Securing and sharing clinical data

Pedro Miguel Baptista dos Reis
pedrombreis@ist.utl.pt

Instituto Superior Técnico, Lisboa, Portugal

October 2013

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 work 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.

Keywords: clinical data, security, privacy, linked data, systems architecture, heterogeneous data

1. Introduction

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 [12, 15]. 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 [3, 4, 11].

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 [8]. Studies show that roughly 80% of the patients treated for cancer die because of drug inadequacy [5]. 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 [13].

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 [10]. By integrating tens of pub-

lic 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 work 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.

2. Background

2.1. Privacy Laws and Principles

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 [7]. 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" [17]. 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 describe the related information security principles, which are: identification, authentication, authorization, confidentiality and accountability.

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.

Authentication 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.

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 $(user, resource, operation)$ 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.

Confidentiality should not be mistaken as privacy. 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. To attain confidentiality measures beyond the technical real must be enforced. For example, a caregiver may disclose information about his patients during informal conversations outside the professional environment, thus requiring legal and administrative procedures to be defined to avoid this kind of disclosure.

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. Access control policies are rules that govern user access to resources in a system. Access control models are the formalization of access control policies. Lastly, access control mechanisms are the implementations of access control models. Figure 1 details the multiphase design of access control.

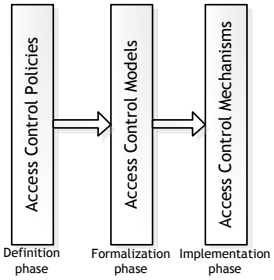

Figure 1: Phases of Access Control Design

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

1. **Least privilege**: users should not be assigned more privileges than those strictly needed to perform the task at hand.
2. **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.
3. **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.

Access control policies are classified in three groups: Discretionary Access Control (DAC), Mandatory Access Control (MAC) and Role-based Access Control (RBAC). Discretionary policies are those at the discretion of the owner or creator of a resource, therefore, the owner defines who else has access to his resources. Mandatory policies are those where a central authority defines who has access to which resources, these policies are typically used in military organizations. Role-based polices are aligned with commercial organization structure, whereby resources are accessed according to the role employees have within the organization.

2.4. Linked Data

The hype surrounding Linked Data began after Tim Berners-Lee gave at the Technology, Entertainment

and Design (TED) conference on February 2009 [6]. This talk also 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. 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.

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:

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

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.

2.5. 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 [20].

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.

Besides the storage formats, a fundamental issue is to understand what is considered private data in healthcare. 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 [6]. 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.

Its 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) [16].

2.6. Ongoing Research

The need for accessing private health data in research has manly to do with phenotype information. Supposing a certain group of individuals has a genetic predisposition for developing lung can-

cer. 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 [9]. 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 [8]. 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, namely, open access, streamlined access, research commons and data analysis servers.

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.

Streamlined access is the simplification of administrative procedures for obtaining access to clinical data.

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

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.

3. Analysis

3.1. Analysis Framework

An analysis framework is fundamental to establish the concepts involved in the design and development of a global healthcare syste. Moreover, it will provide a more concise terminology and semantics between stakeholders. Table 1 details the six dimensions framework.

Table 1: The Six Dimensions Framework

Dimension	Elements
Scale	Local
	Multi-institutional
	Global
Scope	Ethical
	Legal
	Management
	Administrative
	Design
	Technical
Access	Private
	Confidential
	Shared
	Public
Time	Events
Procedural	Processes
	Activities
	Tasks
Operative	Stakeholders

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 the addressed issues [18, 19]. For better contextualization, we detail the identified stakeholders in Table 2

3.2. Requirements

From the analysis of legal documentation a set of eleven requirements are identified:

1. All data related to patients is considered private and owned by the patient.
2. In order to provide treatment (primary use) a partitioner must create, access and change patient data.
3. Any secondary usages of patient data must be consented by the patient.
4. Patients can access and update their personal data.
5. Patients may revoke access privileges to specific practitioners or purposes.
6. Management of access control policies should be possible at local scale.
7. Patients should be able to delegate control over their data in case they expect to become incapacitated.
8. Data should be anonymized to prevent or hamper patient identification, thus allowing scien-

Table 2: System Stakeholders

Stakeholder	Description	Interest
Patients	Individuals seeking medical treatment.	To receive the best treatment with the least amount of risk involved.
Practitioners	Individuals providing medical treatment.	To provide the best treatment according to the patients' characteristics and considering up-to-date knowledge.
Researchers	Individuals provinding knowledge discovery and innovation.	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".
Government	Individuals elected to govern national policies.	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.
Insurance Companies	Organizations providing financial assistance during medical treatment.	To minimise the cost of providing financial assistance. To have no indemnification processes resulting from data breaches.
Security Managers	Individuals responsible for defining and maintaining the security policies.	To minimise the effort and complexity of managing access privileges.
Students	Individuals providing assistance in knowledge discovery and innovation.	To have no hassle in accessing the tools and knowledge required to provide research assistance.

tific studies to be conducted. This is applicable even if the patient enrols in any clinical study.

9. Access restrictions to patient data can be bypassed in case of an emergency.
10. Accountability for perpetrators of any violation.
11. Breach detection.

The resulting architectural design targets only the local scale of the system (Figure 2). Access control is less restrictive for emergencies, depicted by the thinner part of the access control layer.

Figure 2: Architecture for requirements 1 to 11.

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.

3.3. Identification and Authentication

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 [14]. 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.

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.

Patients also require identification so that their data is uniquely identified in the system, enabling researchers conducting clinical studies to identify duplicate entries for the same patient, even if data comes from different data sources. This identifier is known as Universal Healthcare Identifier (UHID).

4. Proposed System

Given the recent nature of international efforts to implement EHRs, there are still some inconsistencies with the related standards [2]. 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 3 and it is applicable to clinical records.

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 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.

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 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.

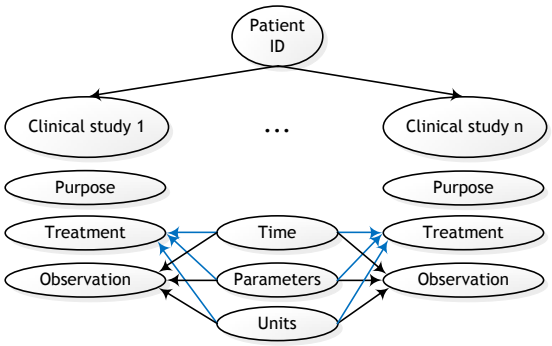

Figure 4: Data structure for clinical studies

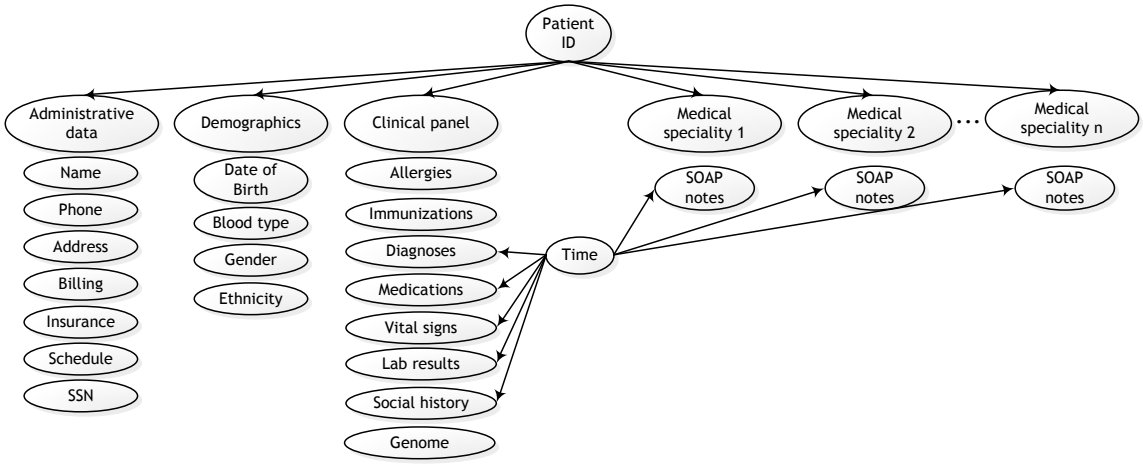

Figure 3: Healthcare data structure

4.1. Authorization and Identification

Authorization can be achieved by using a Role-based access control model without hierarchies, as already proposed in OASIS [1]. However, to enable a fully distributed authentication system FOAF+SSL (WebID) may be used to assure a simple implementation for practitioners, allowing the usage of certificates issued by each healthcare institution without the intervention of national healthcare services. The authorization of patients require a trust network similar to the one resulting from the affiliation between practitioners and healthcare institutions. This may be achieved if a system like the European eID could be implemented at global scale. Nonetheless, the authorization of patients may be bypassed until further political progress is obtained. The crucial factor is to provide unique identification of each patient record in the whole healthcare system. This may be attained with the exclusive usage of natural data instead of circumstantial (Figure 5). This identification method should be performed using a cryptographic hash function like SHA-256 or SHA-512 allowing global identification of patients with low collision probability and also enabling anonymization due to the irreversible properties of hash functions.

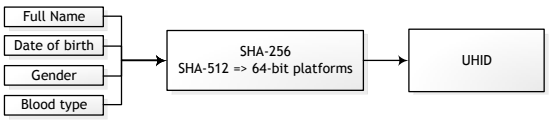

Figure 5: Generating an Universal Healthcare Identifier (UHID).

4.2. Heterogeneity and Integration

Linked Data technologies stand out when considering data heterogeneity in terms of data types (e.g. alphanumerical, electrocardiogram waveforms) and applicability (e.g. cardiology, ophthalmology). Also Linked Data enables simple data integration due to the its atomic format, which allows different ontologies to be linked by adding statements (triples) of the form $(subject, predicate, object)$. Additionally, using techniques like graph groups implemented in some triple stores, it is possible to attain authorization in distinct graphs of the same ontology or different ones.

4.3. Prototype

The prototype uses OAuth instead of WebID for simplification purposes, so that no certificate installation is required. 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 URL: `http://link.inesc-id.pt:8890/oauth/sparql.vsp`
- administrative OAuth token: 091ef5f38fefd809648ac4571a6bee455d561753
- patient OAuth token: 6ffc8ed861416bab1c7739e67e407e929a80359f
- practitioner OAuth token: 0d1d596341332461ebf521c9c45a63d3c4b4cbd1
- every role has the same password: 123

5. Conclusions

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 (WebID) 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 clinical 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.

References

[1] J. Bacon, K. Moody, and W. Yao. A model of oasis role-based access control and its support for active security. *ACM Transactions on Information and System Security (TISSEC)*, 5(4):492–540, 2002.

[2] B. Blobel and P. Pharow. Analysis and evaluation of ehr approaches. *Methods of information in medicine*, 48(2):162, 2009.

[3] K. A. Davenport. Identity Theft That Can Kill You. `http://www.law.uh.edu/healthlaw/perspectives/2006/(KD)IdentityTheft.pdf`, 2006. [Online; accessed 2013-08-17].

[4] Dell SecureWorks. Healthcare security breaches. `http://www.secureworks.com/assets/pdf-store/other/infographic.healthcare.pdf`, Oct. 2012. [Online; accessed 2013-08-17].

[5] J. A. DiMasi, J. M. Reichert, L. Feldman, and A. Malins. Clinical approval success rates for investigational cancer drugs. *Clinical Pharmacology & Therapeutics*, 2013.

[6] European Council Medical Orders. Deontological Guidelines. `http://www.ceom-ecmo.eu/en/deontological-guidelines-144`, 2013. [Online; accessed 2013-09-12].

[7] European Parliament and the Council of the European Union. Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. `http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:EN:HTML`, 1995.

[8] Global Alliance. Creating a Global Alliance to Enable Responsible Sharing of Genomic and Clinical Data. `https://www.broadinstitute.org/files/news/pdfs/GAWhitePaperJune3.pdf`, June 2013. [Online; accessed 2013-08-17].

[9] National Human Genome Research Institute. Workshop on Establishing a Central Resource of Data from Genome Sequencing Projects. `http://www.genome.gov/Pages/Research/DER/GVP/Data_Aggregation_Workshop_Summary.pdf`, June 2012. [Online; accessed 2013-09-12].

[10] Ontotext AD. Linked life data sources and repository overview. `http://linkedlifedata.com/sources.html`, July 2013. [Online; accessed 2013-07-20].

[11] Ponemom Institute. Third annualsurvey on medical identity theft. `http://www.ponemon.org/local/upload/file/Third_Annual_Survey_on_Medical_Identity_Theft_FINAL.pdf`, June 2012. [Online; accessed 2013-08-17].

[12] Ponemom Institute. 2013 cost of data breach study: Global analysis. `https://www4.symantec.com/mktginfo/whitepaper/053013_GL_NA_WP_Ponemon-2013-Cost-of-a-Data-Breach-Report_daiNA_cta72382.pdf`, May 2013. [Online; accessed 2013-08-17].

[13] Presidential Commission for the Study of Bioethical Issues. Privacy and Progress in Whole Genome Sequencing. `http://bioethics.gov/sites/default/files/PrivacyProgress508_1.pdf`, Oct. 2012.

[14] H. Story, B. Harbulot, I. Jacobi, and M. Jones. Foaf+ssl:restful authentication for the social web. In *Proceedings of the First Workshop on Trust and Privacy on the Social and Semantic Web (SPOT2009)*, 2009.

[15] Symantec Corporation. Internet Security Threat Report 2013. `http://www.symantec.com/content/en/us/enterprise/other_resources/b-istr_main_report_v18_2012_21291018.en-us.pdf`, Apr. 2013. [Online; accessed 2013-08-17].

[16] U.S. Department of Health & Human Services. Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf, Oct. 2012. [Online; accessed 2013-09-12].

[17] U.S. Government, Legal Information Institute. Title 44, Chapter 35, Subchapter 111, 3542.

[18] J. Zachman. The zachman framework for enterprise architecture. *Zachman International*, 2002.

[19] J. A. Zachman. John Zachman's Concise Definition of The Zachman Framework. http://zachman.com/about-the-zachman-framework, 2008. [Online; accessed 2013-09-12].

[20] R. Zhang and L. Liu. Security models and requirements for healthcare application clouds. In *Cloud Computing (CLOUD), 2010 IEEE 3rd International Conference on*, pages 268–275. IEEE, 2010.