This is a pre-print version of the following article: Kamateri, E., Kalampokis, E., Tambouris, E., and Tarabanis, K. The Linked Data Access Control Framework. *Journal of Biomedical Informatics*, Vol.50, pp. 213-225 http://dx.doi.org/10.1016/j.jbi.2014.03.002

Available for download from http://kalampok.is

The Linked Medical Data Access Control Framework

Eleni Kamateri[1,2], Evangelos Kalampokis[1,2], Efthimios Tambouris[1,2] and Konstantinos Tarabanis[1,2]

[1]Information Technologies Institute, Centre for Research & Technology - Hellas, 6th km Xarilaou - Thermi, 57001, Thessaloniki, Greece

[2]University of Macedonia, Egnatia 156, 54006, Thessaloniki, Greece

{ekamater, ekal, tambouris, kat}@uom.gr

Abstract

The integration of medical data coming from multiple sources is important in clinical research. Amongst others, it enables the discovery of appropriate subjects in patient-oriented research and the identification of innovative results in epidemiological studies. At the same time, the integration of medical data faces significant ethical and legal challenges that impose access constraints. Some of these issues can be addressed by making available aggregated instead of raw record-level data. In many cases however, there is still a need for controlling access even to the resulting aggregated data, e.g., due to data provider's policies. In this paper we present the Linked Medical Data Access Control (LiMDAC) framework that capitalizes on Linked Data technologies to enable controlling access to medical data across distributed sources with diverse access constraints. The LiMDAC framework consists of three Linked Data models, namely the LiMDAC metadata model, the LiMDAC user profile model, and the LiMDAC access policy model. It also includes an architecture that exploits these models. Based on the framework, a proof-of-concept platform is developed and its performance and functionality are evaluated by employing two usage scenarios.

Keywords

Medical data, data-cubes, linked data, RDF, access policy, privacy.

1. Introduction

Clinical research aims at finding new and better ways to understand, diagnose, prevent, or treat a specific pathological process, e.g., diseases or adverse events. It comprises three main categories: (i) the patient-oriented research that involves human subjects; (ii) the epidemiological and behavioral studies that examine the distribution of disease and the factors that affect health; and (iii) the outcomes and health services research that seeks to identify the most effective and efficient interventions, treatments, and services [1].

Clinical research often requires the integration of medical data coming from multiple datasets that are usually stored across multiple sources such as hospitals, clinical sites, research institutes and pharmaceutical companies [2-5]. Medical data may contain sensitive patient data such as demographics, diagnoses, and medication, as well as radiology images, laboratory test results, doctors' entries and comments [6-7].

In patient-oriented research, the integration of multiple medical datasets enables the identification of a sufficient number of subjects [8]. For example, clinical trial phase III, which assesses the safety and the efficacy of a studied treatment or drug, requires large groups of people matching specific eligibility criteria that cannot be found through a single clinical site. In epidemiological studies, analysis of integrated datasets improves the statistical power of results. For instance, studies of clinical effectiveness or disease biology in rare diseases are only possible through multi-center analyses [9]. The integration of multiple datasets also enables better understanding of relationships between pathological processes and risk factors, or between genotype and phenotype [10-11]. For example, recent genome-wide association studies identified 13 novel loci associated with systolic and diastolic blood pressure as well as hypertension [12-13].

At the same time however, clinical researchers face technical and interoperability [14], as well as ethical and legal [15], challenges in discovering and accessing scattered and heterogeneous medical data. Although the former challenges have been addressed by several standards [16-17], the latter still remain.

In order to overcome these ethical and legal challenges, the approach of aggregating data has been proposed and widely employed. According to this approach, only the counts of subjects having specific characteristics are reported instead of raw record-level data, guaranteeing in this way non-identification and anonymization. Despite that, there is still need for controlling access even to aggregated data, e.g., due to data providers' policies. A promising technology that facilitates data discovery and access at a Web scale is Linked Data. Linked Data refers to *"data published on the Web in such a way that it is machine-readable, its meaning is explicitly defined, it is linked to other external datasets, and can in turn be linked to from external datasets"* [18]. Currently, the most promising implementation of Linked Data involves publishing structured data in RDF using URIs in contrast to the full-fledged Semantic Web vision focusing on the ontological level or inferencing [19].

The objective of this paper is to present the Linked Medical Data Access Control (LiMDAC) framework that capitalizes on Linked Data technologies to enable controlling access to medical data across distributed sources with diverse access constraints. The framework consists of (a) three Linked Data models, namely the LiMDAC metadata model for describing aggregated medical data, the LiMDAC user profile model for describing medical data consumers, and the LiMDAC access policy model, and (b) an architecture that

exploits and orchestrates the three models to enable controlling access to medical data. From a technological perspective, the framework is validated using a proof-of-concept platform that is developed for that purpose.

The remaining of the paper is organized as follows. In Section 2, we provide background knowledge that is necessary for scoping and presenting our work. Section 3 presents the state-of-the-art regarding (a) existing solutions for controlling access to medical data and (b) the use of Linked Data technologies in medical data and for access control. Section 4 describes the proposed LiMDAC framework while Section 5 illustrates a proof-of-concept implementation of the LiMDAC framework. Finally, in Section 6 the results are discussed and in Section 7 conclusions are drawn.

2. Access Constraints

Ethical and legal challenges related to medical data mainly derive from (a) strict regulations that protect personal data and prevent patient re-identification by any means [20], (b) agreements that are specified in consent forms, e.g., patients approve sharing their data only in certain clinical studies [21], and (c) policies of stakeholders owing the data e.g., pharmaceutical companies do not contribute to a clinical research led by competitors, or physicians exclude data derived from studies in progress [22-24].

In general, ethical and legal challenges impose access constraints that can be categorized as follows [25-26]:

- *Secrecy*: Ensures the privacy of patients and the confidentiality of medical data preventing unauthorized disclosures of information.
- *Integrity*: Ensures the integrity of medical data and prevents the unauthorized or improper modifications of data.
- *Availability*: Ensures the availability of medical data only to authorized persons and prevents the unauthorized or unintended withholding of data.

In order to overcome *secrecy* and *integrity* constraints, the approach of aggregating data has been proposed. Aggregated data includes only counts of patients having specific characteristics instead of raw record-level information. Aggregated data is usually structured in the form of multi-dimensional data cubes [27, 28]. In this way, non-identification and anonymization are ensured while the original data remain safe from any modifications. Despite that, there is still a need for controlling access to aggregated data, e.g., due to data provider's policies, and thus *availability constraints* call for appropriate solutions [29].

In order to elaborate on availability constraints, a patient-oriented research and an epidemiological study scenario are described below.

Patient-oriented research

In patient-oriented research, clinical researchers search for subjects that meet certain eligibility criteria related to a clinical study. Initially they identify possible data providers and ask them whether data of relevant subjects is included in their patients' database. Data providers having such data and wishing to participate to the specific clinical study have to perform some intensive tasks. First, they check whether the identified subjects can be included according to the study's eligibility criteria. Then, they match the eligible subjects with the patients' consent forms to identify if they can be enrolled to the specific trial. Finally, they confirm that access to the patient data is permitted without violating any access constraints, e.g., when subjects have been recruited for a different trial. If the number of eligible patients is not sufficient, clinical researchers seek for additional subjects from other sources to meet the recruitment target.

Epidemiological study

In epidemiological studies, clinical researchers perform statistical analyses of medical data in order to conduct secondary clinical research and thus, identify risk factors influencing the occurrence of a pathological process. In order to have accurate and statistically significant results, they need a large number of medical data. To this end, they identify possible data providers and ask for relevant data they can access without violating any access constraints. Data providers wishing to contribute to the specific clinical research have to perform some intensive tasks. First, they modify their data in order to ensure that their data will be transferred in an anonymized and non-identifiable form. To achieve this, they delete all references to the subject and create aggregated data. Data providers confirm that access to the data is permitted without violating any policies and provide the data e.g., data that is used for studies in progress.

In addition to the scenarios, we interviewed stakeholders working in organizations that participate in the EU funded FP7 Linked2Safety project [30]. In particular, we interviewed five clinical researchers, one data manager and three clinical study managers coming from three healthcare organizations maintaining and using medical data for clinical research, namely the Institute of Neurology and Genetics in Cyprus, the Lausanne University Hospital, and ZEINCRO Hellas S.A., a private Contract Research Organisation in Greece. This exercise resulted in a list of user requirements that are related to availability constraints. This list is presented in the Appendix.

These scenarios and requirements enable us to identify that two *abstract roles* related to medical data management are important in clinical research. The *data provider* creates and keeps medical data regarding patient-specific information in order to organize patients' treatment, or conduct a clinical research. The *data consumer* discovers subjects meeting certain eligibility criteria for a patient-oriented research, or medical data to perform an epidemiological study.

In addition, these scenarios and requirements enable us to come up with an *abstract process* that delineates clinical research and consists of the following steps:

1. Data provider modifies and aggregates data in order to ensure anonymity and non-identification.
2. Data consumer searches for data providers
3. Data consumer asks data provider for certain data.
4. Data provider checks whether the requested data is available.
5. Data provider checks whether data consumer is allowed to access the data according to some access constraint policies.
6. Data consumer receives the data.

3. State of the Art

3.1 Access Control of Medical Data

Two different approaches to discover and access medical data from multiple data sources have been proposed: centralized and distributed [3]. The centralized approach enables access to medical data that have been transferred in centralized repositories. Alternatively, distributed medical data networks enable discovery of medical data in their original, disparate locations.

Several frameworks consisting of processes, data models and software tools have been recently developed in order to enable medical data sharing and reuse [31]. For example, the Informatics for Integrating Biology and the Bedside (i2b2) [32-33] open source platform and software implementation allow clinical researchers to search across multiple i2b2 sites, find sets of interesting patients, and reuse medical data while preserving patient privacy and ensuring data integrity. While i2b2 created an analytic platform for a single clinical data repository, the Shared Pathology Information Network (SPIN) [34] has tackled the problem of cross-institution data sharing across a peer-to-peer network, in which each participating institution maintains autonomy and control of its own data. The Shared Health Research Information Network's (SHRINE) [35-36] was built upon i2b2 and SPIN to enable investigators to search the electronic health records of patients across multiple independent sites. In the same context, the Cross-Institutional Clinical Translational Research (CICTR) [8] framework also extended i2b2 in order to enable federated queries across i2b2 sites. In particular, implementations of these two frameworks allow querying distributed hospitals and display aggregate counts of the number of matching patients. The Biomedical Informatics Research Network (BIRN) [37] aggregates imaging, clinical and behavioural data from multiple independent sites using a mediator that re-submits user queries to the relevant sites and aggregates results. The Service-Oriented Interoperability Framework (SIF) [38] targets heterogeneous data sources and employs Web Services standards (e.g., SOAP) to query JDBC databases. In this case users should be aware of all data models in order to form appropriate queries. The Federated Utah Research & Translational Health e-Repository (FURTHeR) [39] is also based on Service Oriented Architecture and

employs Web Services standards in order to perform federated queries across distributed data sources. Finally, the integrating Data for Analysis, anonymization, and SHaring (iDASH) [40] framework covers many aspects of medical data reuse including annotation, compression, anonymization, information extraction from text, sharing in a privacy-preserving manner, and integration.

3.2 Linked Medical Data

Linked Data is an approach for accessing and connecting data using open Web standards. The Linked Data publishing process [41-42] usually begins with existing, structured data in various formats (CSV, relational data, XML, etc.), which are converted to RDF. The publication is based on a *Linked Data model* that can be created either by devising a new local schema (in RDF Schema, OWL, etc.), or by reusing existing, widespread vocabularies (such as FOAF [43], Dublin Core [44], and SKOS [45]). Thereafter, URIs are assigned to the items in the data set on the instance level and interlinks are established with other datasets. There are typically two types of links that can be established, namely *owl:sameAs* used to link URI aliases and other, domain-specific RDF links such as *foaf:knows* and *dc:author*.

The importance of publishing *medical data* as Linked Data becomes apparent in the case that we reuse widely used ontologies or linked datasets. For example, the International Classification of Diseases-11 ontology [46] classifies diseases and other health problems, including signs, symptoms, abnormal findings, etc. In addition, the Experimental Factor Ontology (EFO) [47] combines parts of several biological ontologies, such as anatomy, disease and chemical compounds, to annotate experimental variables. Furthermore, an increasing amount of life science datasets are becoming available as linked data. UniProt is a database of protein sequence [48], Reactome describes biological pathways [49], ChEMBL contains bioactive drug-like small molecules [50], ChemSpider includes chemical compounds [51], and WikiPathways describes pathways [52]. The reuse of such ontologies and linked datasets enables the disambiguation of concepts referring to the same entity, as well as the enrichment of medical data with data coming from disparate sources.

The *RDF Data Cube vocabulary (qb)* is a Linked Data model that defines how to structure multi-dimensional data using RDF [53]. It is of vital importance in Linked Medical Data because it can be used to publish the aggregated data that clinical researchers produce. According to the vocabulary, a data cube *(qb:Dataset)* consists of observations (*qb:Observation*) that are characterised by *dimensions (qb:DimensionProperty), measures (qb:MeasureProperty*) and possibly by additional *attributes (qb:AttributeProperty).* Specifically, the dimension defines the characteristics of the observed properties (e.g., gender, BMI, received medication and weight), the measure describes what has been measured (e.g., cases) and the attribute represents how the observations are expressed (e.g., units, status, etc.) The possible values for each dimension are taken from a *code list*. A code list is a controlled vocabulary such as a list of diseases, or possible age groups.

At the same time, several research efforts have been made so far to control access in data published as Linked Data [54]. Initially, access policies were defined for the entire RDF file stored on a web server [55, 56]. Thereafter, it was attempted to apply access policies on parts of the RDF graph [57-62]. In order to achieve this, the proposed access control frameworks define parts of the RDF graph on which access can be allowed (or denied). These parts are identified by specifying RDF patterns. Whereas the above approaches have primarily focused on RDF patterns, Costabello et al. [63] and Sacco et al. [64] propose access control ontologies over Linked Data. They both employ the SPARQL ASK to determine whether the requester is allowed, or not to access the requested resource. In general, the ASK query form can be used to test whether a query pattern has a solution and returns whether the solution exists. An important issue that may arise is the increase of the overhead produced by evaluating policies in every RDF triple [65].

4. The Linked Medical Data Access Control Framework

The proposed Linked Medical Data Access Control (LiMDAC) framework consists of the following:

a. Three Linked Data models, namely the LiMDAC metadata model, the LiMDAC user profile model and the LiMDAC access policy model.

b. An architecture that exploits and orchestrates the three models to enable controlling access to medical data.

The framework aims at supporting the abstract process of data management in clinical research presented in Section 2. In particular, the steps of the process supported by the LiMDAC framework, along with a mapping to those steps presented in Section 2, are depicted in Table 1. In this table we assume that a platform has been implemented based on the LiMDAC architecture.

Table 1: Abstract clinical research process supported in the LiMDAC framework

Existing Abstract Process	Abstract Process in the LiMDAC framework
Setup phase	
1. Data provider modifies and aggregates data in order to ensure anonymity and non-identification.	1. (This is not supported by the current version of the LiMDAC framework)
	• Data providers publish aggregated data as Linked Data and they use the LiMDAC metadata model to describe them. • Data providers define access constraints using the LiMDAC access policy model.

This is a pre-print version of the following article: Kamateri, E., Kalampokis, E., Tambouris, E., and Tarabanis, K. The Linked Data Access Control Framework. *Journal of Biomedical Informatics*, Vol.50, pp. 213-225 http://dx.doi.org/10.1016/j.jbi.2014.03.002
Available for download from http://kalampok.is

	• Data consumers create profiles based on the LiMDAC user profile model.
Access phase	
2. Data consumer searches for data providers.	2. Data consumer searches for data providers' SPARQL endpoints through the LiMDAC platform.
3. Data consumer asks data provider for certain data.	3. Data consumer search for suitable data based on the LiMDAC metadata model.
4. Data provider checks whether the requested data is available.	4. The LiMDAC platform checks whether the providers' datasets include suitable data.
5. Data provider checks whether data consumer is allowed to access the data according to some access constraint policies.	5. The LiMDAC platform checks data consumer's profile against the available access policies.
6. Data consumer receives the data.	6. Data consumer receives suitable data through the LiMDAC platform.

The rest of this section is structured according to the main parts of the process and framework. In particular, Section 4.1 presents how aggregated medical data should be developed and published as Linked Data. This corresponds to the setup phase of Table 1and is an essential prerequisite for the LiMDAC framework. We note that this is presented here only for clarity. The processes, technologies and tools for performing these tasks are outside the scope of this paper, as they are well documented in the relevant literature, e.g., [27-28]. Section 4.2 describes the LiMDAC metadata model, while Section 4.3presents the LiMDAC user profile model. Section 4.4 elaborates on the LiMDAC access policy model and Section 4.5describes an architecture that exploits these three models.

In order to enhance clarity, we present a research study about childhood obesity [66]. According to the study, six paediatric academic health sites from different regions of the United States have participated in a clinical research related to children's obesity. The dataset maintained by each site involves records about children between 2 and 17 years old. For each child, the health sites store information about the age and the Body Mass Index (BMI). Moreover, the health sites have detected groups of conditions that most commonly co-occur with obesity including hypertension, hyperlipidaemia as well as other rare disorders such as acute leukaemia, multiple sclerosis, and chromosomal anomalies. In this paper, we use the background scenario of this study in order to present our results. We should, however, underline that we have used dummy and not real data from the study.

This is a pre-print version of the following article: Kamateri, E., Kalampokis, E., Tambouris, E., and Tarabanis, K. The Linked Data Access Control Framework. *Journal of Biomedical Informatics*, Vol.50, pp. 213-225 http://dx.doi.org/10.1016/j.jbi.2014.03.002

Available for download from http://kalampok.is

4.1 Linked Medical Data Cubes

Based on the obesity example, Table 2 depicts part of a truncated dummy data cube provided by one of the sites.

Table 2: Children's Obesity Data Cube coming from a data provider

Disease	**Hypothyroidism**				**Diabetes**				**…**			
BMI / **Age**	**15**	**16**	**17**	**…**	**15**	**16**	**17**	**…**	**15**	**16**	**17**	**…**
2	22	22	23		12	23	11					
3	23	22	23		20	23	12					
4	22	22	24		30	24	22					
…												

Figure 1 presents the RDF graph produced from the data shown in Table 2. The graph is modelled based on the RDF data cube vocabulary. In particular, it describes the data structure definition, along with two observations. The SKOS concept collection is used to indicate a set of disease concepts. Figure 1 also presents the links that have been established between concepts and external vocabularies. It is apparent that "ex:disease" is linked to the concept "EFO:disease" from the EFO ontology that has the same meaning. The concept "EFO:body mass index" can be reused for the BMI dimension. Moreover, the measurement of frequencies and the age dimension can be expressed using the Statistical Data and Metadata eXchange standard (SDMX)[1] which is used to publish statistical data on the web.

We repeat here that medical data is transformed in this format during the setup phase shown in Table 1, is a prerequisite for employing the LiMDAC framework, and is not further elaborated in this paper.

[1] http://sdmx.org/

This is a pre-print version of the following article: Kamateri, E., Kalampokis, E., Tambouris, E., and Tarabanis, K. The Linked Data Access Control Framework. *Journal of Biomedical Informatics*, Vol.50, pp. 213-225 http://dx.doi.org/10.1016/j.jbi.2014.03.002

Available for download from http://kalampok.is

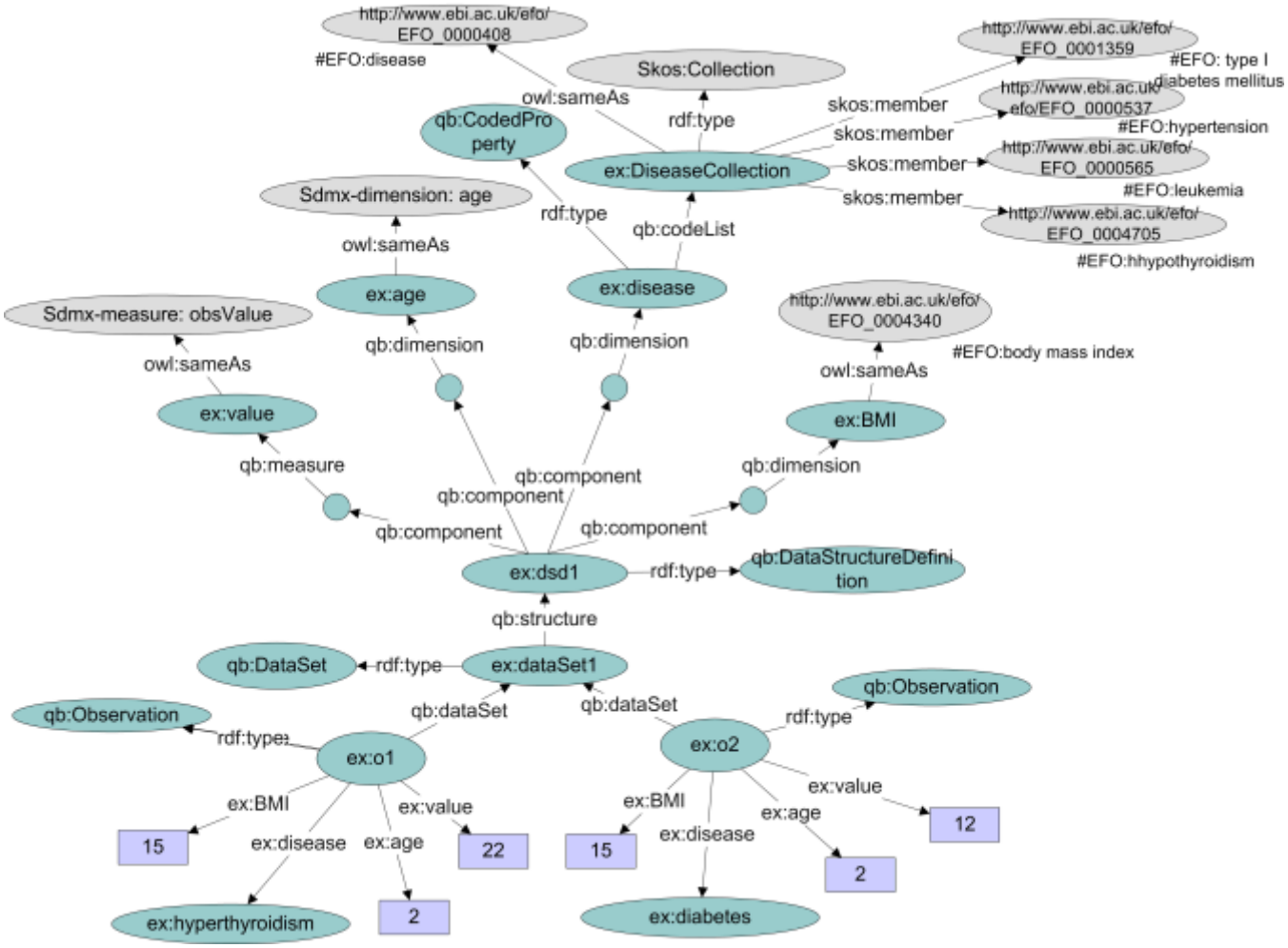

Figure 1: RDF graph of Children's Obesity Data Cube

4.2 The LiMDAC Metadata Model

The LiDMAC metadata model facilitates the improved description of linked data cubes. Metadata enable data providers to express richer access constraints and data consumers to perform more expressive search queries.

The RDF data cube vocabulary includes the *qb:DataStructureDefinition* concept which defines metadata related to cube structure. These metadata include the dimensions of the data cube, along with the values that are measured. However, additional metadata are needed to provide information about the clinical study (e.g., title, purpose, duration, location, subject, responsible personnel for conducting the clinical study etc.) and the aggregation process that has been followed. These metadata were extracted from the requirements described in Section 2 and presented in the APPENDIX.

Table 3 presents the mapping between the requirements and the concepts extracted for describing medical data cubes. Based on these concepts, a conceptual model has been created, which is depicted in Figure 2.

This is a pre-print version of the following article: Kamateri, E., Kalampokis, E., Tambouris, E., and Tarabanis, K. The Linked Data Access Control Framework. *Journal of Biomedical Informatics*, Vol.50, pp. 213-225 http://dx.doi.org/10.1016/j.jbi.2014.03.002

Available for download from http://kalampok.is

Table 3: Concepts of the LiMDAC metadata model as elicited from user requirements

Concepts / Requirements	Data	Study	Agent	Role	Variable	PeriodOfTime	Location
R1	√				√		
R2	√		√	√			
R3	√	√					
R3a		√					
R3b		√					
R3c		√					
R3d		√	√	√			
R3e		√	√	√			
R3f		√	√	√			
R3g		√	√	√			
R3h		√				√	
R3i		√	√				√

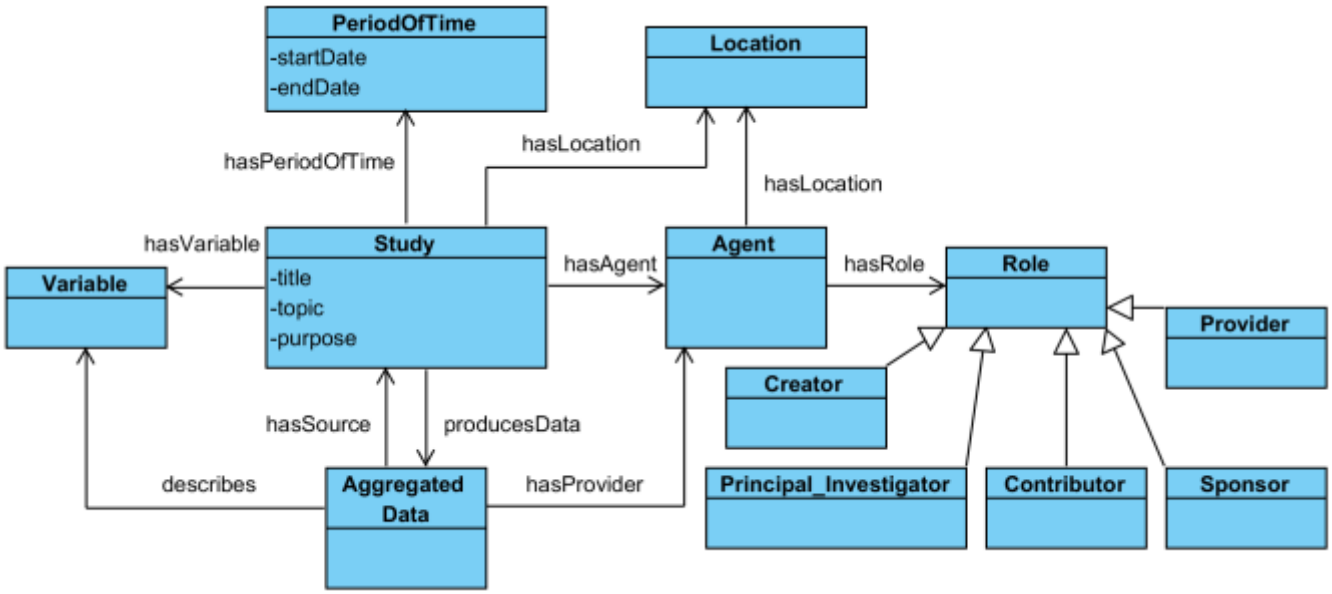

Figure 2: The LiMDAC metadata conceptual model

Figure 3 shows the the LiMDAC metadata model in the form of a linked data vocabulary. Since the model follows the linked-data principles, it reuses concepts from existing vocabularies instead of defining new ones. In particular, the following popular linked data vocabularies are exploited:

- The DDI Discovery Vocabulary [67] that describes research and survey datasets on the Web.

This is a pre-print version of the following article: Kamateri, E., Kalampokis, E., Tambouris, E., and Tarabanis, K. The Linked Data Access Control Framework. *Journal of Biomedical Informatics*, Vol.50, pp. 213-225 http://dx.doi.org/10.1016/j.jbi.2014.03.002
Available for download from http://kalampok.is

- The DCMI Metadata Terms vocabulary [44] that is a specification of all metadata terms used to describe a resource.
- The FOAF vocabulary [43] that describes people and their relationships.
- The SKOS vocabulary [45] that is used to define classifications.

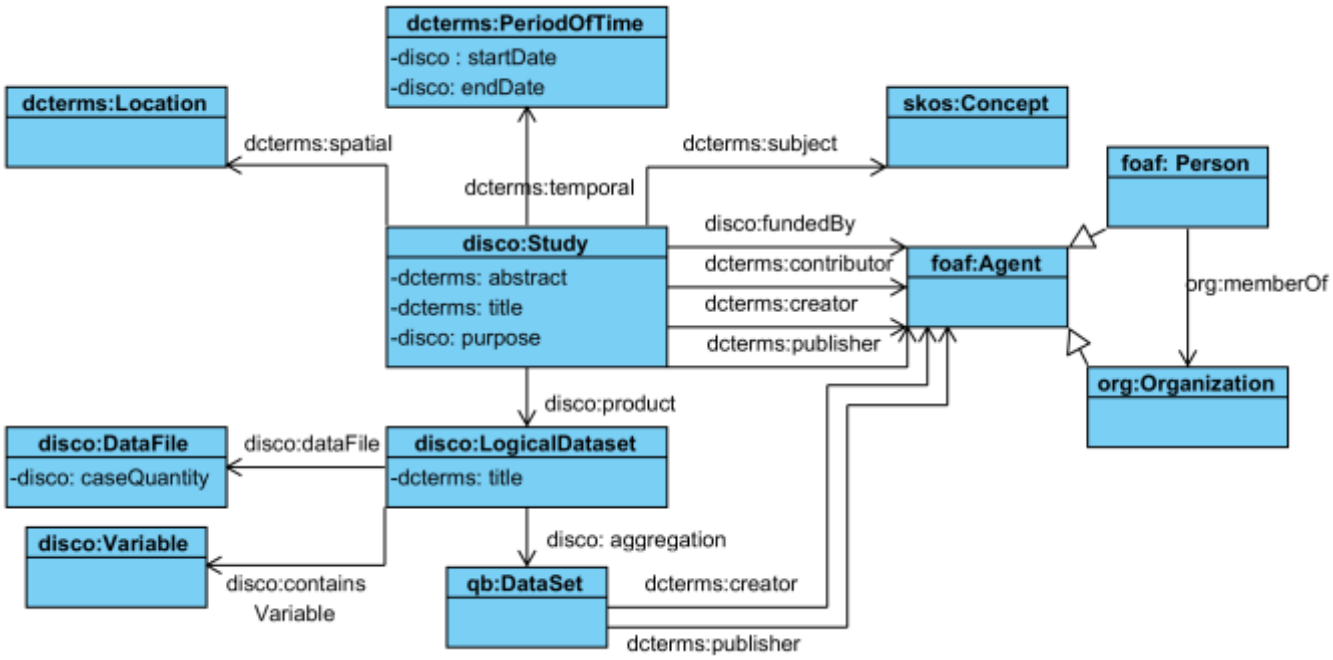

Figure 3: The LiMDAC metadata linked data model

In order to describe the proposed metadata model, we use the Study class (*disco:Study*) as an entry point. A Study represents the process by which a dataset was collected. It may contain a set of literal properties that provide information about the title (*dcterms:title*), the purpose (*disco:purpose*), and other high-level information. In addition, a study includes object properties such as the sponsor of the study (*disco:fundedBy*), and other affiliations such as creators, contributors and publishers of Studies (*dcterms:creator, dcterms:publisher, dcterms:contributor)*. They all are foaf:Agents, which can be either foaf:Persons or org:Organizations whose members are foaf:Persons. Moreover, we can use the subject, the temporal and the spatial properties (*disco: subject, dcterms:temporal, dcterms:spatial*) to describe the respective coverage of studies. For time periods, start (*disco:startDate*) and end dates (*disco:endDate*) can be also attached.

The outcome of a study is a raw record-level data set (*disco:LogicalDataSet*) that may point to multiple variables (*disco:Variable)*. The LogicalDataset contains a property *disco:aggregation* that indicates that a data cube *(qb:Dataset)* was derived by tabulating a LogicalDataset. Furthermore, the Study/LogicalDataset has a DataFile (*disco:DataFile*) that is the distributed file holding that data.

This is a pre-print version of the following article: Kamateri, E., Kalampokis, E., Tambouris, E., and Tarabanis, K. The Linked Data Access Control Framework. *Journal of Biomedical Informatics*, Vol.50, pp. 213-225 http://dx.doi.org/10.1016/j.jbi.2014.03.002

Available for download from http://kalampok.is

In the example of childhood obesity, the Children's Hospital of Colorado is the creator of the study that is titled "Multi-institutional study to access childhood obesity" and its purpose is to associate several health conditions with childhood obesity. Michael G. Kahn is the principal investigator of the study as well as responsible person for creating the data cubes transforming the produced medical data in an aggregated form and publishing them as linked data. The study was funded by the Agency for Healthcare Research and Quality (AHRQ) and run from January 2007 to December 2008.

Taking into account all this information, the Children's Hospital of Colorado enriches the linked data cube derived from this study with the following LiMDAC metadata as shown in Figure 4.

```
:childhoodObesityStudy a disco:Study;                    #metadata about the study
    dcterms:title          "Multi-institutional study to access childhood obesity";
    disco:purpose          "Identify the risk factor of childhood obesity";
    dcterms:temporal       [
      disco:startDate      "2007-01-01";
      disco:endDate        "2008-12-31";]
    dcterms:spatial        :Colorado;
    dcterms:creator        :ChildrenHospitalColorado;
    disco:fundedBy         :AgencyforHealthcareResearchandQuality;
    disco:product          :childhoodObesityLogicalDataSet.
:childhoodObesityLogicalDataSet  a disco:LogicalDataSet; #metadata about the logical dataset
    disco:dataFile         :obesityLogicalDataFile;
    disco:aggregation      :childhoodObesityDataCube1, :childhoodObesityDataCube2, ..
:obesityLogicalDataFile a disco: DataFile;               #metadata about the data file
    disco:caseQuantify     "1000";
:childhoodObesityDataCube1 a qb:DataSet;                 #metadata about the data cubes
    dcterms:creator        :MichaelGKahn;
    dcterms:publisher      :MichaelGKahn;
    qb:structure           :childhoodObesityDataStructureDefinition1 #structure's dimensions:disease,
                                                                      BMI and age
:childhoodObesityDataCube2 a qb:DataSet;
    dcterms:creator        :MichaelGKahn;
    dcterms:publisher      :MichaelGKahn;
    qb:structure           :childhoodObesityDataStructureDefinition2 #structure's dimensions:BMI, age
                                                                      and physical activity
[...]
```

Figure 4: Metadata of Children's Obesity Data Cube

4.3 The LiMDAC User Profile Model

The LiMDAC user profile model is used to describe data consumers. This model is exploited by data providers to define their access constraints and by data consumers to describe their user profiles. The LiMDAC user profile model should be in alignment with the requirements expressed by clinical research stakeholders (APPENDIX). Table 4 presents the mapping between users requirements and the identified concepts that are used in the model (Figure 5).

Table 4: Concepts of the LiMDAC user profile model as elicited from user requirements

Concepts / Requirements	Person	Organization	Position	Location	Activity	Role

This is a pre-print version of the following article: Kamateri, E., Kalampokis, E., Tambouris, E., and Tarabanis, K. The Linked Data Access Control Framework. *Journal of Biomedical Informatics*, Vol.50, pp. 213-225 http://dx.doi.org/10.1016/j.jbi.2014.03.002

Available for download from http://kalampok.is

R13	√					
R14	√			√		
R15	√					
R16	√	√				
R16a		√				
R16b		√				
R16c		√		√		
R16d		√	√			
R17	√				√	
R17a					√	
R17b					√	
R17c					√	
R17d					√	√

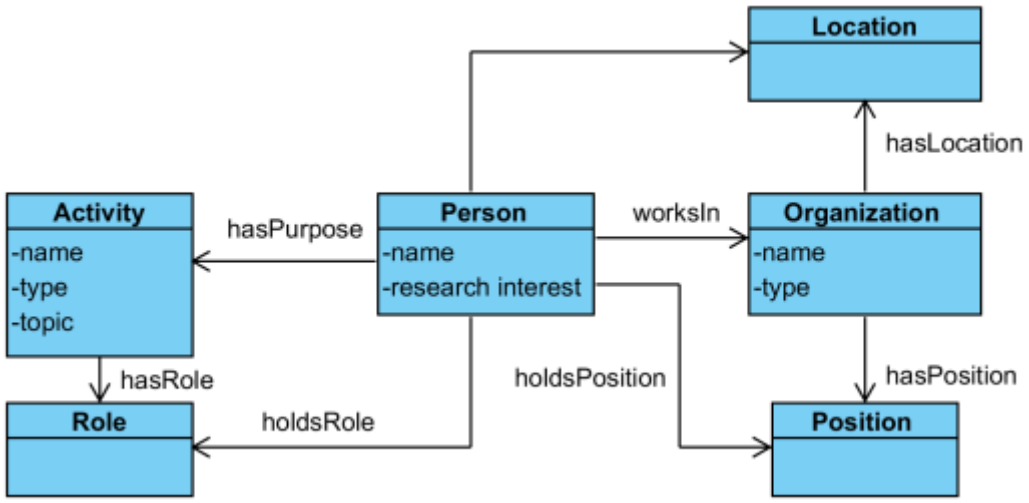

Figure 5: The LiMDAC user profile conceptual model

Figure 6 shows the proposed LiMDAC user profile model in terms of a linked data vocabulary. Again, the model capitalizes on popular linked data vocabularies by reusing existing concepts instead of defining new ones. In particular, it capitalizes on the following linked data vocabularies:

- The FOAF vocabulary [43] that defines the agent.
- The Organization vocabulary [68] that is used to describe organizational structures.
- The PROV Ontology [69] that is used to model provenance information.
- Finally, the SKOS vocabulary [45] that is used to define classifications.

This is a pre-print version of the following article: Kamateri, E., Kalampokis, E., Tambouris, E., and Tarabanis, K. The Linked Data Access Control Framework. *Journal of Biomedical Informatics*, Vol.50, pp. 213-225 http://dx.doi.org/10.1016/j.jbi.2014.03.002

Available for download from http://kalampok.is

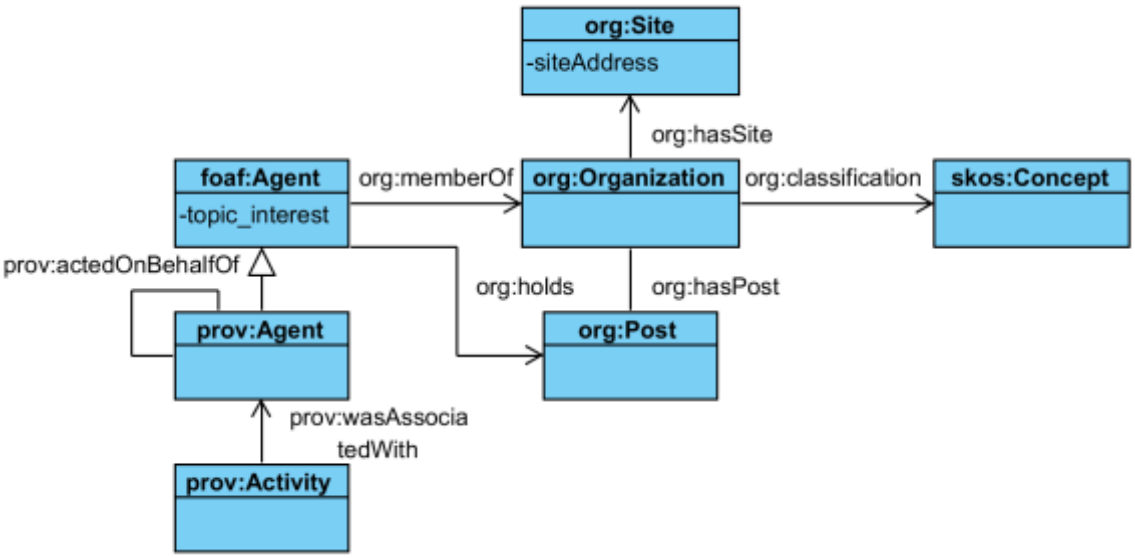

Figure 6: User metadata model

We use the Organization vocabulary to define the Organization (*org:Organization*) that a data consumer works for (*org:memberOf*). The class *org:Post* represents the position that the data consumer holds in the Organization. The *org:Site* denotes the office or other premise at which the Organization is located. In addition, the Site uses the property *org:siteAddress* to indicate the address of the Site. The SKOS vocabulary is used to define the classification of the Organization within some classification scheme (*org:classification*). Furthermore, the PROV concept *prov:Activity* associates an agent (*prov:Agent,)* with a action/activity that he plans, or is responsible to conduct on the extracted data.

In the example of childhood obesity, we could consider two clinical researchers, researcher A and researcher B.

Researcher A works as a biologist. His current occupation is principal trial investigator in AHRQ while his latest research interests are related to the effects of the obesity in children's life. Researcher A has decided to undertake a clinical research related to the effectiveness of a new treatment to children' obesity for children presenting diabetes. To this end, he needs a sufficient number of subjects aged under 18 and diagnosed with diabetes (patient-related research).

Researcher B works as an endocrinologist in the Children's Hospital of Philadelphia. His research interests are related to health habits. Recently, researcher B has started an epidemiological study about health habits and their effects on children's health. In order to receive statistically significant results from the statistical analyses, he needs a sufficient amount of medical data derived from youth patients aged under 18 and containing information about BMI, regular exercise and/or vegetable consumption (epidemiological study).

This is a pre-print version of the following article: Kamateri, E., Kalampokis, E., Tambouris, E., and Tarabanis, K. The Linked Data Access Control Framework. *Journal of Biomedical Informatics*, Vol.50, pp. 213-225 http://dx.doi.org/10.1016/j.jbi.2014.03.002

Available for download from http://kalampok.is

Taking into account all this information, the user profiles of the two data consumers are presented in Figure 7.

```
:ResearcherA a foaf:Agent, provAgent;                        #metadata about the data consumer A
    foaf:topic_interest     :ChildrenObesity;
    org:memberOf            :AgencyforHealthcareResearchandQuality;
    org:holds               :Biologist.
:PatientOrientedResearchforDiabetes a prov:Activity;         #metadata about the purpose/activity
    prov:wasAssociatedWith  :ResearcherA;
    rdfs:comment            "ResearcherA should perform a patient-oriented research about obesity for
                            children having diagnosed with diabetes".
:AgencyforHealthcareResearchandQuality a org:Organization;   #metadata about the organization A
    foaf:homepage           "http://www.ahrq.gov/";
    org:hasSite             [org:siteAddress "540 Gaither Road, Rockville, MD 20850";]
    org:classification      :FederalAgency.
:ResearcherB a foaf:Agent, prov:Agent;                       #metadata about the data consumer B
    foaf:topic_interest     :HealthHabits;
    org:memberOf            :ChildrenHospitalPhiladelphia;
    org:holds               :Endocrinologist.
:EpidemiologicalStudyforHealthHabits a prov:Activity;        #metadata about the purpose/activity
    prov:wasAssociatedWith   :ResearcherB;
    rdfs:comment             "ResearcherB should perform an epidemiological study about health habits
                             and their effects on children's health".
:ChildrenHospitalPhiladelphia a org:Organization;            #metadata about the organization A
    foaf:homepage            "http://www.chop.edu/";
    org:hasSite              [org:siteAddress "34th Street and Civic Center Boulevard, Philadelphia,
                             PA 19104";]
    org:classification       :Hospital.
:HealthOrganization a skos:Collection                        #A skos collection for organizations
    skos:member :Hospital, :ClinicalSite, :ResearchInstitute, :PharmaceuticalCompany, :FederalAgency...
```

Figure 7: User profiles

4.4 The LiMDAC Access Policy Model

Access policies define the data to be protected and to whom access is granted or denied. Thus, each access policy consists of two parts. The first includes the metadata profile of the medical data that will be protected and the second the profile of data consumers that are allowed (or not) to have access to the data.

In the LiMDAC framework we adopt a simplified access policy approach that enables us to assign access policies dynamically on linked data cubes sharing common characteristics. In particular, the LiMDAC access policy model specifies a) an RDF pattern based on the LiMDAC metadata model to limit the application of policies only to data cubes annotated with those metadata and b) a user pattern based on the LiMDAC user profile model to give the access permission only to users described with those attributes.

Figure 8 depicts the LiMDAC access policy linked data model that consists of the following concepts:

- The Dataset: It defines the dataset, where the access policy is applied. The dataset is usually a store containing all linked data cubes of a provider.

- The Data Cube Space: It describes the data cubes, in which the access policy applies. This is achieved through an RDF pattern based on the LiMDAC metadata model that should be satisfied by the metadata of a data cube. If the metadata contain this pattern then the access policy is applied to the data cube.
- The Access Space: It defines the data consumers for which the access policy applies. This is achieved through an RDF pattern based on the LiMDAC user profile model that specifies a user profile.
- The Access Control Privilege: It defines both types of permissions (i.e., grantAccess/denyAccess) and permitted operations (read/write/update). We define it as a subtype of the acl:Access.

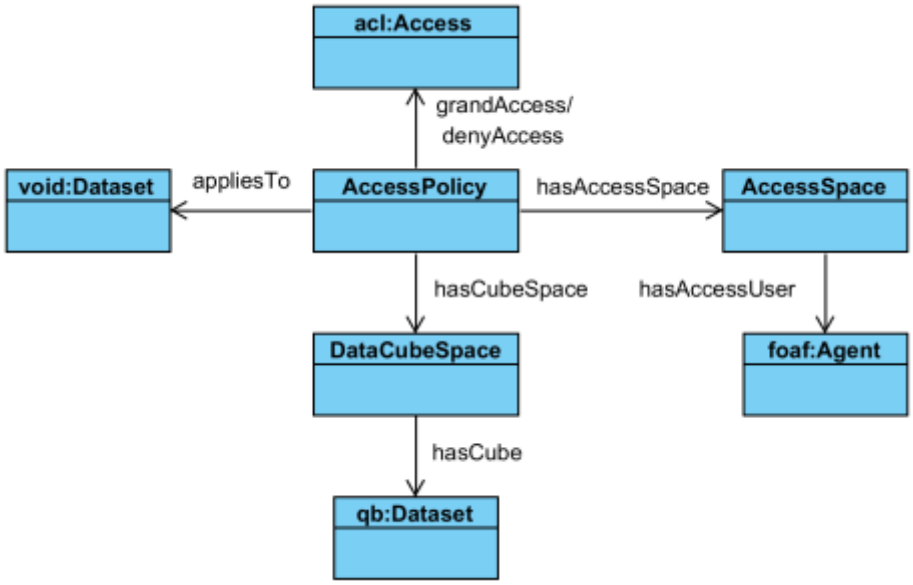

Figure 8: Access policy model

In the example of childhood obesity, each paediatric hospital serves as an individual data provider maintaining in a local repository a large number of linked data cubes coming from various clinical researches. Based on the proposed access policy model, each data provider creates access policies to make available the linked data cubes only to authorized data consumers and under specific conditions. Consider the case of the Children's Hospital of Colorado. Using the proposed access policy model, it creates the following access policies:

1. The data cubes derived from the childhood obesity study and sponsored by AHRQ can be accessed by any user working at AHRQ.
2. The data cubes having in their structure the BMI dimension are authorized for access to data consumers working at the Children's Hospital of Philadelphia and being endocrinologists.

Figure 9 presents the RDF representation of the second access policy.

This is a pre-print version of the following article: Kamateri, E., Kalampokis, E., Tambouris, E., and Tarabanis, K. The Linked Data Access Control Framework. *Journal of Biomedical Informatics*, Vol.50, pp. 213-225 http://dx.doi.org/10.1016/j.jbi.2014.03.002

Available for download from http://kalampok.is

```
:AccessPolicy2 a :AccessPolicy;
    rdfs:label              "This access policy enables data consumers working at Children's Hospital of
                             Philadelphia and being endocrinologists to have 'read' access to linked
                             data cubes having in their structure the BMI dimension";
    :grandAccess            acl:read;
    :hasCubeSpace           [
                            :cubeQuery "ASK {
                            ?cube <http://purl.org/dc/elements/1.1/creator> : ChildrenHospitalColorado.
                            ?cube <http://purl.org/linked-data/cube#structure> ?struct.
                              ?struct <http://purl.org/linked-data/cube#component> ?comp.
                                ?comp <http://purl.org/linked-data/cube#dimension> :BMI.}"^^xsd:string.
     :hasAccessSpace        [
                            :accessQuery  "ASK {
                            ?x <http://www.w3.org/TR/vocab-org/#org:holds> :Endrocrinologist.
                            ?x <http://www.w3.org/TR/vocab-org/#org:memberOf>
                                :ChildrenHospitalPhiladelphia.}"^^xsd:string.
```

Figure 9: Example of access policy

4.5 The LiMDAC Architecture

Figure 10 illustrates an architecture that exploits and orchestrates the three LiMDAC models to enable controlling access to medical data. Apart from the LiMDAC models, the architecture includes an Authorization Mechanism module and an Authorization Interface module.

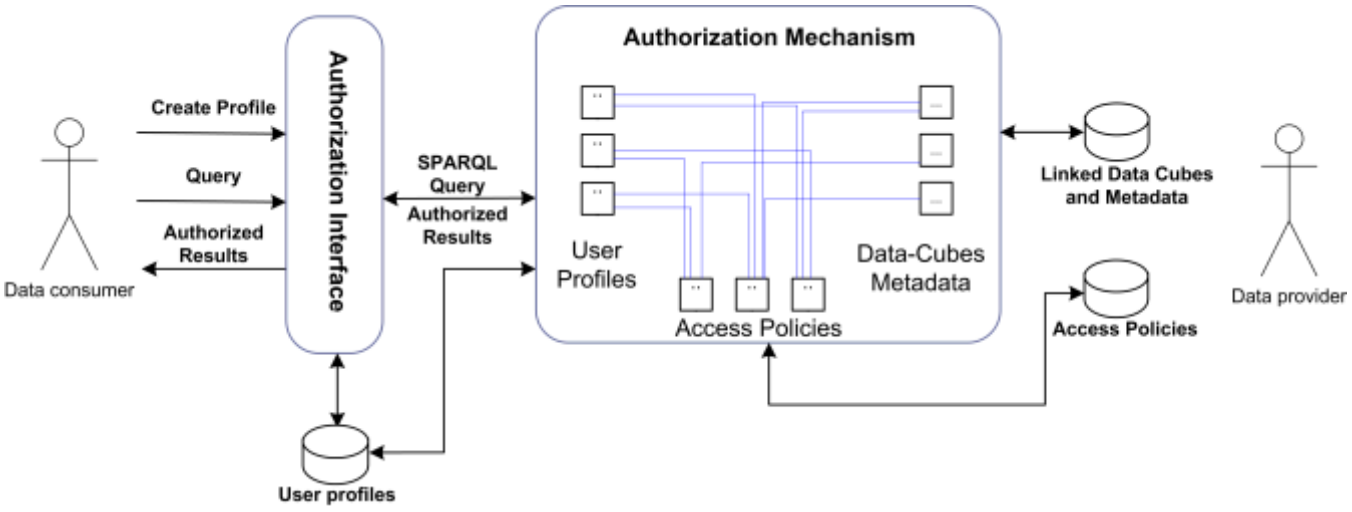

Figure 10: The Linked Medical Data Cubes solution

The Authorization Interface lies between data consumers and providers. It enables data consumers to i) create user profiles based on the LiMDAC user profile model and ii) search for distributed medical data. The data consumer defines the purpose of accessing medical data being either patient-oriented research or epidemiological study. The search criteria are based on the dimensions of the data cubes that are stored in the distributed RDF data stores. Next to each search criteria there is a field for selecting its value from a drop-down list of available codes.

This is a pre-print version of the following article: Kamateri, E., Kalampokis, E., Tambouris, E., and Tarabanis, K. The Linked Data Access Control Framework. *Journal of Biomedical Informatics*, Vol.50, pp. 213-225 http://dx.doi.org/10.1016/j.jbi.2014.03.002
Available for download from http://kalampok.is

Moreover, the Authorization Interface translates data consumers' queries into SPARQL queries and passes them on to the Authorization Mechanism. Initially the Authorization Mechanism retrieves the access policies from distributed data providers. Thereafter, the Authorization Mechanism checks whether the profile of the data consumer matches the user profile defined by each access policy. In case of success, the Authorization Mechanism creates and sends SPARQL queries to distributed data providers. The queries search for data that match both the data consumer's query and the satisfied access policies. In case of success, the resulted datasets are returned to the data consumer via the Authorisation Interface.

The result of this process is either i) (*for patient-oriented research purposes)* the number of patients meeting specific criteria, along with the name of the data provider publishing these data, or ii) (*for epidemiological studies purposes)* the respective linked data cubes.

5. Proof-of-Concept Implementation and Evaluation of the Platform

Based on the framework a proof-of-concept platform is developed and its functionality and performance are evaluated based on two usage scenarios. The platform implements the Authorization Mechanism and the Authorization Interface and exploits the LiMDAC models. For its development we used the Jena Framework[2].

Following on the childhood obesity example, we consider that there are four distributed data providers, each corresponding to a different hospital. Each data provider stores linked data cubes produced in the course of different clinical studies (including the childhood obesity example). Data providers also store metadata and access policies for the cubes based on the LiMDAC metadata and access policy models respectively. To emulate this setting, we have used four distributed RDF data stores (implemented using a Fuseki[3] SPARQL server – part of the Jena Framework). A total of 120.000 linked data cubes are stored in each site. Each cube is structured based on 3 out of 10 dimensions that have been selected for each clinical study. We use around 40 dimensions in total with each receiving values from a pre-defined coded list. Moreover, there exist around 1.5 million triples expressing data cubes' metadata that will be searched by access policies.

In order to evaluate the platform we investigate two usage scenarios. In the example of childhood obesity, researcher A (as defined in Section 4.3) needs medical data for patient-oriented research (purpose) and selects the age dimension with value smaller than 18 and the diabetes dimension. On the other hand, researcher B needs medical data for epidemiological purposes and selects the BMI, Regular_Exercise,

[2] http://jena.apache.org

[3] http://jena.apache.org/documentation/serving_data/

Vegetable_consumption and age dimensions. In the value field of the age dimension he indicates smaller than 18 because so that to investigate effects of health habits in young people.

In the first scenario, the system returns the number of subjects meeting the specified criteria. Specifically, Researcher A is now provided access to data derived from the childhood obesity study, which his organisation has funded.

In the second scenario, the system returns data cubes meeting all specified criteria. We assume that this data comes from the Children's Hospital of Philadelphia and the Children's Hospital of Colorado. Researcher B works at the Children's Hospital of Philadelphia, thus he is provided access to data coming from his organization. Moreover, he is provided access to data coming from the Children's Hospital of Colorado, since the Children's Hospital of Colorado has defined a special access policy so as endocrinologists coming from the Children's Hospital of Philadelphia are permitted access to data cubes related to BMI.

In order to evaluate the proof-of-concept platform, we performed functionality, as well as performance and scalability, testing based on two usage scenarios. In the functionality evaluation testing, we validated and verified that all requirements are implemented by the LiMDAC framework and the proof-of-concept platform.

In order to evaluate the performance and scalability of the platform, we have conducted two sets of experiments measuring the response time. At each running, we have executed a complex (3 search criteria) and a simple (1 search criteria) user query. Table 5 presents an example of these queries. The main difference between them is the number of dimensions requested. In our example, diabetes, BMI and hypertension dimensions are requested in the complex query while diabetes is requested in the simple one:

Table 5: Complex and simple queries

Complex Query Example	**Simple Query Example**

SELECT ?diabetes ?bmi ?hyper
WHERE {
?cube qb:structure ?struct .
?struct qb:component ?comp
?comp qb:dimension :Diabetes
?comp qb:dimension :Hypothyroidism
?comp qb:dimension :BMI
}
	SELECT ?cube
	WHERE {
	?cube qb:structure ?struct .
	?struct qb:component ?comp
	?comp qb:dimension :Diabetes
	}

At the first experimental setting (Table 6), each provider has 1,000 access policies (we assume that 10% of the access policies are satisfied by the user profile). The measurement is repeated for 1,200, 12,000 and 120,000 cubes per provider. The response time varies from 1 sec to 23 sec for the simple user query while for the complex query the response time are much sorter (vary from 0.1 sec to 0.6 sec). The more search criteria are used the less cubes are matched and thus, the quicker the response is. At the second experimental setting (Table 7), the number of cubes is constant, namely 120,000. Here, the variable parameter is the number of access policies per provider. We repeated the measurement for 10, 50, 100, 1,000, 5000 and 10,000 access policies per provider (assuming that 10% of the access policies are satisfied). Again at each case a simple and a complex user query are executed. For the simple query the response times are between 0.1 sec and 23 sec for realistic scenarios until 1000 access policies, while the response time increases significantly as the access policies increase to simulate a web scale. For the complex query the response times are between 0.1 sec and 0.3 sec for realistic scenarios (i.e., until 1000 access policies) while the response time reaches 5 sec and 12 sec for imaginary scenarios that simulate a web scale (5000 and 10000 access policies).

Based on the performance evaluation results, we observe that the authorization mechanism is more sensitive to the number of available access policies than to the number of available cubes. At a real word scenario, it is expected to have a huge number of cubes but the number of policies is not expected to be so high. Based on that, we conclude that the proposed approach is expected to work efficiently for web scale data.

Table 6: First experimental setting with 1,000 access policies per provider

#cubes	Simple query time (ms)	Complex query time (ms)
1,200	832	119
12,000	3,816	262
120,000	23,249	657

Table 7: Second experimental setting with 120,000 cubes per provider

#access policies	Simple query time (ms)	Complex query time (ms)
10	180	150
50	1,672	297
100	2,117	319
1,000	23,249	657
5000	562,141	5,438
10,000	1,547,029	12,239

6. Discussion

Several frameworks have been recently proposed to enable sharing and reuse of medical data. These frameworks consist of data models, processes, architectures, and software tools to address various

challenges of the data value chain, including ethical and legal ones. These challenges impose access constraints that are related to (a) the privacy of patients, (b) the integrity of data, and (c) the availability of data to authorized only persons. Although existing frameworks provide adequate solutions to the first two constraints, they usually follow simple approaches to the latter type of constraints. For example, a common manner to grant access to authorized users is through a static list of IP addresses or particular people.

In this paper, we introduced the Linked Medical Data Access Control (LiMDAC) framework that capitalizes on Linked Data technologies to enable controlling access to linked aggregated medical data across distributed sources with diverse access constraints. Although, the main focus of LiMDAC is access control in a Web-based environment, it caters for all diverse requirements of medical data sharing.

Medical data sharing frameworks usually employ the creation of new data out of the original medical records in order to ensure data integrity. For example, in the i2b2 framework a copy of the medical record is created and thus investigators are free to "clean" and manipulate it for their own purposes with other i2b2 software. These frameworks also employ techniques that prevent unauthorized disclosure of patients' information or identity. For example, a popular technique concerns returning aggregate numbers of patients that satisfy a query to the record-level medical data. In LiMDAC, data cubes that contain aggregate numbers of patients are created from the actual medical data in order to ensure both the privacy of the patients and the integrity of the data. Although the creation of the data cubes is out of the scope of the current LiMDAC version, the details of creating and querying cubes in a linked data environment are described.

The scope of the existing frameworks ranges from single data repositories to multiple distributed data sources inside an institution or at a cross-institutional setting. A suitable data model is critical in these cases so that medical record data and clinical trial data can fit together and thus diseases, genes, and outcomes can be related to each other. For example, the i2b2 "star schema" data model is used to instantiate at a project level the raw medical record data. In the case of LiMDAC the RDF data cube vocabulary, which is a W3C standard, is used to model the aggregated data. In addition, the adoption of the Linked Data paradigm enables reuse of existing widely used vocabularies, datasets and code lists and thus maximizes interoperability and alleviates the burden of semantic alignment. This enables users of LiMDAC to perform queries in multiple sites without having to be aware of the underlying schema of the other sites.

The majority of recent frameworks is based on a Service Oriented Architecture and employ Web Services standards such as SOAP protocol (e.g., in FURTHeR and i2b2) and paradigms such as the Enterprise Service Bus (e.g., in iDASH framework). LiMDAC is to the best of our knowledge the first framework that adopts the Linked Data paradigm to develop the underlying infrastructure for sharing and reusing medical data. This

will enable the easy integration of medical data with other data on the Web (both medical and third party data e.g., government data). This is expected to enhance the possibility to gain innovative insights in epidemiological studies.

Although several Linked Data access control frameworks have been recently proposed, they all suffer from several shortcomings when applied to RDF data cubes that represent aggregated medical data. Medical data providers assign access constraints to cubes, which are frequently updated, even once per week. So, there is a need to assign access constraints dynamically based on domain specific metadata. However, current Linked Data access control frameworks do not support this need. Moreover, most of these approaches restrict parts of the RDF graph having specific RDF characteristics (e.g., triples containing a particular property), or associate access policies to specific RDF data. An RDF data cube can be considered as a small RDF graph made up of many triples. Thus, access policies should apply on the data cube granularity level and access should be restricted based on cubes' metadata instead of RDF properties. LiMDAC satisfies these requirements by employing the SPARQL ASK form in cubes' metadata. It is also important to mention that the LiMDAC framework has been proposed as a simplified solution to address availability constrains on linked medical data cubes.

From a technical perspective, the results of our initial performance evaluation are promising, as they show only a small increase in query processing time and a linear increase as the number of data cubes and satisfied access policies grows. A significant delay has been noticed for simple queries only when the number of access policies exceeds a specific level (5000 and 10000). However, this is not a realistic scenario according to interviews conducted with clinical stakeholders for extracting requirements about availability constraints. Indeed, they suggested that an average of 5 to 20 access policies will be created for each dataset of cubes.

7. Conclusion

In this paper, we introduced the Linked Medical Data Access Control (LiMDAC) framework that capitalizes on Linked Data technologies to enable controlling access to linked aggregated medical data across distributed sources with diverse access constraints. LiMDAC consists of three Linked Data models, namely the LiMDAC metadata model for describing medical data, the LiMDAC user profile model for describing data consumers, and the LiMDAC access policy model for medical data. It also includes an architecture that exploits these models. Based on the framework, we developed a proof-of-concept platform and evaluated its functionality and performance by employing two usage scenarios.

The LiMDAC framework confronts the problems of current linked data access control frameworks providing granularity in data cube structure layer. In particular, it exploits data cubes' metadata to restrict access on cubes satisfying specific characteristics.

In this paper we focused on the application of the LiMDAC framework to simple access policies and queries. In future work, we will look into handling more expressive access policies with more than 3 search criteria, while we will cover also duplication issues that arise when two contrary access policies are applied in the same data. Furthermore, we plan to further investigate the problems of licensing and the terms of medical data reuse. In addition, we will investigate issues that arise when creating and publishing medical data as Linked Data. Last, we are planning to extend the platform's functionality (e.g., to enable data providers to create access policies) as well as to beta test the platform with the help of collaborating organizations.

Acknowledgments:

This work is partially funded by the European Commission within the 7th Framework Programme in the context of the ICT project Linked2Safety (http://www.linked2safety-project.eu/) under grand agreement No. 288328. The authors would like to thank the whole Linked2Safety consortium and especially the trial partners of the project (CING, CHUV and ZEINCRO) as well as Mr. Dimitrios Zeginis that contributed to the implementation of the proof-of-concept platform. Last, we thank the editor and two anonymous reviewers for their constructive comments, which helped us to improve the manuscript.

References

[1] US Department of Health and Human Services, Public Health Service, Grant Application (PHS 398), Available at: http://grants.nih.gov/grants/funding/phs398/phs398.pdf, Revised 2012.

[2] A. Burgun, O. Bodenreider, Accessing and integrating data and knowledge for biomedical research, Yearbook of Medical Informatics. 3 (2008) 91-101.

[3] M.G. Weiner, P.J. Embi, Toward reuse of clinical data for research and quality improvement: The end of the beginning?, Annals of internal medicine. 151(5) (2009) 359-360.

[4] J.C. Maro, R. Platt, J.H. Holmes, B.L. Strom, S. Hennessy, R. Lazarus, J.S. Brown, Design of a national distributed health data network, Annals of internal medicine 151 (5) (2009) 341-344.

[5] H.U. Prokosch, T. Ganslandt, Perspectives for Medical Informatics. Reusing the electronic medical record for clinical research, Methods Inf Med, 48(1) (2009) 38-44.

[6] P.B. Jensen, L.J. Jensen, S. Brunak, Mining electronic health records: towards better research applications and clinical care, Nature Reviews Genetics, 13(6) (2012) 395-405.

[7] E.C. Lau, F.S. Mowat, M.A. Kelsh, J.C. Legg, N.M. Engel-Nitz, H.N. Watson, H.L. Collins, R.J. Nordyke, J.L. Whyte. Use of electronic medical records (EMR) for oncology outcomes research: assessing the comparability of EMR information to patient registry and health claims data. Clinical Epidemiology, 3(1) (2011) 259–272.

[8] N. Anderson, A. Abend, A. Mandel, E. Geraghty, D. Gabriel, R. Wynden, et al., Implementation of a deidentified federated data network for population-based cohort discovery, Journal of the American Medical Informatics Association, 19(e1) (2012) e60-e67.

[9] A.L. Sherborne, K. Hemminki, R. Kumar, C.R. Bartram, M. Stanulla, M. Schrappe, et al, Rationale for an international consortium to study inherited genetic susceptibility to childhood acute lymphoblastic leukemia, Haematologica, 96 (2011) 1049–1054.

[10] T. Tong, H. Zhao, Practical guidelines for assessing power and false discovery rate for a fixed sample size in microarray experiments, Statistics in medicine, 27(11) (2008) 1960-1972.

[11] A.J. McMurry, S.N. Murphy, D. MacFadden, G. Weber, W.W. Simons, J. Orechia, et al., SHRINE: Enabling Nationally Scalable Multi-Site Disease Studies, PloS one, 8(3) (2013) e55811.

[12] C. Newton-Cheh, T. Johnson, V. Gateva, M.D. Tobin, M. Bochud, L. Coin, M. Orho-Melander, Genome-wide association study identifies eight loci associated with blood pressure, Nature genetics, 41(6) (2009) 666-676.

[13] D. Levy, G.B. Ehret, K. Rice, G.C. Verwoert, L.J. Launer, A. Dehghan, C.M. van Duijn, Genome-wide association study of blood pressure and hypertension, Nature genetics, 41(6) (2009) 677-687.

[14] R.D. Kush, E. Helton, F.W. Rockhold, C.D. Hardison, Electronic health records, medical research, and the Tower of Babel, New England Journal of Medicine, 358(16) (2008) 1738-1740.

[15] P. Taylor, Personal genomes: when consent gets in the way, Nature, 456 (7218) (2008) 32-33.

[16] A. Begoyan, An overview of interoperability standards for electronic health records, USA: society for design and process science, (2007).

[17] W, Goossen, A, Goossen-Baremans, M. van der Zel, Detailed clinical models: a review, Healthcare informatics research, 16(4) (2010) 201-214.

[18] C. Bizer, T. Heath, T. Berners-Lee, Linked data – the story so far, International Journal on Semantic Web and Information Systems (IJSWIS), Special Issue on Linked Data, 5(3) (2009) 1–22.

[19] M. Hausenblas, Exploiting linked data to build web applications, IEEE Internet Computing, 13(4) (2009) 68–73.

[20] K, Benitez, B. Malin, Evaluating re-identification risks with respect to the HIPAA privacy rule, Journal of the American Medical Informatics Association, 17(2) (2010) 169-177.

[21] E. J. Ludman, S. M. Fullerton, L. Spangler, S. B. Trinidad, M. M. Fujii, G. P. Jarvik, et al., Glad you asked: participants' opinions of re-consent for dbGap data submission, Journal of empirical research on human research ethics: JERHRE, 5(3) (2010) 9.

[22] N. Anderson, K. Edwards, Building a chain of trust: using policy and practice to enhance trustworthy clinical data discovery and sharing. In Proceedings of the 2010 Workshop on Governance of Technology, Information and Policies, ACM, 2010, pp. 15-20.

[23] L.C. Huang, H.C. Chu, C.Y. Lien, C.H. Hsiao, T. Kao, Privacy preservation and information security protection for patients' portable electronic health records. Computers in Biology and Medicine, 39(9) (2009) 743-750.

[24] R.V. Dhopeshwarkar, L.M. Kern, H.C. O'Donnell, A.M. Edwards, R. Kaushal, Health care consumers' preferences around health information exchange, The Annals of Family Medicine, 10(5) (2012) 428-434.

[25] R.C. Barrows, P.D. Clayton, Privacy, confidentiality, and electronic medical records, Journal of the American Medical Informatics Association, 3(2) (1996) 139-148.

[26] J.L.F Alemán, I.C. Señor, P.Á.O. Lozoya, A. Toval, Security and privacy in electronic health records: A systematic literature review, Journal of biomedical informatics (2013).

[27] L. Lefort, H. Leroux, Design and generation of Linked Clinical Data Cubes, In 1st International Workshop on Semantic Statistics (SemStats), 2013.

[28] K. Perakis, T. Bouras, D. Ntalaperas, P. Hasapis, C. Georgousopoulos, R. Sahay, et al. Advancing Patient Record Safety and EHR Semantic Interoperability. IEEE International Conference on Systems, Man, and Cybernetics, 2013.

[29] K. Caine, R. Hanania, Patients want granular privacy control over health information in electronic medical records. Journal of the American Medical Informatics Association, 20(1) (2013) 7-15.

[30] A. Antoniades, C. Georgousopoulos, N. Forgo, A. Aristodimou, F. Tozzi, P. Hasapis, et al., Linked2Safety: A secure linked data medical information space for semantically-interconnecting EHRs advancing patients' safety in medical research, In IEEE 12th International Conference on Bioinformatics & Bioengineering (BIBE), 2012, IEEE. 2012, pp. 517-522.

[31] B. Malin, D. Karp, R.H. Scheuermann, Technical and policy approaches to balancing patient privacy and data sharing in clinical and translational research, Journal of investigative medicine: the official publication of the American Federation for Clinical Research, 58(1) (2010) 11.

[32] I.S. Kohane, S.E. Churchill, S.N. Murphy, A translational engine at the national scale: informatics for integrating biology and the bedside, Journal of the American Medical Informatics Association, 19(2) (2012) 181-185.

[33] S.N. Murphy, G. Weber, M. Mendis, V. Gainer, H.C. Chueh, S. Churchill, I. Kohane, Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2), Journal of the American Medical Informatics Association, 17(2) (2010) 124-130.

[34] T.A. Drake, J. Braun, A. Marchevsky, I.S. Kohane, C. Fletcher, H. Chueh, et al., A system for sharing routine surgical pathology specimens across institutions: the Shared Pathology Informatics Network, Human pathology, 38(8) (2007) 1212-1225.

[35] G.M. Weber, S.N. Murphy, A.J. McMurry, D. MacFadden, D.J. Nigrin, S. Churchill, I.S. Kohane, The Shared Health Research Information Network (SHRINE): a prototype federated query tool for clinical data repositories. Journal of the American Medical Informatics Association, 16(5) (2009) 624-630.

[36] G.M. Weber, Federated queries of clinical data repositories: the sum of the parts does not equal the whole, Journal of the American Medical Informatics Association, 20(e1) (2013) e155-e161.

[37] Kreator, D. B., Wei, D., Gadde, S., Bockholt, J., Grethe, J. S., Marcus, D., Aucoin, N., and Ozyurt, I. B., Derived Data Storage and Exchange Workflow for Large-Scale Neuroimaging Analyses on the BIRN Grid. *Front Neuroinformatics* 3, 2009, doi: 10.3389/neuro.11.030.2009

[38] Slaymaker, M. Power, D., Russell, D., Wilson, G. and Simpson, A., Accessing and aggregating legacy data sources for healthcare research, delivery and training. In Proceedings of the ACM symposium on Applied computing, Fortaleza, Ceara, Brazil2008 ACM, New York, NY, 1363994, 1317-1324, 2008.

[39] Livne, O. E., Schultz, N. D., & Narus, S. P. (2011). Federated querying architecture with clinical & translational health IT application. *Journal of medical systems*, *35*(5), 1211-1224.

[40] L. Ohno-Machado, V. Bafna, A.A. Boxwala, B.E. Chapman, W.W. Chapman, K. Chaudhuri, et al., iDASH: integrating data for analysis, anonymization, and sharing, Journal of the American Medical Informatics Association, 19(2) (2012) 196-201.

[41] R. Cyganiak, M. Hausenblas, E. McCuirc, Ocial Statistics and the Practice of Data Fidelity. In: Wood, D. (ed.) Linking Government Data, pp. 135-151. Springer (2011).

[42] Heath, T. and Bizer, C. 2011. *Linked Data: Evolving the Web into a Global Data Space Synthesis Lectures on the Semantic Web: Theory and Technology*. Morgan & Claypool.

[43] FOAF Vocabulary Specification (2014) http://xmlns.com/foaf/spec/

[44] DCMI Metadata Terms (2012) http://dublincore.org/documents/dcmi-terms/

[45] SKOS Simple Knowledge Organization System (2009) http://www.w3.org/TR/skos-reference/

[46] T. Tudorache, C. I. Nyulas, N. F. Noy, M. A. Musen, Using Semantic Web in ICD-11: Three Years Down the Road, The Semantic Web–ISWC 2013, Springer Berlin Heidelberg, 2013, pp. 195-211.

[47] J. Malone, E. Holloway, T. Adamusiak, M. Kapushesky, J. Zheng, N. Kolesnikov, et al., Modeling sample variables with an Experimental Factor Ontology, Bioinformatics, 26(8), 2010, 1112-1118.

[48] The UniProt Consortium, Update on activities at the universal protein resource (UniProt) in 2013, Nucleic Acids Research 41 (D1), 2013, D43-D47.

[49] D. Croft, G. O'Kelly, G. Wu, R. Haw, M. Gillespie, L. Matthews, et al., Reactome: a database of reactions, pathways and biological processes, Nucleic acids research, 39(suppl 1), 2011, D691-D697.

[50] E.L. Willighagen, A. Waagmeester, O. Spjuth, P. Ansell, A.J. Williams, V. Tkachenko, J. Hastings, B. Chen, D. J. Wild: The ChEMBL database as linked open data. Journal of cheminformatics, 5(1), 2013, 23.

[51] H.E. Pence, A. Williams, Chemspider: An online chemical information resource, Journal of Chemical Education, 87(11), 2010, 1123-1124.

[52] T. Kelder, M.P. van Iersel, K. Hanspers, M. Kutmon, B.R. Conklin, C.T. Evelo, A.R. Pico: WikiPathways: building research communities on biological pathways. Nucleic acids research, 40(D1), 2012, D1301-D1307.

[53] The RDF Data Cube Vocabulary (2014), http://www.w3.org/TR/vocab-data-cube/

[54] S. Kirrane, A. Abdelrahman, A. Mileo, S. Decker, Secure Manipulation of Linked Data, The Semantic Web–ISWC 2013. Springer Berlin Heidelberg, 2013. pp. 248-263.

[55] J. Hollenbach, J. Presbrey, T. Berners-Lee, Using RDF Metadata to enable Access Control on the Social Semantic Web, In: Proceedings of the Workshop on Collaborative Construction, Management and Linking of Structured Knowledge (CK2009), 2009.

[56] Web Access Controlhttp://www.w3.org/wiki/WebAccessControl

[57] H. Muhleisen, M. Kost, J.C. Freytag, SWRL-based Access Policies for Linked Data. In 2nd Workshop on Trust and Privacy on the Social and Semantic Web, 2010.

[58] G. Flouris, I. Fundulaki, M. Michou, G. Antoniou, Controlling Access to RDF Graphs, In Future Internet-FIS 2010. Springer Berlin Heidelberg, 2010, pp. 107-117.

[59] P. Reddivari, T. Finin, A. Joshi. Policy-Based Access Control for an RDF Store. In IJCAI-07 Workshop on Semantic Web for Collaborative Knowledge Acquisition, 2007.

[60] A. Jain, C. Farkas, Secure resource description framework: an access control model. 11th ACM symposium on Access control models and technologies. ACM, 2006, pp. 121-129.

[61] S. Dietzold, S. Auer. Access control on RDF triple stores from a semantic wiki perspective. ESWC Workshop on Scripting for the Semantic Web, 2006.

[62] L. Kagal, T. Finin, and A. Joshi. A policy-based approach to security for the semantic web. In Proceedings of 2nd International Semantic Web Conference (ISWC'03), LNCS 2870, pages 402-418. Springer, September 2003.

[63] L. Costabello, S. Villata, N. Delaforge. Linked data access goes mobile: Context-aware authorization for graph stores. In LDOW - 5th WWW Workshop on Linked Data on the Web, 2012.

[64] O. Sacco, A. Passant, S. Decker, An Access Control Framework for the Web of Data. In 10th International Conference on Trust, Security and Privacy in Computing and Communications, IEEE, 2011, pp. 456-463.

[65] F. Abel, J.L. De Coi, N. Henze, A.W. Koesling. Enabling advanced and context-dependent access control in RDF stores. The Semantic Web, Springer Berlin Heidelberg, 2007, pp. 1-14.

[66] L.C. Bailey, D.E Milov, K. Kelleher, M.G. Kahn, M. Del Beccaro, F. Yu, et al., Multi-Institutional Sharing of Electronic Health Record Data to Assess Childhood Obesity, PLOS ONE, 8(6) (2013) e66192.

[67] Bosch, T., Cyganiak, R., Gregory, A., Wackerow, J.: DDI-RDF Discovery Vocabulary: A Metadata Vocabulary for Documenting Research and Survey Data. In: LDOW2013, May 14, 2013, Rio de Janeiro, Brazil (2013)

[68] The Organization Ontology (2014) http://www.w3.org/TR/vocab-org/

[69] PROV-O: The PROV ontology (2013) http://www.w3.org/TR/prov-o/

APPENDIX

Table: Requirements related to Availability Constraints

No	Requirements
Requirements regarding the medical data	
Access to medical data is restricted based on metadata about medical data	
1	The content of medical data e.g., data about cancer. The content derives from the variables that have been recorded for each subject in the clinical study
2	The data provider of the medical data e.g., a hospital, a pharmaceutical company or a health agency

3	The clinical study from which medical data has been derived
3a	The title of the clinical study
3b	The research topic of the clinical study e.g., children's obesity
3c	The purpose of the clinical study e.g., for an epidemiological study
3d	The principal investigator of the clinical study
3e	The health institution where the clinical study has been conducted e.g., a clinical site, a hospital, a private medical centre, a clinical laboratory. It can also be denoted as the creator of the clinical study
3f	The contributors of the clinical study if additional subjects were needed
3g	The sponsor of the clinical study e.g., a pharmaceutical company or a public health authority
3h	The period of time that the clinical study has been run
3i	The location where the clinical study has been conducted. This usually refers to the location of the healthcare institute performed the clinical study
Requirements regarding the data consumer	
Access is restricted based on metadata about the data consumer	
13	The name of the data consumer
14	The location/origin of the data consumer e.g., a country
15	The research interest of the data consumer e.g., oncology or metabolic syndrome.
16	The organization where the data consumer is working
16a	The name of the organization
16b	The type of the organization e.g., pharmaceutical company
16c	The location of the organization
16d	The occupation/position in the organization that the data consumer holds e.g., biologist, epidemiologist, or endocrinologist
17	The activity that the data consumer needs to perform. It can also be denoted as the purpose of the data consumer
17a	The name of the activity/purpose
17b	The type of the activity e.g., clinical trial, epidemiological study, publication
17c	The topic of the activity e.g., health habits, breast cancer
17d	The role that the data provider plays during the activity e.g., clinical researcher or data analyst