAC$_0$ model is known as the basic RBAC model and introduces sessions to support the least privilege principle. The user is able to have several active sessions, each one in conformity with the required role to per-
form a certain task. Privileges in this model are always positive, meaning the policies must follow the ‘closed’ principle. Moreover, the resulting privileges within a session are the union of all associated permissions expressed in the PA relationship. Figure 2.7 shows the RBAC0 model.

Figure 2.7: The RBAC0 Model.

RBAC1 extends the basic model by adding role hierarchies, which were considered as the natural way to represent organizational structure. The notion supporting role hierarchies is that senior roles always include the privileges of a junior role in a transitive manner. Figure 2.8 shows the RBAC1 Model where an additional one-to-many relationship (denoted RH) enables role hierarchy.

Figure 2.8: The RBAC1 Model.

RBAC2 extends RBAC0 and adds constraints to help cope with other principles: prerequisites, separation of duties and cardinalities. All constraints are expressed in the form of a predicate, returning either true or false.

Prerequisites determine role assignment to a user if a certain condition is satisfied. For example, the assignment of the role “Financial Manager” can only be done if the user already has the “Employee” role.

Separation of Duties (SoD) is a fundamental principle in organization structure to prevent fraud. Typically by assigning different roles to task execution and task approval, the chance of fraud is less probable because a collusion between two or more people is required. SoD policies can either be static (administration time) or dynamic (runtime). Static Separation of Duties (SSoD) is obtained by identifying conflicting roles and registering them in the model as constraints to be applied in the UA relationship, this will avoid the assignment of mutually exclusive roles to users. Dynamic Separation of Duties (DSoD) occurs during runtime and it prevents the user from activating conflicting roles at the same time. This type of constraint is usually applied to the relationship between sessions and roles.

Cardinalities are constraints that limit the number of sessions a user might have at the same time, or the number of users assigned to a role. This kind of constraint is usually applied to the relationship between users
and sessions, but it can also be applied to the UA and RH relationships. Figure 2.9 illustrates the RBAC$_2$ where it is noticeable that the broad applicability of constraints results in relationships scattered throughout the model.

Finally, RBAC$_3$ is the model that combines both RBAC$_1$ and RBAC$_2$, resulting in a model where constraints can also be applied to the RH relationship.

2.3.2.6 Open Architecture for Secure Interworking Services

The Open Architecture for Secure Interworking Services (OASIS) is an architectural proposal for supporting healthcare in the United Kingdom. It includes a role-based access control model targeting the facilitation of access in distributed systems. This decentralized approach to access control withdraws the need for the National Health Service (NHS) to centrally manage and register all doctors. Instead, it proposes that hospitals perform role definition and user association, the integration between hospitals will be undertaken by services that map external roles into internal ones during federated accesses. The distributed authorization in OASIS is supported by X.509 certificates.

Unlike other RBAC models, OASIS does not support delegation. Instead, it introduces the concept of appointments, whereby a certain user may issue appointment certificates to other users, enabling them to activate one or more roles [4]. One example of appointment practicality versus delegation, is to consider a visiting doctor in a specific hospital to be required to perform some operation. He or she would require access to the patient’s electronic health record (EHR). Following the appointments rationale, a certificate may be issued to the visiting doctor allowing him, or her, access to EHRs for the visiting period of time. However, if delegation was applied, some resident doctor would have to grant his privileges to the visiting doctor, making a direct privilege association. The appointment strategy does not establish privileges association to a role, like delegation does, but merely allows someone to certify that some user may activate a specific role, following certain activation rules. Moreover, it does not require the user issuing the certificate to have the same role or privileges to be issued.

Another important rule OASIS states is that there are no explicitly defined role hierarchies. This is exactly to avoid uncontrolled privileges association and better support the least privilege principle.
2.4 Linked Data

The term Linked Data was coined by Tim Berners-Lee on a note published in 2006 [5]. The note discussed design issues of the Web and its evolution towards the Semantic Web. Nevertheless, the hype surrounding Linked Data was only born after the talk Berners-Lee gave at the Technology, Entertainment and Design (TED) conference on February 2009 [6]. However, this talk brought confusion on the notion of Linked Data as Open Data.

Linked Data refers to the usage of Web technologies to enable data publishing and querying on the web. It is also one of the forms of enabling the Web to evolve into Semantic Web by linking data together, and also, enriching that data with metadata or ontologies. Linkage and enrichment will provide an increasing semantic context to the data and the expected result is that queries can be easily disambiguated, allowing users to identify the precise thing they are searching for, and also, navigate the relationships of that thing with other things already published.

The technical requirements on Linked Data are: (1) the usage of the Uniform Resource Identification (URI) to identify things; (2) the usage of Hypertext Transfer Protocol (HTTP) as the scheme for URIs; (3) make data compliant with the Resource Description Framework (RDF) standard and (4) link data with other data. However, the Web is mostly used to host Hypertext Markup Language (HTML) documents and URIs generally point to these documents, making data not readily accessible and identified on the Web. The required technique to identify data (or things) instead of documents on the Web is known as URI dereferencing. It consists in using a part of the URI to determine the thing being searched for. Using a browser to navigate to http://dbpedia.org/resource/Portugal, the result will be a dynamically built HTML page showing data on the Web referring to the country Portugal. DBpedia uses the string ‘Portugal’ to search for related data on other Linked Data repositories and responds to the browser with a HTTP redirect pointing to the generated page.

On the other hand, Linked Open Data (LOD) derived from Linked Data and the Open Data Movement. The notion of open data consists in the free usage and redistribution of data. The arguments to support the openness of data lie in the fact that government and scientific data are financed with public taxes and therefore should be publicly available. However, not all open data is Linked Data as well as not all Linked Data is open data. The 5-star classification system for LOD clearly shows that the first thee stars are related with the openness of data, while the last two are related with linked data:

- ★ Available on the Web independently of its format. The only requirement is an open licence.
- ★★ Available in a structured machine-readable format, instead of scanned documents.
- ★★★ Published in a non-proprietary format.
- ★★★★ Use World Wide Web Consortium (W3C) standards to identify and publish the data.
- ★★★★★ Link the published data with other data available on the Web.

This distinction between Linked Data and LOD is a crucial factor to the evolution of research in the healthcare sector. If patients agree to have some of their data openly published on the Web there would be no problems relating to privacy, however the general public still has little awareness about the impacts of harmful...
usage of sensitive data. People are usually more worried with breaches involving financial data, such as credit card numbers, but dire consequences may happen when personal healthcare data is exposed [11, 49].

Linked Data is appropriate to handle both private and public data, since it does not commit to any degree of openness. Moreover, it facilitates the linkage and integration of data at a global scale.

### 2.4.1 Serialization Formats

The basis of Linked Data is the RDF triple, which simplicity and flexibility eases the representation of heterogeneous data and also facilitates data enrichment with metadata or ontologies. Figure 2.10 shows the structure of triples, depicting the connection of resources (subject and object) by a predicate, forming a directed graph.

![An RDF triple.](image)

Figure 2.10: An RDF triple.

Subjects are always identified by an URI, whereas the objects may also be URI identified or simply correspond to RDF literals (e.g. boolean or string). The predicate denotes the relationship between the subject and the object. This structure of RDF triples is what enables heterogeneous data to be supported. A subject or object in a triple may reference an X-Ray image or a Sequence Alignment Map (SAM) file containing sequencing data.

There are several types of serialization formats for triples, but the first recommendation by the W3C was the RDF/XML format in 1999 [46]. Listing 2.1 shows an example of how a reference to the business man Elon Musk on DBpedia can be represented in RDF/XML format.

```xml
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:dbpprop="http://live.dbpedia.org/property/">
  <rdf:Description rdf:about="http://live.dbpedia.org/page/Elon_Musk">
    <rdf:type rdf:resource="http://live.dbpedia.org/ontology/Person"/>
    <dbpprop:name>Elon Musk</dbpprop:name>
  </rdf:Description>
</rdf:RDF>
```

Listing 2.1: RDF/XML example

In March 2001 the W3C submitted a more human readable serialization format, named Notation3 or N3. This format also has the advantage of being less verbose than RDF/XML, thus taking up less bandwidth in data transferences. Listing 2.2 shows the same reference to Elon Musk in N3 format.

```n3
@prefix dbpprop: <http://live.dbpedia.org/property/>.
```

Listing 2.2: N3 example

Other formats have been proposed, like Terse RDF Triple Language (Turtle), N-Triples, TriG, TriX, RDFa and JavaScript Object Notation for Linked Data (JSON-LD). Some formats have more specific applications, like TriG
and TriX that enable support for multiple named graphs on the same file. TriG is essentially Turtle syntax enclosed in curly brackets to denote a graph.

Choosing one of them is related with the required use and the proper support of adopted programming languages or frameworks.

2.5 Private Health Data

Healthcare data from a patient might be stored in Personal Health Records (PHR), Electronic Health Records (EHR) and Electronic Medical Records (EMR). Although, EHR and EMR may sometimes be used interchangeably and as representing the same thing, they are not. The purpose of EHRs is to provide a full coverage of patient health data independently of medical specialities, and also, to allow health data to be accessible across healthcare institutions. However, EMRs are proprietary, difficult to integrate and not owned by the patient; but still, they are the legal record of the medical history of patients [58].

On the other hand, the PHR is kept by the patient, its information also exists in EMRs or EHRs and it may not be digital. It merely serves as a facilitator, quickening the administrative processes in case the patient needs to consult with other practitioners in other healthcare institutions.

As described, the EMRs violate the legal principles that govern data protection laws. The owner of the medical data should be the patient and not the healthcare institution. Although it is usual for the institutions to provide storage and protection for healthcare data, they should not own the data and force patients to request access to their records. The issues raised by this conflicting nature of EMRs and data protection laws must be solved in case historical data is transferred to new standard platforms. However, these data migration activities are not considered in this dissertation, meaning that the presented solution will only target the principles supporting EHRs.

Even though, the actual existing formats play an important role in the healthcare information availability and protection, it is yet more relevant to understand what is in fact considered private data in the health sector. For a long time there have been deontological ethics involved in the practice of medicine defending professional secrecy. This principle states that all collected information about any patient must be held in absolute secrecy, even after the patient had died [16]. For any patient, the purpose of disclosing their personal data is to receive treatment, also known as the ‘primary use’. Any secondary use of data should only happen if the patient consents it, requiring a process that would involve an informed consent. The usual circumstances where the informed consent is used to allow treatment, is when the patient suffers from cancer or a child diagnosed with cancer requires the parents permission to be treated.

It is clear, from the deontological ethics, the ‘primary use’ and data protection laws, that all data collected by medical practitioners is private and should be kept in secret. In some occasions, in an event of cancer or other rare diseases, the treatment the patient must undergo will have a considerable death risk, thus requiring the patient to make an informed choice about his future. These are the circumstances where doctors, faced with information scarcity, may seize the opportunity to motivate the patient into allowing his data to be used
for research purposes.

Nevertheless, the patient data not directly related with diagnosis and treatment, like name and social security number, should never be disclosed even for research purposes. Removing personal identification data from medical records is known as de-identification or anonymization. The standard methods for de-identification are defined in the United States (US) through the Health Insurance Portability and Accountability Act (HIPAA) [51]. In Europe there are no equivalent standards defined.

In conclusion, personal identification data should never be disclosed and may be anonymized to prevent direct identification. Health data is always private existing merely to enable patient treatment (the primary use). Any secondary usages imply the informed consent from the patient.

## 2.6 Ongoing Research

The need for accessing private health data in research has mainly to do with phenotype information. Supposing a certain group of individuals has a genetic predisposition for developing lung cancer. It is then, essential to determine what are the characteristics leading to the development or non-development of cancer in subsets of these individuals. These characteristics may include age, smoking or prenatal smoking, and even facts like their mothers receiving artificial hormone diethylstilboestrol during pregnancy to prevent miscarriage.

By integrating data across several medical specialities and analysing it, is thus possible to improve prognosis, prevention and treatment of rare diseases. Different research groups have been discussing the subject of privacy and progress. The Workshop on Establishing a Central Resource of Data from Genome Sequencing Projects, held on June 2012 has started the formal debate and also brought forward some discussion on the whole genome sequencing [34]. On June 2013, a Global Alliance composed of more than 70 institutions worldwide was formed. The purpose of this alliance is to develop a global effort to enable secure sharing of genomic and clinical data [21]. An important effort has been done through the European Life-Sciences Infrastructure for Biological Information (ELIXIR) consortium that has been implementing an European research infrastructure. Although it has been initially focused on biological information, one of the main stockholders, the European Molecular Biology Laboratory (EMBL) has already joined the global alliance and privacy issues related with clinical data are expected to be addressed in ELIXIR.

Despite the number of institutions involved, the published integration strategies amount to four different ones. Although five approaches may also be referred, we do not consider the “Registered user system” as a concrete alternative, but as merely a reference to authorization and auditing, not leading to infrastructure, administrative and management impacts when compared with other published strategies. These strategies are not mutually exclusive and may be applied simultaneously in a global solution.

### 2.6.1 Open Access

With open access, patients simply donate their data through an informed consent briefing them on the risks involved. The data, although anonymized, will be available online and downloadable.
The advantages with this strategy are strictly related with the absence of technical complexity for its implementation. All the technology to support the indiscriminate access to data on the Web is already freely available. Also, the time required to implement this kind of access would be insignificant when compared to any other strategy. This low-cost and rapid implementation characteristics would promptly lead universities, medical and research institutions to take advantage of the information to achieve new insights.

However, given the indiscriminate access to sensitive data, it is expected that the general public will not be much receptive to this approach. Studies show that a considerable amount of people will not seek medical help due to privacy concerns, specially in cases of mental illness and sexually transmitted diseases [12]. Without public trust, there will be no scientific progress in healthcare research, and more importantly, people will be more isolated and without any care or treatment for their illness.

Other disadvantage of open access is the high duplication of effort and data, since institutions would make data available, that will be downloaded, integrated and stored in other institutions. This will likely lead to several institutions performing exactly the same data integration and analysis, instead of developing a coordinated effort to enhance knowledge discovery. Moreover, because data would be anonymized there will be no confidence in establishing relationships between newly obtained data and already existing data. For example, new psychiatric data might be available for a specific individual, whose data on several other medical specialities is already available. Relating the previous data with the new one will be unlikely.

The same happens when considering the data quality of the datasets resulting from integration. It is likely that the inability to identify individuals and their data, will result in the same data being integrated more than once, thus, invalidating the statistical results used for data analysis.

### 2.6.2 Streamlined Access

The streamlined access is similar to what has been common practice in the healthcare sector. The purpose of this strategy is to make clinical and genotype data more rapidly available by changing and simplifying the administrative procedures for obtaining access to clinical data.

This strategy will hardly introduce any integration benefits, by perpetuating the already existing and flawed system.

### 2.6.3 Research Commons

The research commons is based on the pre-authorization of patient data for research purposes, through informed consents. It will require a central authority to authenticate researchers and other participants and clinical data will be available if authorized access is granted.

In this scenario, data might be distributed and accessible though federated queries, or it could be retrievable from different servers for consequent integration. By not imposing a central administration and management of the infrastructure and data, this is the better candidate for implementing a fully distributed system, based
on the recent open standards. Moreover, it enables distributed processing if the architectural design considers this feature.

This strategy does improve the current status of healthcare services by enabling the EHRs to be available across several institutions, and also, it would enhance research by allowing clinical data to be easily integrated and analysed. A distributed system is what better fits the distributed nature of clinical data and its decentralized management. Moreover, it would not require central management, since each institution would perform the necessary maintenance and management of the hardware, software, data and users in order to keep the system working.

2.6.4 Data Analysis Servers

Data analysis servers will enable a simplified and convenient way of integrating data and controlling access. However, they will not provide full access to clinical data but to previously computed results based on that data. For example, researchers would not have access to the whole genome of a patient, but rather to its genotype numbers or p-values.

They could also provide means to enable distributed processing, but because this would be a more central solution, and not fully distributed in each institution, this will likely lead to bottlenecks in demanding data analysis procedures. As well as the research commons, this solution will also benefit from centralized authority, but will require a central management effort and a dedicated team for its maintenance.

Additionally, it considers a special kind of access to medical records, which would be stored in another family of servers. This kind of access and its infrastructure requirements would make the system more expensive and more troublesome to manage.

Because this is a centrally managed infrastructure, either standards for data would be required from participants, or the effort of transforming all kinds of data into a standard structure would be supported by the team responsible for managing and maintaining these systems.
This chapter will detail the subjects underlying the requirements for a globally distributed healthcare system. We start with the definition of an analysis framework, in Section 3.1. Next, in Section 3.2, the fundamental requirements are presented, as well as the consequent architectural designs. Section 3.3 introduces a part-by-part approach to evaluate the most adequate strategies and tools to solve the nuclear issues of identification and authentication. Section 3.4 will discuss the known anonymization models. Lastly, Section 3.5 succinctly presents the analysis results.

### 3.1 Analysis Framework

The design of a global healthcare system with intricate security requirements demands a formalization step so that every dimension of the problem can be properly analysed. The broad scope of subjects involved, ranging from legal principles to technical details will require some degree of decomposition and segregation into different classes.

To contextualize every issue within problem at hand, we propose the use of a framework comprised of six dimensions. This will breakdown the complexity of the system into more easily analysable parts, enabling a more detailed level of concerns. Moreover, it will provide a more concise terminology and semantics throughout the reminder of this thesis. Table 3.1 details the proposed six dimensions framework.

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Elements</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scope</td>
<td>Ethical, Legal, Management</td>
</tr>
<tr>
<td></td>
<td>Administrative, Design</td>
</tr>
<tr>
<td></td>
<td>Technical</td>
</tr>
<tr>
<td>Scale</td>
<td>Local, Multi-institutional, Global</td>
</tr>
<tr>
<td>Access</td>
<td>Private, Confidential, Shared</td>
</tr>
<tr>
<td>Time</td>
<td>Public</td>
</tr>
<tr>
<td>Procedural</td>
<td>Events</td>
</tr>
<tr>
<td>Operative</td>
<td>Processes, Activities, Tasks</td>
</tr>
<tr>
<td></td>
<td>Stakeholders</td>
</tr>
</tbody>
</table>
These dimensions are aligned with the Zachman Framework, allowing a more precise distinction as to the what, how, where, who, when and why of each of the addressed issues [56, 57]. For example, if one asks why laws are made to consider privacy principles, then the answer lies in the “scope” dimension of the framework, specifically, because it is an ethical principle intrinsic to the belief system of any person.

The scope dimension provides a layering of the concerns involved in the analysis. It follows the rationale that some concerns are of interest to society, while others are fundamental aspects of how institutions are organized. Hence, this dimension enables layering concerns from society level down to the individual level, but also reflects the usual responsibilities within an institution. Below, is presented the definition of each of the elements that constitute the scope dimension:

- **Ethical**
  The concerns directly related with society and under the influence of cultural aspects. Examples of ethical concerns are privacy, equity, liberty, justice and safety.

- **Legal**
  Legal concerns are the institutionalization of rules that govern social behaviour and the establishment of the consequences or punishments if rules are not obeyed.

- **Management**
  Management is composed by the planning, control of execution and corrective decisions targeting a well defined objective. An important factor that is commonly referred as being part of management is motivation, however this assumption is incorrect\(^1\). Examples of management concerns are: to reduce system complexity, to minimise operating costs, to improve product quality, etc.

- **Administrative**
  Administrative concerns are those directly related with the operation of an institution. For example, different roles in the administrative level are accountants and human resources officers. A concrete case in the healthcare system, is that of informed consents. Producing, explaining and signing an informed consent is an administrative task.

- **Design**
  Design concerns are those considered before implementation, but addressed during or afterwards the study and analysis of a system. It provides a formal description of a system, identifying its composing parts and respective interactions from different perspectives. To the several perspectives and abstractions of a system expressed through design, we call architecture.

- **Technical**
  Technical concerns are those associated with concrete domains of knowledge requiring a high level of

---

\(^1\) The author considers management to be a natural activity not exclusive to humans. Moreover, it considers motivation as an element of leadership. For more detailed information and demonstrations please read the Appendix A.
specialization from the individuals addressing them. For example, performing eye laser surgery or implementing consistent backup in a distributed environment are technical concerns.

The scale dimension represents the organizational segregation level to which the data is stored and handled. At the local level, only a single institution is considered in the analysis, design or development. This will enable handling concerns related with the scale of the system in a bottom-up manner, analysing the system from the local scale to the global one.

The access dimension represents the different data accessibility levels within the healthcare system from the patients’ perspective. The private level represents the exclusive individual knowledge about a piece of information. Confidential information is accessed by two or more identifiable individuals. The shared level happens when information is no longer associated with identifiable individuals, but is accessible to anyone responsible for specific tasks. The public level implies that no access restrictions are applied to the information, thus, being available to everyone. Figure 3.1 illustrates the access levels and the events triggering the transition between them.

![Figure 3.1: Access Levels and Transitions.](image)

Information in the initial state is only held at private level and at this phase, an individual is not yet a patient. When an individual participates in a medical appointment it will become a patient and his information will also be known by the practitioner assisting him. When information is confidential the patient knows exactly who else has access to his data. If the patient enrols in a clinical study his data will be shared among researchers for a specified purpose. This transition always involve an informed consent and may also encompass a donation of data. This last aspect usually refers to donating data for research or to science [10, 54]. Lastly, if the patient decides to publish his data then it will no longer have any kind of control over it.

The order by each of the access levels happens in the system is arbitrary, therefore an individual may simply decide to publish his genome without ever being considered a patient in a healthcare institution. Also, the informed consent is reversible, meaning that the patient has the right to revoke the access at any time.

The time dimension is merely composed of events. These will determine the transitions in access levels, the assignment of certain tasks to different stakeholders or the triggering of specific activities within a process.
The procedural dimension addresses the level of detail in operations leading to the achievement of a goal or objective. These correspond to processes, activities and tasks. Processes are the series of actions that traverse the whole organization. Activities cross whole departments and tasks are usually executed within a department. Figure 3.2 illustrates this separation in procedural levels.

![Figure 3.2: Processes Activities and Tasks.](image)

The operative dimension has several types of stakeholders, each one of them with different interests in the system. Table 3.2 enumerates the stakeholders of a global healthcare system and describes their interests within the system.

<table>
<thead>
<tr>
<th>Stakeholder</th>
<th>Description</th>
<th>Interest</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patients</td>
<td>Individuals seeking medical treatment.</td>
<td>To receive the best treatment with the least amount of risk involved.</td>
</tr>
<tr>
<td>Practitioners</td>
<td>Individuals providing medical treatment.</td>
<td>To provide the best treatment according to the patients’ characteristics and considering up-to-date knowledge.</td>
</tr>
<tr>
<td>Researchers</td>
<td>Individuals providing knowledge discovery and innovation.</td>
<td>To provide scientific evidence of new discoveries that enhance quality of life. To access the largest amount of existing knowledge to avoid “reinventing the wheel”.</td>
</tr>
<tr>
<td>Government</td>
<td>Individuals elected to govern national policies.</td>
<td>To provide the most appropriate policies so that diverging interests may be covered with little compromise and conflict, achieving sustainable progress in the interest of all.</td>
</tr>
<tr>
<td>Insurance Companies</td>
<td>Organizations providing financial assistance during medical treatment.</td>
<td>To minimise the cost of providing financial assistance. To have no indemnification processes resulting from data breaches.</td>
</tr>
<tr>
<td>Security Managers</td>
<td>Individuals responsible for defining and maintaining the security policies.</td>
<td>To minimise the effort and complexity of managing access privileges.</td>
</tr>
<tr>
<td>Students</td>
<td>Individuals providing assistance in knowledge discovery and innovation.</td>
<td>To have no hassle in accessing the tools and knowledge required to provide research assistance.</td>
</tr>
</tbody>
</table>

### 3.2 Requirements

Performing requirements elicitation for healthcare at a global scale will require several years to be accomplished, due to the size of the system and to the conflicting requirements each healthcare institution may have established. In this thesis the specification of the fundamental requirements is grounded on the legal and security principles governing healthcare data, while using the presented analysis framework to provide the different
3.2. REQUIREMENTS

architectural perspectives.

We start by presenting a list of requirements from all the already presented information. This list will evolve as analysis deepens.

Req1 All data related to patients is considered private and owned by the patient.

Req2 In order to provide treatment (primary use) a partitioner must create, access and change patient data.

Req3 Any secondary usages of patient data must be consented by the patient.

Req4 Patients can access and update their personal data.

Req5 Patients may revoke access privileges to specific practitioners or purposes.

Req6 Management of access control policies should be possible at local scale.

Req7 Patients should be able to delegate control over their data in case they expect to become incapacitated.

Considering the seven mentioned requirements, it is possible to design a simplified diagram about what is expected from the system at a local scale. Figure 3.3 illustrates this simplified view.

![Figure 3.3: Architecture for requirements 1 to 7.](image)

From these seven requirements and the simplified local architecture, it is possible to realize that roles are required to perform access control in the system. The layer controlling access should be intermediating all the roles in the system and the data. From Req6 it is possible to conclude that an access control model allowing management is required. This fact narrows down the type of access control model to RBAC.

Requirements Req3, Req5 and Req7 need the administrative scope to be included. It should be defined an activity that consults the patient for informed consent in the case of Req3. From Req7 we deduce that patients should have access to the subset of roles accessing their data in order to revoke access to certain practitioners. However, to deny access to specific purposes an administrative activity must be designed into the system. As for Req7, supposing patients would access the system with credentials like username and password, even if they
would let someone know their credentials, changes in access regarding purpose still require an administrative activity (derived from Req5).

Based on already discussed and described evidences, a new set of requirements is considered:

**Req8** Data should be anonymized to prevent or hamper patient identification, thus enabling scientific studies to be conducted. This is applicable even if the patient enrolls in any clinical study.

**Req9** Access restrictions to patient data can be bypassed in case of an emergency.

**Req10** Accountability for perpetrators of any violation.

**Req11** Breach detection.

By adding these new four requirements, the local architecture changes significantly, as shown in Figure 3.4. A more detailed local architectural diagram is available in the Appendix D. Access control is less restrictive for emergencies, depicted by the thinner part of the access control layer. It denotes that any health professional will bypass some of the access control mechanisms in the event of an emergency.

![Figure 3.4: Architecture for requirements 1 to 11.](image)

Practitioners delivering emergency care services will be identified and authenticated but no restrictions will be applied while accessing the patient record. Therefore, the emergency role will have read access to sensitive areas of the patient record (e.g. allergies, drug or alcohol addictions, diagnosed diseases). Nonetheless, every practitioner performing an emergency must go through the audit trail layer, implying that their actions will be recorded into the system. This will discourage them to use the emergency role to peek into patient data. Moreover, auditing authorized accesses will permit accountability to be implemented, although it also requires administrative tasks and activities to be introduced into the system. The system, at its technical scope, can only identify who has performed certain tasks that may have violated the data protection principles. But still, accountability requires inquiring the responsible parties with the purpose of determining if intentional or unintentional harm was done.
The anonymization layer required for researchers should prevent the patient name, birthday and full address to be known. Typically, to perform research only the age of the patient and the state or some other coarse geographical reference is needed.

The breach detection requirement exists due to legal principles demanding that patients should be informed in case their personal data was disclosed. Moreover, precautionary measures must be taken into account to detect any fraud that may have resulted from the breach. It is up to the patients to judge if they should take legal action towards the healthcare institution or its practitioners.

Detecting a breach may be impossible while it is happening, and even some time later it might still go undetected. Breaches are hard to detect because the intruders usually resort to social engineering techniques. These techniques generally include tricking authorized users to perform certain tasks or quietly obtaining their authorization credentials. Therefore, the alert layer should be configurable by the security managers to lookup certain irregular accesses to data, for example, unknown computers accessing the system.

### 3.2.1 Multi-Institutional and Global Scales

By introducing the interoperability between two different healthcare institutions other concerns require attention. The process of patients accessing their data should be transparent even if the patient is visiting two different practitioners in two distinct institutions. Because the system is distributed, data may not be integrated when the patient accesses the system of a single institution. It may be easier, for the patient, to simply access the system of the institution where he knows the data resides. This raises the issue of identifying the patient in a distributed system. When only the local scale was considered, security managers could simply create a user with the patient role and let him define his own password.

For practitioners to access medical records from a specific patient, stored in another institution, they should just require authorization from the patient. This is a scheme that follows the OASIS architecture proposition of appointments. By allowing the patient to issue a certificate that allows a certain practitioner to access his data, it would not only address the issue of only requiring the patient’s permit to access the data, it will also reduce management and administrative activities in each of the institutions.

Nonetheless, it is essential for practitioners to have means of being identified between institutions. This raises the question if identification should be provided centrally, by an National Health Service (NHS). Again, as the OASIS architecture proposes, practitioners identification should be decentralized, this leads to a more faster process of creating an identification. However, the OASIS identification model may be limited for global use, as explained in the following section detailing identification and authentication issues.

### 3.3 Identification and Authentication

As seen in the requirements section, as soon as the scale of analysis changes from local to multi-institutional or global, it raises the question of how to manage identification between institutions. Since authentication
relies on identification, these two principles cannot be analysed separately. These concerns are applicable to patients, practitioners and researchers, thus requiring a more detailed analysis in each case.

### 3.3.1 Practitioners and Researchers

The identification model, proposed in OASIS, for practitioners is based on Role Membership Certificates (RMC). It uses a three level hierarchy model composed of the NHS, the healthcare institutions and the practitioners. The NHS issues RMCs for healthcare institutions, which in their turn issue RMCs for practitioners. This model has the advantage of having a certification chain that assures the legitimacy of any individual as being a practitioner, thus enabling not only proper identification, but also an authentication mechanism. Also, it provides central revocation if, for any reason, the practitioner must have its licence withdrawn.

However, this model may break if taken outside of a single country, in this case OASIS was proposed for the United Kingdom NHS. If a Portuguese institution needs to authenticate some practitioner from the UK, it could fail to identify the certificate validity due to a untrustworthy root authority. This could be solved if the Portuguese institution directly trusted the UK NHS. But, this would imply that every healthcare institution in the world would directly trust each of the others countries NHS. This is somewhat impractical and would require some organization, like the World Health Organization (WHO) to be considered the root certificate authority (CA), which would then issue certificates for each of the NHSs.

Another alternative, is to have a complete distributed identification mechanism for practitioners. This will remove the need for a tree of issuing authorities and each of the practitioners could have their own self-signed certificate. This raises the issue of potentially everyone being a practitioner by creating its own self-signed certificate.

WebID, is a protocol that uses the Friend of a Friend (FOAF) ontology and the Secure Socket Layer (SSL) to enable users to manage their own identification on the Web without the need of a CA [44]. This protocol does not remove the threat of impersonation by itself, but since it is based on FOAF it can be used to build a Web of Trust (WOT), which will enable authentication. The WebID protocol relies on SSL to establish an encrypted channel of communication together with a two-way identification and authentication handshake, in opposition to typical SSL utilization where only the server certificate can be authenticated. For the server to rely on the identify provided by the client it requires more than just a self-signed certificate. This is where FOAF complements SSL by enabling the server to check the user profile and relationships. In order to have a trustful relationship in FOAF, the client cannot only state that he knows someone that might serve as a trust anchor. The latter must also state that he or she knows the client. This bidirectional validation scheme based on trust, can be easily accomplished on the healthcare sector if the institutions implement management and administrative processes to simplify the production and storage of public WebID profiles for practitioners, thus forming a distributed authentication mechanism.

Healthcare institutions are likely to interact amongst themselves, thus creating a WOT that would spread from regional to national, leading to a global scale network of trust. Practitioners would directly benefit from this by being identified using transitive closure.
All the assumptions for practitioners are also valid for researchers. Some of the researchers may also practice a medical speciality, thus simplifying the process of identification and authentication. Nonetheless, researchers typically are affiliated with some institution, which can easily participate in a WOT. Practitioners that are also researchers, do raise conflict of interest issues in the system. For example, a practitioner/researcher (PR) may use his access to patients’ records to conduct research exposing private data to other elements on the research team. These cases can only be detected through audit trail and can be only dealt with at the legal and management scopes of the system.

Figure 3.5 exemplifies how the WebID helps in enabling a distributed authentication mechanism to support healthcare data to be accessible across multiple institutions. From step 1 to step 3 of the protocol, the laboratory is able to determine if the sent WebID is in fact a practitioner. From step 4 to 6 the laboratory determines if the sender is in fact the holder of the private key, thus concluding that the sender must in fact be the owner of the presented WebID which is confirmed by the hospital as being a active practitioner. The lab results are then sent to the practitioner encrypted with its public key, guaranteeing that only him or her may be able to read them.

![Image of WebID utilization diagram](image)

Figure 3.5: Example of WebID utilization.

### 3.3.2 Patients

The law establishes that patients may access their data to perform corrections. Since patients may seek medical help anywhere it implies that their data should also be accessible globally. Moreover, if a patient enrols in clinical studies, his data is expected to be available to every research team working on the same subject of those clinical studies. This raises two main concerns regarding patient identification, one targets the authentication of the patient in every medical system, the other targets medical record uniqueness. To easily distinguish between these two paradigms, we will call the global authentication identifier an Universal Patient Authentication Identifier (UPAID), and the identifier for uniqueness in clinical records we call Universal Healthcare Identifier (UHID).

Providing means of authenticating patients deals mostly with legal, management and administrative
scopes. This arises from the lack of patient affiliation to a specific institution, part of the healthcare system, that can prove that the owner of certain credentials is in fact the patient. The suggested WebID strategy for practitioners is not directly applicable to patients due to the lack of a WOT.

There is, however, a proposal that may evolve into a global authentication framework for citizens, namely the European eID, in which the Portuguese identity card “Cartão do Cidadão” is included. This proposal is based on the Public Key Infrastructure (PKI) standards, alongside the WebID, but it relies on a smart card to hold the private key and personal data. This enables a two-factor authentication method based on possession and knowledge. Moreover, it enables portability and security of the private key, benefiting from the trust network that each national civil registry service brings to the whole system.

It is suffice to say that the technologies and protocols for establishing a global citizen authentication platform are already available and are known by governing parties. The real problem is the scale of modernizing identification and authentication mechanisms, leading to considerable amounts of time and financial investments to have every citizen in this new paradigm. Considering these non-technical requirements for patient authentication, the focus of this analysis will be the UHID instead of the UPAID. Note that this will have impact in the design of a proof of concept, since patients will be left out of the system’s architecture as direct actors or users.

The American Society for Testing and Materials (ASTM) defined a standard guide for implementing UHID, the ASTM E1714 [2]. This guide defines thirty properties that should be considered while designing an universal identifier for patients. This form of identification is fundamental for achieving data integration, record uniqueness, statistical integrity and incident prevention. This last point is critical given the harm resulting from identification errors in healthcare, these estimates range from 44000 to 98000 deaths per year in the US alone [53].

Some proposals based on the E1714 guides have emerged, most of them have considerable issues. For example, the Enhanced Social Security Number (ESSN) proposes extending the SSN with alphanumeric characters and check digits, however, it raises the issue of indirect identification by using the SSN part of the identifier to determine the patient identification in any organization that stores SSNs, moreover, in a global healthcare system it forces every other country to follow the US numbering scheme [33]. The most recent proposal introduces the voluntary act of universal identification as opposed to the governmental and centralized model, called the Voluntary Universal Healthcare Identification (VUHID) [22]. This identifier is a 32 digit number structured in three different parts: a prefix of 16 digits; a one digit delimiter; a 8 numbers as check digit and a 7 digit privacy class. The issue with VUHID is that a patient should have only one Open Voluntary Identifier (OVID), in which the privacy class is all zeroed, and may have several Private Voluntary Identifiers (PVID). This strategy completely invalidates global integration of data considering uniqueness principles, fundamental to perform most of the statistical analysis or data mining processes to enable research.

Nonetheless, if we consider the ideal UHID scenario, the obvious fact is that it should have irreversible properties like those used in hash functions. Considering that the Population Reference Bureau has estimated $107.6 \times 10^9$ to be the number of *Homo sapiens* to have ever lived and that the current highest estimate from the United Nations for world population in 2100 is $16 \times 10^9$ [38, 50], we can clearly conclude that the 256-bit
Secure Hash Algorithm (SHA), denoted as SHA-256, will be more than enough to hold identifiers for all historical medical records. If we consider the current yearly birth rate, which is approximately $1.3 \times 10^9$, SHA-256 will provide us with $8.9 \times 10^{67}$ years of support just considering its result space, but not its collision probability. More specifically, this time frame for SHA-256 applicability is an optimistic prevision, that requires no collisions to occur.

A proposal for a possible European UHID points to the use of the SHA-2 algorithm and suggests that SSNs are used as input value for the hash function in order to provide uniqueness and anonymity. At the European scale the same article proposes the addition of personal patient data (e.g. first name, date of birth, etc.) or biometric values as inputs [40]. The usage of the SSN may be applicable to Europe, but it leaves out less developed countries in which research studies may be conducted.

Given the irreversible properties of hash functions, there is no reason why the full name, date of birth, gender and blood type should not be used as inputs for producing an UHID (Figure 3.6). Collisions on these inputs are more likely on cultures where the parents rarely know the exact date of birth of their children, or naming traditions show a regular adoption of a small set of names. Notwithstanding the foregoing, the inputs for the hashing function should be natural and not circumstantial, leaving out residential addresses, phone numbers and also the SSN. The latter should be excluded due to naturalization rights for foreign citizens, since different identifiers would be produced for the same patient in distinct countries.

The use of the SHA-512 secure hash function is recommended for 64-bit platforms and, if needed, it may be truncated to 256 or 224-bit [48].

### 3.4 Anonymization

Anonymization is the process of removing any link between the patient data and the patient, so that it is impossible to identity someone from published data. This technique is not useful for primary use of healthcare data. For example, some research studies reveal facts useful to treat some of the patients whose data is included in the research. If the data was anonymized before conducting research, then it will be impossible to identify the patients who would benefit from these discoveries.

De-identification, on the other hand, provides re-identification of patient, thus being reversible. There is some common confusion regarding both definitions when it comes to the Health Insurance Portability and Accountability Act (HIPAA) guidance methods for de-identification [51]. It is typically assumed that HIPAA has defined anonymization guidance methods, when in fact they are considering re-identification as a possibility.
The irreversible assumption related with anonymization is a misleading concept that fails to be applicable in practical terms. Because most data is published with educative purpose and uses descriptive statistics to summarize a sample, it always reveals some attributes to classify the results. These attributes may be used to link two or more different data sources and achieve re-identification.

Data protection laws establish that indirect identification is a concern, following these principles the HIPAA has defined two methods for de-identification that demonstrate concerns regarding re-identification:

1. **Expert Determination** - consists in the application of statistical or scientific principles to render information as not individually identifiable.

2. **Safe Harbor** - consists in 18 identification attributes to be removed from clinical data, from which the simplest ones are shown in table 3.3.

Table 3.3: HIPAA identification attributes (excluding 2 and 3)

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Names</td>
</tr>
<tr>
<td>4</td>
<td>Telephone numbers</td>
</tr>
<tr>
<td>5</td>
<td>Fax numbers</td>
</tr>
<tr>
<td>6</td>
<td>Email addresses</td>
</tr>
<tr>
<td>7</td>
<td>Social security numbers</td>
</tr>
<tr>
<td>8</td>
<td>Medical record numbers</td>
</tr>
<tr>
<td>9</td>
<td>Health plan beneficiary numbers</td>
</tr>
<tr>
<td>10</td>
<td>Account numbers</td>
</tr>
<tr>
<td>11</td>
<td>Certificate/license numbers</td>
</tr>
<tr>
<td>12</td>
<td>Vehicle identifiers and serial numbers, including license plate numbers</td>
</tr>
<tr>
<td>13</td>
<td>Device identifiers and serial numbers</td>
</tr>
<tr>
<td>14</td>
<td>Web Universal Resource Locators (URLs)</td>
</tr>
<tr>
<td>15</td>
<td>Internet Protocol (IP) addresses</td>
</tr>
<tr>
<td>16</td>
<td>Biometric identifiers, including finger and voice prints</td>
</tr>
<tr>
<td>17</td>
<td>Full-face photographs and any comparable images</td>
</tr>
<tr>
<td>18</td>
<td>Any other unique identifying number, characteristic, or code, except as permitted by re-identification</td>
</tr>
</tbody>
</table>

Summarizing the second identification attribute, it relates to geographical divisions smaller than a state, like ZIP codes and street names, etc. It also mentions an exception regarding a threshold of 20000 habitants for showing the first three digits of the ZIP code. The third identification attribute relates to time and suggests the removal of any date elements with exception of the year. It also mentions the creation of a specific class for citizens with more than 89 years old. The second and third identification attributes defined by HIPAA, do reflect concerns regarding indirect identification, which are consistent with the usual inference by linkage of different data sources or previous knowledge.

Because it is impossible to determine *a priori* how someone may use his lifelong acquired knowledge and arrive at conclusions that lead to privacy exposure of others, the anonymization of data is somewhat an utopia. Nonetheless, several methods targeting anonymization have been proposed. The discussion of these methods requires the previous definition of the following concepts:

- **Key attributes (KA)** - the attributes that can directly identify a person, for example name, address and phone number;

- **Quasi-identifiers (QID)** - the attributes that linked together can potentially identify a person, for example, ZIP code, date of birth and gender;
• **Sensitive attributes (SA)** - the attributes that can lead to social discrimination, for example, clinical information, tax information and salary;

• **Equivalence class** - the set of records having the same values for the quasi-identifiers.

### 3.4.1 k-Anonymity

Proposed by Latanya Sweeney in 2002 [45], it starts by addressing re-identification via linking data sources, and provides a first person real case example. Sweeney was able to identify William Weld (governor of Massachusetts in 1997) on the medical records published by the Group Insurance Commission (GIC), by linking the GIC data with the voter registration list for Cambridge Massachusetts. Figure 3.7 shows two hypothetical data sets, similar to those Sweeney had access to. The green shaded lines represent the two linkable records that allow anyone to conclude that Alice suffers from ovarian cancer.

The example demonstrates that even without key attributes in the one of the datasets, it is still possible to link them via quasi-identifiers and discover some sensitive information. The k-Anonymity model determines that the information of each person contained in the dataset cannot be distinguished from at least \( k - 1 \) individuals whose information also appears. Simply put, for each personal record in a dataset, there should be at least one or more records with exactly the same quasi-identifiers.

The two methods proposed to achieve k-Anonymity are generalization and suppression. The former replaces quasi-identifiers values with less specific ones until the desired \( k \) is achieved. The suppression method occurs when generalization causes complete information loss, which happens typically with the distribution outliers. Figure 3.8 shows a generalized dataset where each of the differently shaded records show a different set of \( k \) individuals. Generalization was achieved by changing the age to birth year and removing the last digit from the ZIP code.

The k-Anonymity establishes that if someone knows a quasi-identifier from one individual, it will not be able to make an assertion with a confidence greater than \( 1/k \). However, this model has limitations, some of which were published by Sweeney herself, namely the unsorted matching attack, the complementary release attack and the temporal attack. The first refers to the publishing of similar datasets having the same exact records in the same exact order; the second refers to the publishing of datasets where the quasi-identifiers may have different generalizations, but complement each other; the temporal attack refers to the adding of new records that will stand-out from previous releases of the same dataset.
<table>
<thead>
<tr>
<th>Birth Year</th>
<th>Gender</th>
<th>ZIP</th>
<th>Diagnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>1965</td>
<td>M</td>
<td>0214*</td>
<td>short breath</td>
</tr>
<tr>
<td>1965</td>
<td>M</td>
<td>0214*</td>
<td>chest pain</td>
</tr>
<tr>
<td>1965</td>
<td>F</td>
<td>0213*</td>
<td>hypertension</td>
</tr>
<tr>
<td>1964</td>
<td>F</td>
<td>0213*</td>
<td>obesity</td>
</tr>
<tr>
<td>1964</td>
<td>M</td>
<td>0213*</td>
<td>chest pain</td>
</tr>
<tr>
<td>1964</td>
<td>M</td>
<td>0213*</td>
<td>obesity</td>
</tr>
<tr>
<td>1967</td>
<td>M</td>
<td>0213*</td>
<td>short breath</td>
</tr>
<tr>
<td>1967</td>
<td>M</td>
<td>0213*</td>
<td>chest pain</td>
</tr>
</tbody>
</table>

Figure 3.8: Generalized dataset where \( k = 2 \).

### 3.4.2 I-Diversity

The I-Diversity model addresses the issue of diversity within each quasi-identifier. This issue leads to what is known as the homogeneity attack on k-Anonymity [31]. Supposing that Eve knows that Bob has 25 years old and lives on ZIP code 02131, if she receives a dataset like the one on Figure 3.9 then she will conclude that Bob suffers from heart disease.

<table>
<thead>
<tr>
<th>Health Insurance Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>QID</td>
</tr>
<tr>
<td>-----</td>
</tr>
<tr>
<td>0213*</td>
</tr>
<tr>
<td>0213*</td>
</tr>
<tr>
<td>0213*</td>
</tr>
<tr>
<td>0213*</td>
</tr>
</tbody>
</table>

Figure 3.9: Homogeneity issue

Another known attack that motivated I-Diversity is the background knowledge attack. This attack can be simply summarized as inference by exclusion, in the sense that the attacker by knowing some fact, it may discard some values from a quasi-identifier and arrive at some sensitive information. For example, Eve knows that Dave practices several fitness sports. Eve also knows that Dave lives in ZIP code 02132 and that he is 35 years old. From the Figure 3.9 she would conclude that Bob has prostate cancer, because heart disease would be much less probable.

There are three known methods to achieve I-Diversity:

- **Probabilistic I-Diversity** - each sensitive attribute value must have \( l \) distinct values;

- **Entropy I-diversity** - the entropy of the sensitive attribute values in each equivalence class is at least \( \log(l) \);

- **Recursive \((c,l)\)-diversity** - ensures that the less frequent sensitive attribute values are not too scarce and that the most frequent sensitive attribute values are not the most common. This is done for each equivalence class by recursively ensuring that the frequency of the most frequent sensitive value is less than the summation of all other less frequent sensitive values.
### 3.4.3 t-Closeness

The t-Closeness model was proposed to solve skewness attacks and similarity attacks on l-Diversity [30]. The former happens when the distribution of sensitive attribute values differs from the global distribution of the dataset, while the latter, are based on the semantic similarity of sensitive values within an equivalence class. This model states that the distance between the distribution of a sensitive attribute value in an equivalence class should not have more than a threshold \( t \) from the distribution of the same attribute on the whole dataset.

### 3.5 Summary

Considering the identified requirements and the six dimensions framework, we summarize their intersection on Table 3.4, followed by a short revision of some architectural considerations.

<table>
<thead>
<tr>
<th>Requirement</th>
<th>Scale</th>
<th>Scope</th>
<th>Access</th>
<th>Time</th>
<th>Procedural</th>
<th>Operative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Req1</td>
<td>Any</td>
<td>Management Administrative Design Technical</td>
<td>Confidential</td>
<td>Each appointment</td>
<td>All</td>
<td>Patients Practitioners Security Managers</td>
</tr>
<tr>
<td>Req2</td>
<td>Any</td>
<td>Management Design Technical</td>
<td>Confidential</td>
<td>Each appointment Each analysis</td>
<td>Activities Tasks</td>
<td>Practitioners Security Managers</td>
</tr>
<tr>
<td>Req3</td>
<td>Any</td>
<td>Administrative</td>
<td>Shared</td>
<td>Enrolment</td>
<td>All</td>
<td>Patients Practitioners Researchers</td>
</tr>
<tr>
<td>Req4</td>
<td>Any</td>
<td>Management Design Technical</td>
<td>Any</td>
<td>Anytime</td>
<td>Task</td>
<td>Patients Security Managers</td>
</tr>
<tr>
<td>Req5</td>
<td>Any</td>
<td>Design Technical</td>
<td>Confidential Shared</td>
<td>Anytime</td>
<td>Task</td>
<td>Patients</td>
</tr>
<tr>
<td>Req6</td>
<td>Local</td>
<td>Management Administrative Design Technical</td>
<td>Confidential Shared</td>
<td>Anytime</td>
<td>Activities Tasks</td>
<td>Security Managers</td>
</tr>
<tr>
<td>Req7</td>
<td>Any</td>
<td>Administrative Design Technical</td>
<td>Confidential Shared</td>
<td>Anytime</td>
<td>Activities Tasks</td>
<td>Security Managers Patients Practitioners Researchers</td>
</tr>
<tr>
<td>Req8</td>
<td>Any</td>
<td>Design Technical</td>
<td>Confidential Shared</td>
<td>Clinical studies</td>
<td>All</td>
<td>Security Managers Patients Researchers Students</td>
</tr>
<tr>
<td>Req9</td>
<td>Any</td>
<td>Design Technical</td>
<td>Confidential Shared</td>
<td>Emergencies</td>
<td>Activities Tasks</td>
<td>Security Managers Patients Practitioners</td>
</tr>
<tr>
<td>Req10</td>
<td>Any</td>
<td>Legal Management Design Technical</td>
<td>Confidential Shared</td>
<td>Anytime</td>
<td>All</td>
<td>Security Managers Patients Practitioners Government Insurance Companies</td>
</tr>
<tr>
<td>Req11</td>
<td>Any</td>
<td>Management Design Technical</td>
<td>Confidential Shared</td>
<td>Anytime</td>
<td>Activities Tasks</td>
<td>Security Managers</td>
</tr>
</tbody>
</table>
We note that the issues of identification and authentication are still problematic due to the time any global paradigm change requires to take place. Nonetheless, it is possible to make use of the PKI standards together with a trust network to enable both security principles.

Anonymization, although desirable is not completely achievable due to the unpredictability about the future data to be published and the previous knowledge anyone might have to infer sensitive information. The computational cost for achieving k-anonymization is NP-hard for $k \geq 3$ [30, 32], its also NP-hard for l-diversity [55] as well as t-closeness [1], thus rendering any anonymization model out of the question for global scale, where most presumably large datasets will be used. The HIPAA guidelines are a simple and deterministic way to anonymize data that could be combined with k-anonymization with $k = 2$.

The alert layer will be mostly a pattern matching algorithm that looks into the audit trail files. It is difficult to foresee the rules to which these patterns should be compliant. Every healthcare institution will have different equipment, mobile devices and network topologies, implying that each security manager must define the patterns he considers the most appropriate. Nevertheless, by design it must be considered in the architecture so that it may be implemented accordingly.
Proposed System

A perfection of means, and confusion of aims, seems to be our main problem. – Albert Einstein

This chapter explores the functional requirements for practitioners and researchers that are naturally excluded from the legal scope of requirements. From these requirements we demonstrate the adequacy of Linked Data to provide functionality for the two denoted roles, and still, guarantee the security principles already analysed and discussed in the previous chapter.

4.1 Data Heterogeneity

Healthcare systems are rich in type diversity and applicability, not being limited to alphanumerical data and general practice. Type diversity include discrete numerical values (e.g. systolic and diastolic blood pressure), alphabetical values belonging to structured categories (e.g. ordinal measurements) and several media formats like Electrocardiogram waveforms (ECG), Magnetic Resonance Imaging (MRI), Computed Tomography (CT) and X-Ray images. The time dimension is commonly used in healthcare and clinical studies due to the fact that appointments occur several times during patient treatment and in clinical studies all observed measurements are frequently registered during the study.

In terms of applicability, each medical speciality requires distinct parameters to be observed. For a cardiologist an ECG is a tool for diagnosis, while the equivalent for a dentist will be an orthopantomogram. Also, a cardiologist will be looking for symptoms like hypertension or palpitations and will most certainly collect the blood pressure, the heart rate, the weight and height of the patient. A dentist will not require most of these parameters.

As a result, each application developed to aid practitioners in their daily activities ended up creating their own data schemas and formats leading to a non-interoperable and non-integrated information system in the healthcare sector, which was the case with EMRs [3]. Essentially, EMRs ensured the intraoffice operability from patient scheduling to billing, however, every task that required interoperability with external organizations would still need to be processed in a paper-based manner. Several EMR vendors did support type diversity, implementing features like direct editing of pictures, that enabled practitioners to mark an area of interest in an image containing the relevant parts of human anatomy.

The support for data heterogeneity, in both data types and applicability, is technologically built-in into Linked Data. The support of several media formats is implicit due to the Web representation of binary types
for images, videos and audio files. The applicability should be seen as the possibility of integrating several different data models, which is the main purpose of Linked Data. Each data model, or ontology, may represent a specific medical speciality, or some abstract concepts required to provide inference rules.

4.2 Case Study

During the years 2011 and 2012 an information system was developed to support two FP7 financed European projects, namely BIOHYPO\(^2\) and Pneumopath\(^3\). Both these projects required biological data to be stored and specific ontologies had to be developed to ensure that data produced by all the international research groups was integrated. Moreover, data heterogeneity was also a requirement, thus leading to the adoption of Linked Data technologies. Initially, it was unclear if there should be any access control mechanisms involved in data access, because programme committees were considering to enable public access to research results, given the fact that part of the financing came from public funds. Nonetheless, requirements towards the development of an information system using project based authorization were established. Therefore, the same information system could hold data from different projects and participants would only have access to data from their project.

The resulting system was sdlink\(^4\) which served as a case study for implementing authentication delegation and access control. Authentication delegation was achieved by using OpenID, whereby no data related to user credentials was stored and authentication occurred in a remote service. Access control was implemented using Friend of a Friend and Web Access Control (FOAF+WAC) ontologies. The URIs from user accounts together with the available projects formed an ACL.

However, the granularity of access control implemented on sdlink is too coarse to be directly usable in an healthcare system. This is due to the single named graph relationship between projects and respective ontologies. Standards and derived technologies have evolved and new techniques may be considered to implement the required degree of access control in a global healthcare system, as described in the following subsections.

4.3 Data Structure

Given the recent nature of international efforts to implement EHRs, there are still some inconsistencies with the related standards [8]. As a result, our approach is to provide an abstraction over the eventual ontologies to be used in the real system. This structure is depicted in Figure 4.1.

Clear separation of data considering the roles involved in the system will provide simpler implementation of access control mechanisms. Moreover, it will become clearer to whoever defines security policies what are in fact the purposes of each data property in the whole dataset. For example, it is usual to find the patient’s

\(^2\)BYOHYPO information available at: [http://kdbio.inesc-id.pt/biohypo/](http://kdbio.inesc-id.pt/biohypo/)

\(^3\)Pneumopath information available at: [http://www2.le.ac.uk/projects/pneumopath/](http://www2.le.ac.uk/projects/pneumopath/)

name and phone number classified as demographical data, when in fact, demographics are related to age, gender, ethnicity, income level and education level. Also, by creating a division between administrative and demographical data, it will be simpler to define an administrative role that will not have access to the clinical panel or demographics, which are not required to perform the daily tasks of contacting, scheduling and billing patients.

Each time a patient has an encounter (i.e. appointment or emergency), the collected data will be in the standard form of a SOAP note. SOAP stands for Subjective, Objective, Assessment, and Plan, which are the components of the note. The subjective component aims to register the main complaint or purpose of the encounter, and because it will be made of free text and some ordinal qualitative properties it is called subjective in opposition to objective. The objective component will register all the required quantitative measurements (e.g. vital signs). The assessment contains the medical diagnosis for the specific encounter and it may also contain notes in relation to previous encounters, for example, if the patient’s condition is improving. Finally, the plan component refers to the treatment plan, future encounters and lab analysis.

All the information that is collected in the SOAP notes will feed the clinical panel, also known as the patient’s chart. The clinical panel constitutes the core of the EHR, thus being the most relevant information required by most practitioners when dealing with any patient. On a patient’s first encounter, it is essential that the practitioner collects most of these elements. In future encounters the patient will not have to repeat himself to other practitioner, providing the same information over again. Moreover, SOAP notes may be read by people outside the healthcare system, like teachers. In essence the SOAP note not only registers the patient’s history, it also serves as proof that the patient is indeed receiving medical care, justifying absence in school or work.

Additionally, there are privacy concerns regarding the genomic data stored in the clinical panel. The study that raise awareness about identification of individuals from parts of their genome was conducted by Homer et al. in 2008 [26]. The troubling point in this study is that given a number of single nucleotide polymorphisms (SNP) it is still possible to identify an individual. Recent studies have demonstrated the identification of individuals from their genome using simpler methods of combining several publicly available data [24]. These
concerns are valid but they are more applicable to publicly available data. In the proposed system, the data is expected to have controlled access and even knowing that identification is possible, the users that have access to data also have legal responsibilities. These aim to deter any user from doing any other tasks not foreseen in their roles within the healthcare system. The benefits of having genotype-phenotype association for clinical studies is the fundamental aspect for designing a system targeting the controlled access of healthcare data, therefore, genomic data must be included.

On the other hand, the system must address the research functional requirements resulting from clinical studies. These studies always involve humans as subjects and the number of observed parameters may range from tens to hundreds. For clinical studies the most common observed parameters are quantitative, which are preferable to use with statistical tools. There are some studies that may require qualitative measurements, for example, if a study on Alzheimer’s disease requires the patient’s interaction abilities to be registered. Additionally, clinical studies may collect non-alphanumerical data periodically, like ECG waveforms or MRI scans, these types of data, usually called high density data, typically will not serve as input for statistical tools or data mining algorithms, requiring some metadata to be also stored. For example, keeping all the MRI scans from patients with brain tumours during a clinical study will be useful, but in order to make these easily comparable, the area or volume of the tumour should be stored as metadata. Figure 4.2 illustrates the generalized structure of data for clinical studies. High density data and the respective metadata, as well as any qualitative notes are considered part of an observation.

Concerns regarding control groups must be addressed by the leading researcher, if the study is a placebo-controlled one, then only a few elements on the research team will know who are the patients that are treated with real medication and the ones taking placebos. From the design perspective, information about patient-group association should not be stored together with observation data while the study is being conducted, which would raise another security concern. Instead, that information should only be added when the study is over, results are published and the collected data has become available to other research groups.
4.4. Access Control

4.3.1 Evolution and Integration

Because systems are expected to evolve and research is more valuable when it integrates data from several related studies, the technologies selected to implement a global healthcare system must take this requirements into account. Again, from all the available technologies, Linked Data standards stand out. It is possible to make use of already existing ontologies to model specific parts of the system, leading to a more rapid evolution. Also, Linked Data was conceived to enable integration at large scale.

Regarding the purpose of each clinical study, although it is important to note what is the objective of the study, it is relevant to be able to integrate data from several clinical studies that may not be equal in purpose. Patient data collected in clinical studies have been confined to a single institution and within that institution only a few people would have access to the data, typically the researchers involved in the study. This caused information silos between institutions and between different studies. However, it may be important to correlate various pathologies together, for example, it has been shown that breast and ovarian cancers have strong genetic correlations [15]. This integrated analysis between studies with different purposes is fundamental in rare diseases given the lack of statistical significance in each isolated study. Also, more correlations between different health issues may be discovered. This broader applicability of clinical studies must be addressed in the informed consents and patients should be aware that their data may be used for scientific progress and not only in the interest of specific research groups.

4.4 Access Control

Although there are several other functional requirements that could be elicited, the main purpose is to demonstrate that it is possible to address the legal and security concerns of healthcare, while still providing global scale integration and authorized access to heterogeneous data.

A clear separation of data into functional requirements, as shown in Figure 4.1, simplifies not only the purpose of the data, but also the access control required in authorization. This last aspect enables a clearer distinction between the data access each roles has. Table 4.1 identifies roles and their privileges. Users of the system must have at least one role, also, users are distinct from stakeholders in the sense that they handle healthcare data directly.

Address is considered an element of administrative data, however some geographical information regarding the patients is also required as demographical data. The addresses, upon creation or update, should be directly transformed to valid demographical data according to HIPPA rules or any other de-identification rules that respects each nations population distribution. The most generalized form of any patient’s address should be “Country”.

Access control with Linked Data is achieved with graph groups, since data is stored as triples and their composition yields graphs. Each role will have access to a specific collection of graphs and some rules should also be considered in specific user accesses. For example, all practitioners will have read access to the clinical panel and the demographics data, but each practitioner will only be able to create their own SOAP notes.
Table 4.1: List of roles and respective privileges

<table>
<thead>
<tr>
<th>Role</th>
<th>Operation</th>
<th>Resource</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patients</td>
<td>read</td>
<td>all data</td>
</tr>
<tr>
<td></td>
<td>read</td>
<td>demographics</td>
</tr>
<tr>
<td></td>
<td>create</td>
<td>clinical panel</td>
</tr>
<tr>
<td></td>
<td>update</td>
<td>clinical panel</td>
</tr>
<tr>
<td>Practitioners</td>
<td>read</td>
<td>SOAP notes</td>
</tr>
<tr>
<td></td>
<td>create</td>
<td>SOAP notes</td>
</tr>
<tr>
<td>Researchers</td>
<td>read</td>
<td>anonymized demographics</td>
</tr>
<tr>
<td>Administrative Assistants</td>
<td>read</td>
<td>administrative data</td>
</tr>
<tr>
<td></td>
<td>create</td>
<td>administrative data</td>
</tr>
<tr>
<td></td>
<td>update</td>
<td>administrative data</td>
</tr>
<tr>
<td>Management</td>
<td>read</td>
<td>administrative data</td>
</tr>
</tbody>
</table>

The graph structure of data also facilitates the access at multi-institutional and global scales. For example, it is possible to integrate EHRs stored in any healthcare institution by knowing the user role and using an URL structure like https://<organization domain>/ehr/<patient id>/.. Synchronization protocols should be implemented in detriment of simple access to external EHR sources. More specifically, each time a patient goes to a new institution, the first step will be to obtain his personal data to create a medical record. However, having the ability to access EHRs outside the institution, data retrieval should be done to gather all alpha-numerical data and eventually repeated if high density data was required during any encounter. The advantage would be that relevant data to perform medical care would be readily available and access control would be done inside the local system. The synchronization protocol could consider transmitting some information related with billing status, for instance, if a patient is in debt with other institutions.

### 4.4.1 Practical Example

Implementing access control and audit trail layers of the system will enable most of the required security. The anonymization layer may provide a simple generalization of the address and date of birth for research purposes, given the fact that researchers will not access administrative data. For example, addresses can be generalized as ZIP codes and dates of birth as age. For publishing purposes, the data should be further anonymized.

The data in this architecture should be stored in a triple store, while having audit trail and access control intermediating each user access. Another option, is to have access control and audit trail together in the triple store. We will follow the latter approach and provide the practical example using the Open Link Virtuoso. The environment used in this example is available online. The respective usage details are described in Appendix B. The scripts used to define the user profiles and insert the respective RDF data into the triple store are available in Appendix C.

The graph groups feature is a specific extension of Virtuoso to facilitate the creation of SPARQL Protocol and RDF Query Language (SPARQL) queries. The usual form to query several different graphs can be achieved in other triple stores with the usage of multiple “FROM” clauses, as shown in Listing 4.1.
Our concern is not to follow specific ontologies targeting EHR or healthcare data representation, but considering the proposed data structure, the RDF data may be represented by blank nodes to illustrate the access control mechanism over Linked Data. Authorization is ensured by the OAuth 1.0 standard. Table 4.2 exemplifies the data loaded into the triple store.

Table 4.2: Example dataset

<table>
<thead>
<tr>
<th>predicate (property)</th>
<th>object (value)</th>
<th>graph</th>
</tr>
</thead>
<tbody>
<tr>
<td>name</td>
<td>Alice</td>
<td><a href="http://hcorg/patient/adm">http://hcorg/patient/adm</a></td>
</tr>
<tr>
<td>phone</td>
<td>+351 111 222 333</td>
<td><a href="http://hcorg/patient/adm">http://hcorg/patient/adm</a></td>
</tr>
<tr>
<td>address</td>
<td>Rua da Avenida, 1, 1 Dir</td>
<td><a href="http://hcorg/patient/adm">http://hcorg/patient/adm</a></td>
</tr>
<tr>
<td>dateOfBirth</td>
<td>1970-01-01T00:00:00</td>
<td><a href="http://hcorg/patient/dem">http://hcorg/patient/dem</a></td>
</tr>
<tr>
<td>bloodType</td>
<td>O+</td>
<td><a href="http://hcorg/patient/dem">http://hcorg/patient/dem</a></td>
</tr>
<tr>
<td>gender</td>
<td>F</td>
<td><a href="http://hcorg/patient/dem">http://hcorg/patient/dem</a></td>
</tr>
<tr>
<td>datetime</td>
<td>1975-01-01T00:00:00</td>
<td><a href="http://hcorg/patient/cp">http://hcorg/patient/cp</a></td>
</tr>
<tr>
<td>allergies</td>
<td>Penicillin</td>
<td><a href="http://hcorg/patient/cp">http://hcorg/patient/cp</a></td>
</tr>
<tr>
<td>diagnoses</td>
<td>Breast Cancer</td>
<td><a href="http://hcorg/patient/cp">http://hcorg/patient/cp</a></td>
</tr>
</tbody>
</table>

To determine which data is accessible by the patient, administrative assistant and practitioner roles, a select all SPARQL query is used (Listing 4.2).

Listing 4.2: Select all SPARQL query

SELECT * WHERE { ?s ?p ?o }

The obtained results, shown in Figure 4.3, demonstrate that it is possible to achieve access control with Linked Data. The results show that the administrative role has the least amount of results, all of them being from the administrative data graph. The patient as complete access to his data, while the practitioner has access to demographical data and the clinical panel.

Figure 4.3: Results from select all SPARQL query with access control

It is desirable to have one-to-one, or one-to-many, relationships between roles and graphs, nonetheless, some exceptional cases are possible. For example, when a specific role may require access to a single property within a graph. These cases will require additional design effort so that the properties are grouped together in the same graph, thus assuring the non-violation of the desired level of authorization.
Furthermore, each time an healthcare institution has a new administrative assistant, security managers will only need to create a new user account with the administrative role. If a new patient, without any EHR needs to be added, then the activity should not require the intervention of security managers. Essentially the software providing the administrative interface should create a new user account with the UHID, add the new patient’s administrative graph to the administrative graph group. Next, upon the selection of a practitioner, the software should add the respective demographics and clinical panel graphs to the graph group of the respective practitioner. Finally, the software should create a graph group for the patient containing all of his data graphs.
Conclusions and Future Work

I think, therefore I am. – René Descartes

In this thesis we have covered the issue of securing healthcare data. The major concerns being protecting patients from the consequences of unauthorized access and enable researchers to use patient data in their clinical studies. In order to tackle this concerns, legal directives and related security principles had to be identified to understand the data protection requirements.

Although international efforts to enable electronic healthcare records have been made in the last few years, the awareness about the inadequacy of clinical studies for rare diseases is very recent. Thus, the need to enable data integration to assure statistical significance for clinical studies had also become a concern.

We have shown that it is possible to design a fully distributed and global healthcare system, using already existing technology. The identification and authentication of practitioners is possible with the usage of FOAF+SSL or with a chain of trust and digital certificates. The authentication of patients is possible with electronic identification like the European e-ID, although it is still not globally implemented.

Using a six-dimensional framework to analyse the healthcare requirements allowed the identification of the stakeholders and roles, leading to the adoption of an RBAC model to ensure access control. Also, it allowed a clear distinction between all the different scopes of responsibility within the system, thus, enabling the identification of technical requirements. We also provide some recommendations towards requirements that traverse technical, administrative and legal scopes.

Lastly, we have demonstrated the applicability of Linked Data to achieve controlled access to healthcare data, while proposing a segregation approach to enable direct mapping of roles into graph groups. Considering the requirements for data heterogeneity and large scale data integration, we rely on other publicly available projects to substantiate the usage of Linked Data.

As for future work, we propose the design and implementation of data leakage detection mechanisms to prevent situations where a practitioner or a researcher may export patient’s data and deliver it to third parties in open formats. Additionally, the study of efficient anonymization algorithms should be seen as a priority, since anonymizing large datasets is computationally expensive and the consequences from identity theft have alarmingly damaging effects.
Bibliography


Decomposing Management

Every known management method follow the Plan, Do, Check and Act (PDCA) cycle. In the military institutions this is called the Command and Control method, but nonetheless it will map to a cycle comprised of control of execution and corrective decision that may or not lead to replanning. If we consider management only, then the “Do” part of the PDCA cycle will not be applicable, because a manager usually does not execute most of the planned tasks. Notwithstanding, the “Check” and “Act” parts of the PDCA cycle are described in this thesis as the “control of execution” and “corrective decision”. Additionally the inclusion of the word “control” is essential in any management method, since it describes its true nature, and also, its root problem and main cause of failure.

A.1 Management as a Natural Phenomenon

A hungry lioness naturally establishes the objective of feeding herself. In order to do this, she will have to hunt while respecting some tactical aspects, like approaching the prey against the wind so that the prey does not smell her presence, thus following a plan. While hunting the lioness will evaluate her success probability against the amount of meat she will obtain from the prey, thus having control of execution. If the prey is small and demanding too much effort, she will give up the hunting and will target a new prey, thus demonstrating a corrective decision.

A.2 Management and Leadership

Jan Hoogervorst’s book Enterprise Governance and Enterprise Engineering discusses relevant issues that show the root causes of failure in organizations [27]. It identifies these root problems as originating from what he calls the Western Thought. By comparing the western and oriental thoughts it becomes clear that the assumption of control, typical in management, is a fundamental aspect of enterprise strategy failure. In the opposite side of control is leadership that accepts uncertainty. Table A.1 summarises the different aspects between management and leadership identified by Hoogervorst, thus demonstrating why motivational or inspirational factors may not be included as a management elements.
### Table A.1: Differences between management and leadership

<table>
<thead>
<tr>
<th></th>
<th>Management</th>
<th>Leadership</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assumed context</td>
<td>Stable, orderly</td>
<td>Dynamic, chaotic, uncertain</td>
</tr>
<tr>
<td>Primary focus</td>
<td>Control, routinising</td>
<td>Vision, direction, values</td>
</tr>
<tr>
<td>Relation with employees</td>
<td>Transactional</td>
<td>Transformational</td>
</tr>
<tr>
<td>Primary elements in relation</td>
<td>Money</td>
<td>Shared values, goals, trust</td>
</tr>
<tr>
<td>Communication</td>
<td>Top-down</td>
<td>Two-way</td>
</tr>
<tr>
<td>Style</td>
<td>Authoritative</td>
<td>Coach, guiding</td>
</tr>
</tbody>
</table>
Virtual Environment

The prototype is available online at: http://link.inesc-id.pt:8890/oauth/sparql.vsp. It consists in a Virtuoso SPARQL endpoint requiring authentication via OAuth, which is used instead of WebID for simplification purposes, so that no certificate installation is required. The select all SPARQL query is already defined as the default query for asserting access control over linked data.

In order to use the prototype the requirements are: (1) to copy the provided OAuth tokens (below) to the respective textbox on the SPARQL endpoint page; (2) to authorize the user agent accessing the SPARQL endpoint by providing the role password when requested and clicking the authorize button.

- Access http://link.inesc-id.pt:8890/oauth/sparql.vsp
- administrative OAuth token: 091ef9f38fe9fd809648ac4571a6bee455d561753
- patient OAuth token: 6ffc8ed861416bab1c7739e67e407e929a80359f
- practitioner OAuth token: 0d1d596341332461ebf521c9c45a63d3c4b4cbd1
- every role has the same password: 123
Preparing the example dataset can be achieved through the isql application, executing as root the following command: virtuoso/bin/isql. Listing C.1 contains the necessary Virtuoso commands to generate the example dataset.

```sql
-- By default no users should have access to data
DB.DBA.RDF_DEFAULT_USER_PERMS_SET ('nobody', 0);

-- Create the administrative assistant role
DB.DBA.USER_CREATE ('administrative', '123');
DB.DBA.USER_CREATE ('patient', '123');
DB.DBA.USER_CREATE ('practitioner', '123');

-- Assign internal grants
GRANT SPARQL_UPDATE to "administrative";
GRANT SPARQL_UPDATE to "patient";
GRANT SPARQL_UPDATE to "practitioner";

-- The created users should not have access to any graph by default
DB.DBA.RDF_DEFAULT_USER_PERMS_SET ('administrative', 0);
DB.DBA.RDF_DEFAULT_USER_PERMS_SET ('patient', 0);
DB.DBA.RDF_DEFAULT_USER_PERMS_SET ('practitioner', 0);

-- Define user privileges (1 = read, 2 = write, 3 = read and write)
DB.DBA.RDF_GRAPH_USER_PERMS_SET ('http://hcorg/patient/adm', 'administrative', 3);
DB.DBA.RDF_GRAPH_USER_PERMS_SET ('http://hcorg/patient/adm', 'patient', 1);
DB.DBA.RDF_GRAPH_USER_PERMS_SET ('http://hcorg/patient/dem', 'patient', 1);
DB.DBA.RDF_GRAPH_USER_PERMS_SET ('http://hcorg/patient/cp', 'patient', 1);
DB.DBA.RDF_GRAPH_USER_PERMS_SET ('http://hcorg/patient/adm', 'practitioner', 3);
DB.DBA.RDF_GRAPH_USER_PERMS_SET ('http://hcorg/patient/cp', 'practitioner', 3);

-- create administrative patient data
SPARQL
INSERT IN <http://hcorg/patient/adm> {
  _:a <name> "Alice" .
  _:a <phone> "+351 111 222 333" .
  _:a <address> "Rua da Avenida, 1, 1 Dir"
};

-- create demographical patient data
SPARQL
INSERT IN <http://hcorg/patient/dem> {
  _:a <dateOfBirth> <1970-01-01T00:00:00> .
  _:a <bloodType> "O+" .
  _:a <gender> "F"
};

-- create the patient's clinical panel
SPARQL
INSERT IN <http://hcorg/patient/cp> {
  _:a <datetime> <1975-01-01T00:00:00> .
  _:a <allergies> "Penicillin" .
  _:a <diagnoses> "Breast Cancer"
};

-- Create the patient's graph group
DB.DBA.RDF_GRAPH_GROUP_CREATE ('http://hcorg/patient', 1);
```
-- Associate the respective graphs to the patient's graph
DB.DBA.RDF_GRAPH_GROUP_INS ('http://hcorg/patient', 'http://hcorg/patient/adm');
DB.DBA.RDF_GRAPH_GROUP_INS ('http://hcorg/patient', 'http://hcorg/patient/dem');
DB.DBA.RDF_GRAPH_GROUP_INS ('http://hcorg/patient', 'http://hcorg/patient/cp');

-- Create the practitioner's graph group
DB.DBA.RDF_GRAPH_GROUP_CREATE ('http://hcorg/patient/practitioner', 1);

-- Associate the respective graphs to the practitioner's graph
DB.DBA.RDF_GRAPH_GROUP_INS ('http://hcorg/patient', 'http://hcorg/patient/dem');
DB.DBA.RDF_GRAPH_GROUP_INS ('http://hcorg/patient', 'http://hcorg/patient/cp');

Listing C.1: Example creation