DATA INTEROPERABILITY THROUGH FEDERATED SEMANTIC METADATA REGISTRIES

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

ALİ ANIL SINACI

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
COMPUTER ENGINEERING

JULY 2014

Approval of the thesis:

DATA INTEROPERABILITY THROUGH FEDERATED SEMANTIC METADATA REGISTRIES

submitted by **ALİ ANIL SINACI** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Computer Engineering Department, Middle East Technical University** by,

Prof. Dr. Canan Özgen _____
Dean, Graduate School of **Natural and Applied Sciences**

Prof. Dr. Adnan Yazıcı _____
Head of Department, **Computer Engineering**

Prof. Dr. Nihan Kesim Çiçekli _____
Supervisor, **Computer Engineering Dept., METU**

Prof. Dr. Asuman Doğaç _____
Co-supervisor, **SRDC Ltd.**

Examining Committee Members:

Prof. Dr. Özgür Ulusoy
Computer Engineering Department, Bilkent University _____

Prof. Dr. Nihan Kesim Çiçekli
Computer Engineering Department, METU _____

Prof. Dr. Ali Hikmet Doğru
Computer Engineering Department, METU _____

Prof. Dr. Ahmet Coşar
Computer Engineering Department, METU _____

Assoc. Prof. Dr. Pınar Karagöz
Computer Engineering Department, METU _____

Date: _____

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: ALİ ANIL SINACI

Signature :

ABSTRACT

DATA INTEROPERABILITY THROUGH FEDERATED SEMANTIC METADATA REGISTRIES

Sınacı, Ali Anıl
Ph.D., Department of Computer Engineering
Supervisor : Prof. Dr. Nihan Kesim Çiçekli
Co-Supervisor : Prof. Dr. Asuman Doğaç

July 2014, 97 pages

In this study, a unified methodology together with the supporting framework for the problem of data interoperability is introduced which brings together the power of metadata registries and semantic web technologies. A federated architecture of semantic metadata registries which are purely based on ISO/IEC 11179 standard leads to the Linked Open Data integration of data element repositories where each element can be uniquely identified, referenced and processed to enable the syntactic and semantic interoperability. Proposed interoperability architecture is applicable to every domain where information extraction and exchange is possible and is a requirement between applications. Although this study takes its motivation from the interoperability requirements between clinical research and clinical care domains focusing on postmarketing surveillance studies, a case study is also presented which applies the proposed solution for the interoperability of electronic business documents. In eHealth, the use of electronic health record systems in clinical care domain

is rapidly increasing and vast amount of data, which is very valuable for clinical research, is accumulating in these systems. During the implementation and deployment of this study, the main objective is to enable automatic information extraction and exchange of data residing in the electronic health record systems of clinical care domain and the data residing in the electronic data capture systems of clinical research domain. As a result; the analysis, implementation and demonstration of the interoperability architecture through federated semantic metadata registries are fully performed to enable the secondary use of electronic health records for post market drug surveillance activities. In addition, the eBusiness case study presents that proposed framework enables automatic data extraction from electronic business documents with the use of semantic metadata registries while eliminating the burden of message translation between different document standards.

Keywords: Interoperability, Metadata Registry/Repository, Semantic Web, Linked Data, Common Data Elements

ÖZ

VERİ BİRLİKTE İŞLERLİĞİNİN FEDERE, ANLAMSAL ÜSTVERİ KÜTÜKLERİ İLE SAĞLANMASI

Sınacı, Ali Anıl
Doktora, Bilgisayar Mühendisliği Bölümü
Tez Yöneticisi : Prof. Dr. Nihan Kesim Çiçekli
Ortak Tez Yöneticisi : Prof. Dr. Asuman Doğaç

Temmuz 2014 , 97 sayfa

Bu çalışmada, üstveri kütüklerinin ve anlamsal ağ teknolojilerinin güçlü oldukları yönler alınarak birlikte işlerlik probleminin çözümüne yönelik, destekleyici altyapısıyla birlike; yeni ve birleştirici bir metodoloji sunulmaktadır. ISO/IEC 11179 standardına uygun olan anlamsal üstveri kütüklerinin oluşturduğu bu federe yapı ile üstveri kütük ve ambarları anlamsal ağ çerçevesinde, bağlı veri yapısına kolayca entegre olmaktadır. Bu kütük ve ambarlarda bulunan her eleman özebir olarak tanımlanabilmekte, erişilebilmekte ve işlenebilmektedir. Bu sayede farklı alanların veri elemanları ve modelleri arasında sentaktik ve anlamsal birlikte işlerlik sağlanmaktadır. Sunulan birlikte işlerlik yapısı bilgi çıkarımı ve değişiminin gerekli ve mümkün olduğu her alan için kullanılabilir. Bu çalışma, motivasyonunu kilinik araştırma ve hasta sağlığı alanlarının birlikte işlerlik gereksinimlerinden almış olduğu halde elektronik iş dokümanlarının birlikte işlerlik problemini çözmeye yönelik

bir vaka çalışması da yapılmıştır. E-Sağlık alanında, elektronik sağlık kayıtlarının sağlık bilgi sistemlerinde kullanılması hızla artmakta ve hastaneler gibi sağlık kuruluşlarında biriken bu elektronik veri kilinik araştırmalar için çok büyük önem arz etmektedir. Bu çalışmanın gerçeklenmesi ve çalışır hale getirilmesindeki ana hedef, sağlık kuruluşlarında bulunan bilgi sistemlerindeki elektronik sağlık kayıtlarının klinik araştırma birimlerindeki bilgi sistemleriyle birlikte işlerliğini sağlamaktır. Sonuç olarak, federe anlamsal üstveri kütükleri üzerinden sağlanan birlikte işlerlik yapısının analizi, geliştirilmesi ve uygulanması e-Sağlıkta; elektronik sağlık kayıtlarının ikincil kullanımı ile ilaç etki takibi çalışmalarında kullanılarak tamamlanmıştır. Ek olarak yapılan e-İş vaka çalışmasıyla, sunulan yapının elektronik iş dokümanlarından otomatik veri çıkarımı yapabileceği deneylenmiş ve bu çözümün farklı standartlar arasında veri geçişi için yapılan mesaj çevirme işlemine üstünlüğü gösterilmiştir.

Anahtar Kelimeler: Birlikte İşlerlik, Üstveri Kütükleri, Anlamsal Web, Bağlı Veri, Ortak Veri Elemanları

To my dear wife, Selcan

ACKNOWLEDGMENTS

I would like to express my sincere gratitude and appreciation to Prof. Dr. Asuman Doğaç for her encouragement and support throughout this study. I would like to thank my supervisor Prof. Dr. Nihan Kesim Çiçekli for her constant support, guidance and friendship. I would also like to convey thanks to the committee members for their valuable comments on this thesis.

I am deeply indebted to my friends Gökçe Banu Laleci Ertürkmen and Suat Gönül whose help, stimulating suggestions and encouragement helped me in all the time of research for and writing of this thesis. I am highly thankful to Anıl Paçacı and Alper Çınar, and all the other colleagues at SRDC Ltd., for their invaluable support throughout this study.

I am deeply grateful to my dear wife Selcan, my brother Cem Berkay and my parents for their love, patience and continued motivating support. Without them, this work could not have been completed.

When I was a teaching assistant in METU-CENG, I had the chance of meeting my valuable friends Erdal Sivri, Can Eroğul, Selma Süloğlu, Hande Çelikkanat, Özgür Kaya, Gökdeniz Karadağ, Onur Deniz, Çelebi Kocair and Sinan Kalkan. They continuously motivated and supported me whenever I needed encouragement in order to successfully continue my study.

Finally, my special thanks go to my friends Birkal, Ferhat, Onur, Yiğityürek, Şerife, Zülfükar, Rojhat, Ufuk, Senem, Özge, Andaç and Orhan Utku for their help, support and cheerful presence through the course of this study.

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7) under grant agreement no ICT-287800, SALUS Project (Scalable, Standard based Interoperability Framework for Sustainable Proactive Post Market Safety Studies).

TABLE OF CONTENTS

ABSTRACT v

ÖZ vii

ACKNOWLEDGMENTS x

TABLE OF CONTENTS xi

LIST OF TABLES xiv

LIST OF FIGURES xv

LIST OF ABBREVIATIONS xviii

CHAPTERS

1 INTRODUCTION 1

1.1 Motivation 5

1.2 Objectives 8

1.3 Summary of the contributions 10

2 BACKGROUND 13

2.1 Metadata & Metadata Registry 13

2.2 ISO/IEC 11179 16

2.3 SALUS Project 19

2.4 IHE Data Element Exchange Profile 20

2.5 Semantic Web & Linked Data 20

3 SEMANTIC METADATA REGISTRY/REPOSITORY 23

3.1 Proposed extensions to ISO 11179 Standard to achieve federated metadata management for semantic interoperability 24

3.1.1 MDRs in the LOD Cloud 24

3.1.2 Linking CDEs to Terminology Systems 26

3.1.3 Linking CDEs to other CDEs 26

3.1.4 Linking CDEs to Extraction Specifications . . 27

3.1.5 Federation through Linked Data Principles . . 30

3.2 Design & Implementation of the Semantic MDR 31

3.2.1 Ontology of the ISO/IEC 11179 Metamodel . . 31

3.2.2 MDR Knowledge Base 32

3.2.3 RESTful interface 34

3.3 Exploiting linked metadata registries for semantic interoperability 35

4 POST MARKETING SAFETY STUDY TOOL 39

4.1 Secondary Use of Electronic Health Records for Postmarketing Surveillance 40

4.2 Objective of the Post Marketing Safety Study Tool (PMSST) 41

4.3 PMSST Use Case 44

4.4 PMSST System Description 48

4.4.1 Data flow between components 49

4.4.2 CDE mappings 51

5 E-BUSINESS CASE STUDY 55

5.1 Semantic Representation of the Core Components . . . 59

5.2 Semantic MDR in eBusiness 63

6 RELATED WORK 69

7 CONCLUSION 75

7.1 Discussion 78

7.1.1 Limitations 79

7.2 Future Work 80

REFERENCES 81

CURRICULUM VITAE 93

LIST OF TABLES

TABLES

Table 3.1 SPARQL script to retrieve the severity information for the Allergy on a Patient. The target content model is SALUS Common Information Model [1, 2] which can serve data through RDF graphs 29

Table 3.2 Mapping of ISO/IEC 11179 metamodel constructs to OWL constructs 32

Table 4.1 Result schema details for the PMSST use case. 46

Table 4.2 Mappings of the Common Data Elements: SDTM – SALUS CDE set – HITSP C154 Data Dictionary 52

Table 5.1 A part of the UN/CEFACT Core Component Library. Core Components are published through spreadsheets. 60

Table 5.2 A part of the N3 representation of the Party. Postal Address element exported from the Semantic Metadata Registry 62

LIST OF FIGURES

FIGURES

Figure 1.1 Semantic Metadata Registries within the Linked Open Data cloud 5

Figure 2.1 Importance of Metadata: Same data annotated through different metadata can lead inconsistencies 14

Figure 2.2 Metadata management in an ISO/IEC 11179 based Metadata Registry 15

Figure 2.3 Decomposition of a data element according to ISO/IEC 11179 17

Figure 2.4 Actors and transactions of the IHE DEX profile. 20

Figure 2.5 Linking Open Data cloud diagram as of September 2011 . . . 22

Figure 3.1 Federated semantic MDR framework. Within the LOD cloud, each MDR maintains a set of CDEs together with the corresponding components and relations. The CDEs are linked to CDEs of other MDRs through KOSs and annotated with terminology systems. . . . 25

Figure 3.2 Annotations and links of a CDE and its Object Class (OC) inside a Semantic MDR. The OC is annotated with a concept (term) from SNOMED-CT which is maintained under BioPortal through owl:sameAs property. The CDE has an "Extraction Specification" which is an XPath expression pointing the exact place of the CDE in the HL7 CCD models. These annotations and links are modeled through Classification Scheme and Classification Scheme Item elements of the ISO/IEC 11179 metamodel. 27

Figure 3.3 The components of a CDE together with their classifications for the LOD links to other CDEs and components residing in a different Semantic MDR. The CDE - AE.AEREL.Text - has a skos:exactMatch relation with the CDE - AdverseEventRelation - in the Semantic MDR which holds the BRIDG CDEs. If needed, CDE mappings to other CDEs can be given through a Context in which some pre-conditions and rules can be specified. 28

Figure 3.4 High-level view of the architecture of the Semantic MDR Service Layer. At the bottom, there is a triple store serving as a backend for the MDR Knowledge Base. Above the triple store, there is a 3 layered API to perform semantic operations on this Triple Store. Semantic Data Manipulation API is a direct implementation of the ISO/IEC 11179 metamodel which reflects the operations to the underlying RDF graph. MDR API an abstraction layer which hides the complex details of the ISO/IEC 11179 metamodel and provides easy-to-use methods for the data manipulation. 33

Figure 3.5 Step-by-step representation of the scenario in which the federated semantic MDR framework is used for the interoperability of clinical research and clinical care domains. A properly annotated study design document can be automatically populated through the information retrieval process of HL7 CCD based content models with the help of the Federated Query Service. The service makes use of the simple REST interfaces of the Semantic MDRs 38

Figure 4.1 Overall architecture of the PMSST integrated with the Semantic MDR based interoperability approach. 43

Figure 4.2 A snapshot of PMSST while the researcher defines a result schema. On the right hand side, domains of SDTM forms a circle; if selected, then CDEs of that domain forms the circle. On the left hand side, a schema item: Date_HbA1C_Average1YBeforeACS is created out of 4 SDTM elements. Below that, a list of other schema items are shown. 49

Figure 4.3 Step-by-step representation of the data flow between different components. A clinical researcher uses PMSST in order to define a result schema so that when patient data is retrieved from the underlying EHR source(s), data will be automatically transformed to that schema. 50

Figure 5.1 Interplay of the major eBusiness document standards. Electronic document exchange is problematic across boundaries. . . 58

Figure 5.2 Decomposition of a Core Component – Address. Postcode. Code – according to the Object Class, Property, and Value Domain constructs of the ISO/IEC 11179 meta-model. 61

Figure 5.3 The hypothetical ERP (Enterprise Resource Planning) component which implements the introduced automatic data extraction mechanism. This architecture can be extended within the LOD cloud by using several linked Semantic Metadata Registries. . 64

Figure 5.4 Web based graphical user interface of the Semantic Metadata Registry. Core Components can be managed according to the ISO/IEC 11179 metamodel through web-based actions. 66

Figure 7.1 A part of the SAS result executed on the simulated data through PMSST 77

LIST OF ABBREVIATIONS

ADE	Adverse Drug Event
ADR	Adverse Drug Reaction
BRIDG	Biomedical Research Integrated Domain Group
CDE	Common Data Element
CDISC	Clinical Data Interchange Standards Consortium
CDASH	Clinical Data Acquisition Standards Harmonization
SDTM	Study Data Tabulation Model
CRO	Clinical/Contract Research Organization
DEX	Data Element Exchange
eBusiness	Electronic Business
EDC	Electronic Data Capture
EDI	Electronic Data Interchange
eHealth	Electronic Health
EHR	Electronic Health Record
FDA	U.S. Food and Drug Administration
HITSP	Health Information Technology Standards Panel
HL7	Health Level 7
CDA	Clinical Document Architecture
CCD	Continuity of Care Document
RIM	Reference Information Model
I2B2	Informatics for Integrating Biology and the Bedside
IHE	Integrating the Healthcare Enterprise
ISO	International Organization for Standardization
LOD	Linked Open Data
MDR	Metadata Registry/Repository
METeOR	Metadata Online Registry
NIEM	National Information Exchange Model
OMOP	Observational Medical Outcomes Project

OWL	Web Ontology Language
PMSST	Post Marketing Safety Study Tool
REST	Representational State Transfer
RDF	Resource Description Framework
S&I	The Standards and Interoperability (S&I) Framework
SALUS	Scalable, Standard based Interoperability Framework for Sustainable Proactive Post Market Safety Studies
SKOS	Simple Knowledge Organization System
UML	Unified Modeling Language
URI	Uniform Resource Identifier
XML	Extensible Markup Language
XSD	XML Schema Definition

CHAPTER 1

INTRODUCTION

Interoperability is one of the major challenges of the computer and software systems since the early days of their development. Indeed, interoperability is a challenge between individuals, between societies, between countries, continents and will probably include the planets in the future. Hence, the requirements of the interoperability come from real life. Two entities that are capable of transmitting and receiving information and interpreting information need well-organized methods and rules to succeed interoperability so that they can exchange the information and use that information once it has been exchanged.

Data interoperability mostly refers to the upper layers in the computer networking stack models where information exchange between applications are in focus. For example, upmost layers of both TCP/IP and OSI models are named as *Application Layer*. Addressing different data interoperability challenges, there are numerous standards which enables data exchange between applications. Standardization can be counted as the backbone of the data interoperability concept. It is the adoption of the well-established standards which makes the modern Internet possible, powers service oriented architectures on the web and enabling access to the databases of different systems using very divergent information models.

Looking from a general perspective, we can divide the approaches of the researchers to the data interoperability problem into two. The early and more practical one is defining formal interfaces and common information models. These models are depicted with unambiguous rules so that it can be ensured

whether the data conforms to the rules or not. Compared to the second, common data element based approach; this exhibits a top-down vision in which information models, entities and their interactions are strictly modeled according to the pre-defined rules. These rules are sometimes given in text-based documents or formalized with some rule languages. In any case, the objective is to give the interoperating systems precise knowledge about the information being exchanged so that it can be syntactically and semantically processed by the applications.

The second approach is the use of common building blocks for the information models of the applications which need to exchange data. These building blocks are referred as the common data elements and can be defined as the smallest meaningful data container in a given context. In literature, the benefit of adopting common data elements in information systems for the sake of data interoperability is well recognized. The objective is to reduce start-up times and accelerate data sharing among interoperating systems. In different domains and their sub-domains, there are several initiatives and standardization bodies who try to publish abstract common data element definitions so that the interfaces can be defined and the information models can be created out of those common data elements in a bottom-up fashion. For example, National Information Exchange Model (NIEM) [3] is the nation wide data element registry for the United States and achieved a mature level of implementation or Metadata Online Registry (METeOR) [4] is the data element repository of Australia for health, housing and community services statistics and information. Throughout this study, we present a brief analysis of the available common data element models and content models in eHealth and eBusiness domains and will show that the main deficiency is that they are not machine-processable.

ISO/IEC 11179 [5] addresses the management of the semantics of the common data elements: it provides a standard metadata model for the representation of the data elements and provides a methodology for the registration of the descriptions of the data elements through this standard model to a metadata registry. The aim is to facilitate accurate common understanding of the data

elements over time, space and applications. In ISO/IEC 11179 metamodel, a data element is represented through its components, basically through a triple: Object Class, Property and Value Domain. In this study, unambiguous semantics of all ISO/IEC 11179 metamodel components is formally defined. In this way, the management the data elements and their components, and the reuse of these components is also facilitated.

Semantic web technologies and Linked Open Data paradigm are the address of the parallel and recent effort trying to solve the problem of interoperability. In the world of semantics, we can also classify the interoperability approaches as top-down and bottom-up. Designing common ontologies to serve as an intermediary during message translation or building ontology based schemas as the common information models of the interoperating parties is the early top-down approach in this field. Bottom-up approaches start with lower granularity and follow the Linked Data principles which resemble to the common data element based interoperability methodologies. In both cases, semantic web technologies add more power in terms of easy integration, adaptability, extensibility and inference-ability. There are numerous ontologies and knowledge organization systems designed for specific domains and applications in the Linked Open Data world. For instance, Simple Knowledge Organization System (SKOS) [6] or Friend of a Friend (FOAF) [7] are two examples of commonly used knowledge organization systems. Moreover, "health/medical types" is one of the top-level ontologies in schema.org [8] which is commonly linked by the eHealth related ontological models.

In this thesis, a unified methodology and the supporting architecture is introduced which brings together the power of metadata registries and semantic web technologies within the Linked Open Data principles, by eliminating the weak points of top-down and bottom-up approaches in both settings. A federated architecture of semantic metadata registries which are purely based on ISO/IEC 11179 leads to the Linked Open Data integration of data element repositories where each element can be uniquely referenced and processed in order to enable the syntactic and semantic interoperability. Figure 1.1 shows a very high level schematic view of the proposed solution. A triple

store based implementation of the ISO/IEC 11179 standard has been integrated with the extensively used knowledge organization systems and ontologies so that the common data element definitions can be accessed and processed semantically within the Linked Open Data cloud. With this work, we show that the problem of data interoperability can be solved in an upper level with the use of common data element phenomenon [9] on top of semantic web technologies.

We present that the machine-processable definitions of the common data elements across domains can be shared, reused and semantically interlinked with each other to address the semantic interoperability challenge. We introduce the notion of *extraction specifications* which are the implementation specific pointers of the abstract, implementation independent common data element definitions. The connection between the abstract data element and the concrete data (i.e. a message instance) is the extraction specification; when it is executed, it extracts data - associated with the abstract data element definition - from the instance by processing the location pointed by the extraction specification. For example, surname of a person can be considered as an abstract data element and an XPath [10] expression becomes its extraction specification if the information model addressed by that data element is depicted with an XML Schema [11] and the data is serialized in XML.

Proposed interoperability architecture is applicable to every domain where information extraction and exchange is possible and is a requirement between the sub-domains. Although a case study is performed for electronic business document interoperability in eBusiness domain, this study takes its motivation from the interoperability requirement between clinical research and clinical care domains. The objective is to enable automatic extraction and exchange of data residing in the Electronic Health Record (EHR) systems of clinical care domain to be used in the Electronic Data Capture (EDC) systems of clinical research domain.

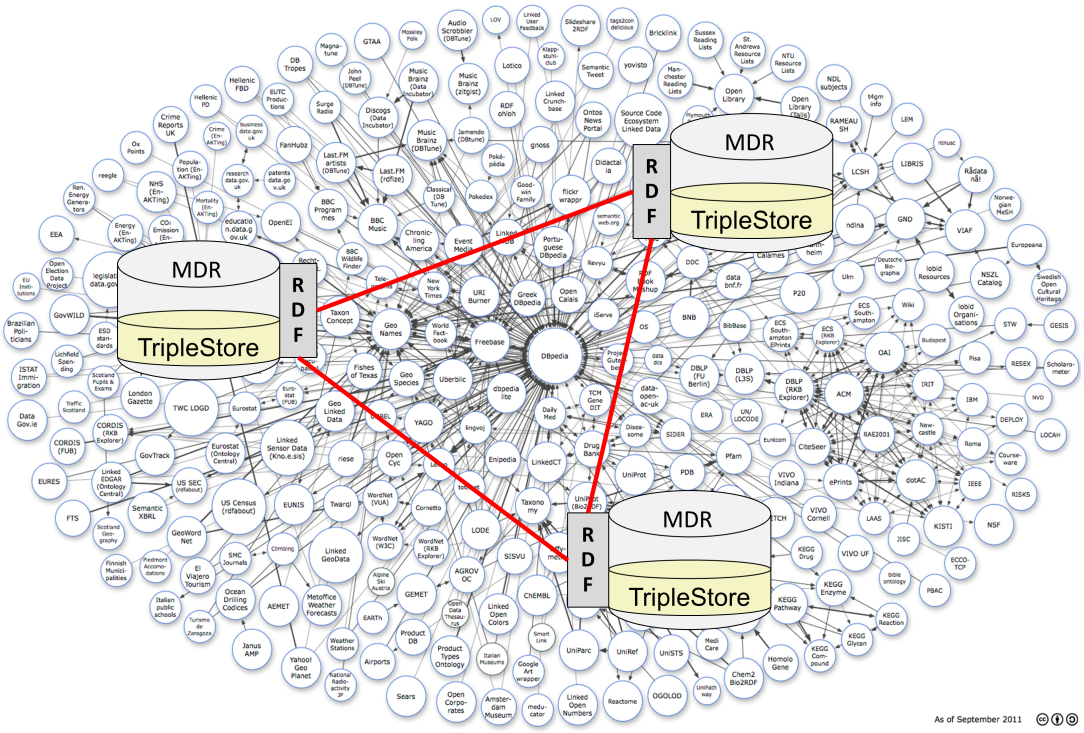


Figure 1.1: Semantic Metadata Registries within the Linked Open Data cloud

1.1 Motivation

In clinical care, as the adoption of Electronic Health Records (EHR) increases, there has been a growing potential of exploiting this data both for enabling better care of patients by sharing the collected data across care organizations, and also for enabling clinical research and quality assessment studies through the secondary use of EHR data. It is a well-accepted fact that, one of the challenges to be addressed to fulfil this great potential is enabling syntactic and semantic interoperability.

A major barrier to repurposing clinical data of EHR systems for clinical research studies (clinical trial design, execution and observational studies) is that information systems in both domains – patient care and clinical research – use different information models and terminology systems. This means that data within each system is stand-alone and not interoperable. As stated by ISO [5], "One of the prerequisites for a correct and proper use and interpretation of data is that both users and owners of data have a common

understanding of the meaning and descriptive characteristics of that data. To guarantee this shared view, a number of basic attributes have to be defined".

In line with this vision, many of the efforts which try to facilitate the exchange of Electronic Health Records for better care of the patient or to enable the secondary use of the EHRs for supporting clinical research and patient safety studies have already been developing common data element models. A few examples can be summarized as follows:

- The Health Information Technology Standards Panel (HITSP) has defined the C154: Data Dictionary Component [12] as a library of the HITSP defined data elements to facilitate the consistent use of these data elements across various HITSP selected standards. These data elements are served through PDF documents and spreadsheets. For example, HITSP C32 [13] which describes the HL7/ASTM Continuity of Care Document (CCD) [14] content for the purpose of health information exchange, marks the elements in CCD document with the corresponding HITSP C154 data elements to establish common understanding of the meaning of the CCD elements.

- The Federal Health Information Model (FHIM) [15] develops a common computationally independent model for EHR systems.

- The Transitions of Care Initiative (ToC) [16] maintains the S&I Clinical Element Data Dictionary (CEDD) [17] as a repository of data elements to improve the electronic exchange of core clinical information among authorized entities in support of meaningful use and improvement in the quality of care. The Query Health [18] initiative extends this data dictionary and establishes Query Health CEDD to enable an architecture for querying distributed EHR systems in order to aggregate healthcare data for collecting quality measures and monitoring disease outbreaks.

- The Clinical Data Interchange Standards Consortium (CDISC) provides common dataset definitions

  (a) in Study Data Tabulation Model (SDTM) [19] for enabling the submission of the result data sets of regulated clinical research

studies to the U.S. Food and Drug Administration (FDA)

(b) and in Clinical Data Acquisition Standards Harmonization (CDASH) [20] for integrating SDTM data requirements into the Case Report Forms.

- The Biomedical Research Integrated Domain Group (BRIDG) [21] developed the Domain Analysis Model (DAM), which harmonizes CDISC data standards with the HL7 Reference Information Model (RIM) [22]. The BRIDG model unifies the concepts in the clinical care and research domains and creates a shared generic representation for each data element.

- Observational Medical Outcomes Project (OMOP) is a public-private partnership which tries to create a Common Data Model (CDM) [23] to be used in pharmacoepidemiology activities specifically for post market drug monitoring.

- Mini-Sentinel [24] is a pilot project to create an active surveillance system to monitor the safety of FDA-regulated medical products by accessing pre-existing electronic healthcare records. It proposes a Common Data Model (CDM) on top of a distributed architecture so that analytic applications can run on a uniform model. This model is maintained in a PDF document and collaborating EHR systems are expected to translate the EHR data to this common model.

There are other similar efforts to define common data elements and accompanying data models like GE/Intermountain Healthcare Clinical Element Models [25], National E-Health Transition Authority (NEHTA) Detailed Clinical Models [26] and I2B2 data model [27]. These are defined either as data dictionaries or through abstract data models which try to ensure interoperability within the boundaries of the associated initiatives. For instance, the query services, analysis methods or data exchange protocols envisioned by these initiatives can seamlessly run on top of the agreed common data element models. However, when it comes to achieving a broader range of interoperability, these efforts fall short: proliferation of common data element

models does not help to solve the interoperability problem. Exchange of EHRs for the care of patients or secondary use of EHRs is not directly possible across these initiatives. For example, it is not directly possible to query an EHR database which conforms to FHIM model through the query services provided by Query Health unless a mapping to Query Health CEDD is achieved first. When a researcher defines the data set to be collected for an observational study through CDISC SDTM variables, it does not become readily possible to extract these data sets from the EHRs which can provide medical summaries of eligible patients through HL7 CCD based patient summaries. The use of different sets of common data elements such as CDISC SDTM variables and HITSP Data Dictionary elements does not solve the problem of interoperability; yet it is not practical to expect all of these diverse initiatives and projects to stick to the same common model, and to use the same set of common data elements.

In this thesis, a federated metadata registry framework is presented where machine-processable definitions of the common data elements across domains can be shared, reused and semantically interlinked with each other to address this semantic interoperability challenge.

1.2 Objectives

In order to solve the interoperability problem within/between clinical research and care domains, several organizations are publishing common data element dictionaries and common information models as described above. Although these efforts ensure interoperability within the selected domain for the selected use cases, interoperability across application domain boundaries is not automatically possible. These stem from the following facts:

- It is an experienced fact that data requirements on clinical research side, and data availability and quality on EHR side are subject to change in time. As this happens, new initiatives propose new common data models into which collaborating EHR sources have to transform and transfer data,

regardless of the systems' central or distributed nature.

- Common data element model development efforts are most of the time carried out disparately. Although previous efforts are examined, predominantly, a common model is created from scratch.

- Most of the time, the specifications for these common data element sets and common models are in unstructured text files.

- Some of these efforts examine previous ones and reuse some common data elements proposed by the others, and sometimes provide partial mappings to other common data element dictionaries. For example, S&I CEDD reuses the elements from HITSP C154, NEHTA and FHIM; HITSP C32 provides extraction specifications between HITSP C154 data elements to the elements of HL7 CCD. However, these are maintained in several different spreadsheets or in PDF documents. Hence, it is not possible to process or query this data.

We believe there is a need for a more coordinated approach that would allow machine-processable definitions of the common data elements defined by different efforts to be searched, allow the common data elements to be reused and to be linked with each other and the mappings/links/relations between different data elements in different domains can be queried to address semantic interoperability [9]. In this thesis, we present a framework that facilitates all of these through the use of federated semantically enabled metadata registries conforming to the ISO/IEC 11179 standard [5] where common data elements maintained in different metadata registries can be uniquely identified, queried and linked with each other through Linked Data principles. We design and implement the Semantic Metadata Registry (Semantic MDR) as the backbone of this federated framework.

On top of the introduced interoperability model, we design and implement the Post Marketing Safety Study Tool (PMSST) which can extract any needed information from a patient record after it is retrieved as a result of an eligibility query or it is directly accessed from EHR database within a data

mining routine of the postmarketing surveillance studies. PMSST lets the clinical researcher to be able to define what need to be extracted from the patient records with the help of the common data elements accessed from a Semantic Metadata Registry. With this dynamic behavior, the researcher writes the surveillance methods on the schema/template which is created based on the data elements that he/she manipulates. With the help of the underlying interoperability framework, postmarketing surveillance methods do not have to be restricted to the data model of the EHR source.

1.3 Summary of the contributions

Having described the objectives in Section 1.2, the contributions of this thesis can be shortly highlighted with the following list:

- We introduce a new solution for the problem of data interoperability with a unified methodology; using the strong points of top-down and bottom-up approaches.
- We formalize and clearly separate the abstract definitions of the common data elements and their implementation dependent extraction specifications.
- We model and implement a federated framework of semantic metadata registries for managing disparate Common Data Elements within the Linked Data principles.
- A fully featured, ISO/IEC 11179 based Semantic Metadata Registry (Semantic MDR) has been developed and is being maintained as an open source project at https://github.com/srdc/semanticMDR
- Introduced framework has been fully implemented in eHealth domain for the data interoperability problem between clinical research and clinical care applications. With our solution, drug surveillance routines can access heterogeneous patient data automatically. That is, post market

drug surveillance studies can be developed independent of the underlying EHR systems.

- The Post Marketing Safety Study Tool (PMSST) enables the secondary use of electronic health records for clinical safety studies with the use of the introduced interoperability framework. Our tool has been built on a use case about Congestive Heart Failure in diabetic patients and is being used in real life settings on top of huge EHR databases in the context of the SALUS project. SALUS project is the provider of the patient data in this work.

- In order to show the applicability of the introduced framework to different domains, we have implemented a case study in the eBusiness domain for the interoperability of electronic business documents conforming to different electronic document standards. UN/CEFACT Core Components have been modeled as abstract common data elements and their extraction specifications have been defined to three different document standards.

This thesis is structured as follows: Chapter 2 gives brief information about the enabling technologies and background concepts of this study. Chapter 3 goes into the details of the Semantic Metadata Registry/Repository design and implementation, and describes the federated framework of the registries within the Linked Data cloud. Chapter 4 introduces the Post Marketing Safety Study Tool which utilizes the introduced interoperability architecture in order to enable the post market surveillance studies on existing EHR systems. Chapter 5 presents a case study that shows the applicability of the introduced interoperability framework in eBusiness domain. Chapter 6 outlines an analysis of the related work in terms of similar research activities. Finally, Chapter 7 concludes the thesis by giving final remarks and future work on this study.

CHAPTER 2

BACKGROUND

The federated architecture of the Semantic Metadata Registries is based on several concepts and enabling technologies from the research fields on metadata management, semantic web and Linked Data. This chapter briefly introduces the major concepts along with the objective of the data interoperability.

IEEE Standard Computer dictionary defines the interoperability as follows: "ability of two or more systems or components to exchange information and to use the information that has been exchanged". We can read this definition in two parts. The first one is the ability of exchanging information. This is defined as syntactic interoperability. In computer systems terminology, we can think of a process which listens on a port and can receive character streams. What comes after is the processing of the received character stream so that the exchanged information can be used by the parties. This corresponds to the second part of the above definition and defined as the semantic interoperability.

2.1 Metadata & Metadata Registry

Metadata has a common definition: "data about data". However, this is a very generic, and deprecated definition. Today's systems make a distinction between *structural* and *descriptive* metadata. Structural metadata gives information about the syntactic nature of the data (data about the containers of data) while descriptive one provides semantics for the data.

Metadata and metadata management is very important for the data

interoperability between different applications. To be able to exchange data and process data once it has been exchanged, the metadata should be agreed on by the interoperating systems. Figure 2.1 presents an example about the use of data and metadata within an application. In the figure, data about a person is presented through some fields like citizenship number, surname and gender. In this example, *Gender* is the metadata and *Male* is the data to indicate the value represented through the semantics of the metadata, *Gender*. During the data exchange between two different applications, when *Male* is received in one hand, it is crucial that the application should know that this data indicates the *Gender* of the person. Apart from that, the application also needs to know that this *Gender* data is indicating the gender of a *Person*. All this syntactic and semantic information is coded with the associated metadata. Hence, interoperating applications should agree on metadata before they start data exchange.

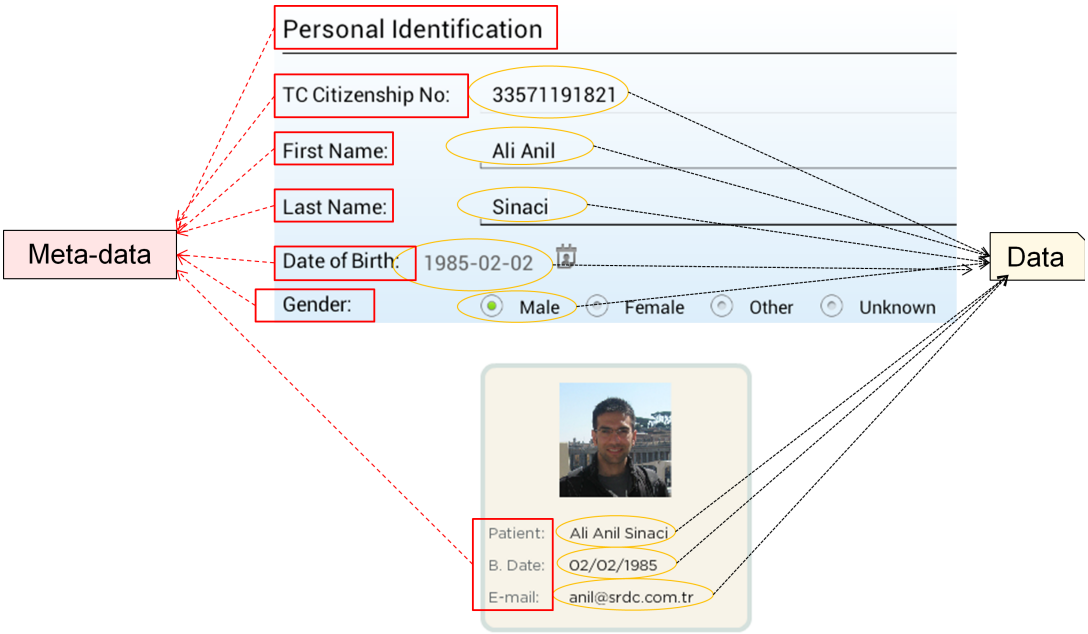


Figure 2.1: Importance of Metadata: Same data annotated through different metadata can lead inconsistencies

Most of the systems implement its own information model; hence each application has its own metadata. Even, it is highly probable that two different interfaces of the same application can consume data through different metadata. This situation is illustrated in Figure 2.1 where birth date

information of a person is annotated with “Date of Birth” in one interface and “B. Date” in the second interface. The human user can automatically associate that these two fields annotated with different metadata tags actually give the same data. Hence, humans can interpret the appropriate links between the inappropriately used metadata; however, in order to promote data interoperability between applications, this should be done by the applications themselves. That is, this information should be machine-processable.

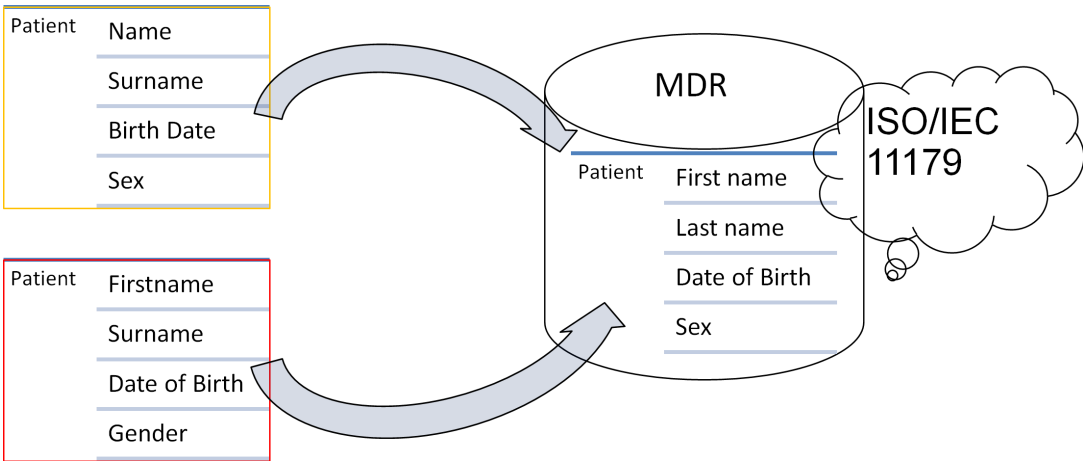


Figure 2.2: Metadata management in an ISO/IEC 11179 based Metadata Registry

In order to agree on the metadata, it should be available to the interoperating applications. Since metadata is data on the metadata level, it needs to be managed through well-established mechanisms. ISO/IEC 11179 defines the required mechanism to manage that data within the Metadata Registry/Repositories as briefly described in section 2.2. Figure 2.2 illustrates metadata management where the structural and descriptive metadata about the data itself (i.e. Patient information structured with First name, Last name, Date of Birth, Sex fields) is managed under an ISO/IEC 11179 based metadata registries. In this kind of a setting, the structure of each entity (i.e. Patient) such as the fields it contains is described and managed under the metadata registry together with the meaning of each information field such as the Sex field.

2.2 ISO/IEC 11179

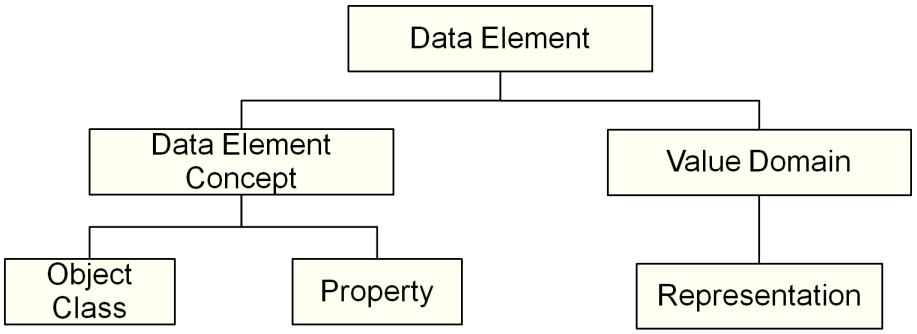
ISO/IEC 11179 family of specifications introduces a standard model for the metadata registries to increase the interoperability of applications with the use of data elements. The main idea is to make disparate systems use the same set of data elements with very well-defined methodologies so that different data models can be made through the aggregation and association of the same data elements. The standard defines a metadata registry; describes how to describe data, store data, classify data and manage data. That is, ISO/IEC 11179 comes in six different parts in order to address the semantics, representation and registration of the data elements (the metadata). These are listed as follows in the 2nd edition of the standard:

1. Framework: Contains an overview of the standard and describes the basic concepts
2. Classification: Describes how to manage a classification scheme in a metadata registry
3. Registry metamodel and basic attributes: Provides the basic conceptual model, including the basic attributes and relationships, for a metadata registry
4. Formulation of data definitions: Rules and guidelines for forming quality definitions for data elements and their components
5. Naming and identification principles: Describes how to form conventions for naming data elements and their components
6. Registration: Specifies the roles and requirements for the registration process in an ISO/IEC 11179 metadata registry

This standard addresses the management of the semantics of the data elements: it provides a standard metadata model for the representation of the data elements and provides a methodology for the registration of the descriptions of the data elements through this standard model to a metadata

registry. The aim is to facilitate the accurate common understanding of the data elements over time, space and applications.

ISO/IEC 11179 exhibits a relational data model which describes the metadata registries through entity-relationship diagrams. This metamodel is designed to be generic; hence any data element model can be represented regardless of the level of granularity. In Figure 2.3, decomposition of a data element is presented according to the metamodel of ISO/IEC 11179. The main advantage of this metamodeling is the clear separation of the concepts of the entity from its representation. That's why, it is possible to represent a data element in several different forms and formats while they all logically present the same data. This clear separation of the concept and representation of an entity lets appropriate links between the concepts and representations of different data elements while increasing the reuse. Figure 2.3 corresponds to a very small part of the metamodel exposed by the ISO/IEC 11179 standard. Apart from this decomposition; the metamodel includes the machinery to manage the administration and identification, different contexts, naming and definition, and classification.



$$\left(\begin{matrix}Object\\Class\end{matrix} + Property = \begin{matrix}Data\ Element\\Concept\end{matrix}\right) + \begin{matrix}Value\\Domain\end{matrix} = \begin{matrix}Data\\Element\end{matrix}$$

Figure 2.3: Decomposition of a data element according to ISO/IEC 11179

Applying ISO/IEC 11179 specifications throughout the metadata management provides several improvements in terms of data interoperability. The standard lists them as follows:

- Standard description of data

- Common understanding of data across organizational elements and between organizations
- Re-use and standardization of data over time, space, and applications
- Harmonization and standardization of data within an organization and across organizations
- Management of the components of data
- Re-use of the components of data

ISO/IEC 11179 is becoming a norm for the metadata registries, especially in eHealth [28]. The metamodel of the standard (2nd edition) is used in several projects [29]. Major ones can be listed as in the following:

- Metadata Online Registry (METeOR) by the Australian Institute of Health and Welfare [4].
- Data Dictionary by Canadian Institute for Health Information [30]
- Cancer Grid Metadata Registry by UK Cancer Grid [31]
- Cancer Data Standards Repository (caDSR) by US National Cancer Institute [32, 33]
- Environmental Data Registry by US Environmental Protection Agency [34]
- US Health Information Knowledgebase (USHIK) by the Agency for Healthcare Research and Quality [35]
- US National Information Exchange Model (NIEM) by US Department of Homeland Security (DHS) and US Department of Justice (DOJ) [3]
- Global Justice XML Data Model (GJXDM) by US Department of Justice [36]

2.3 SALUS Project

Despite pre-marketing clinical trials, Adverse Drug Events (ADEs) impose a remarkable burden on the health care systems: in Europe, they are estimated to be responsible for 6.5% of hospital admissions, complicate at least 1 in 7 in-patient episodes, and account for considerable morbidity, mortality, and extra costs. An impact assessment carried out for the EU Commission has estimated that ADEs cause 197,000 deaths per year in the EU, at a total cost of €79 billion [37]. As a consequence, post-marketing surveillance of drugs and prevention of ADEs still remains a major public health issue.

SALUS (Scalable, Standard based Interoperability Framework for Sustainable Proactive Post Market Safety Studies) is an R&D project co-financed by the European Commission's 7th Framework Programme (FP7) [2]. The SALUS project is exploring new ways of accessing and analyzing data found in electronic health records to provide an infrastructure that enables the execution of the safety studies for mining and analyzing real-time patient data. In this way, patient safety can be ensured through early detection of rare adverse events; the pharmaceutical industry can provide faster medication innovation by decreasing time to market for new, safe and effective drugs, and at the same time the load of the overwhelmed medical practitioners can be reduced.

SALUS is developing the functional and semantic interoperability architecture in order to connect heterogeneous EHR data sources through a semantic layer. SALUS has analyzed the available content models and published a set of abstract SALUS Common Data Elements. On top of it, a SALUS Common Information Model has been created. This is a semantic, RDF based information model and the objective is to mediate the data exchange through this Common Information Model. Since it is a semantic model, it is possible to perform semantic reasoning and infer implicit facts to be used during data analysis. However; although it is a semantic model, it is yet another content model for the sake of data interoperability. Hence, applications can interoperate only if they can process data in the form of SALUS Common Information Model.

2.4 IHE Data Element Exchange Profile

Integrating patient care and clinical research domains requires a standard-based expressive and scalable semantic interoperability framework, allowing dynamic mappings between data elements and semantics of varying data sources. This can be achieved through a metadata registry architecture where machine processable definitions of data elements across domains can be shared, re-used, and semantically interlinked with each other to address this interoperability challenge to move towards EHR-enabled research. DEX enables retrieving "extraction specifications" for a data element defined in a selected domain (like SDTM [19] data elements), from an implementation dependent content model in another domain (like HL7 CCD [14]).

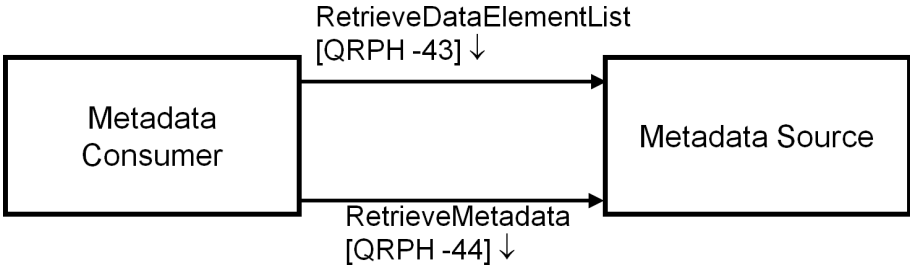


Figure 2.4: Actors and transactions of the IHE DEX profile.

This thesis contributes to the development of the IHE DEX profile under the Quality, Research and Publich Health Domain of IHE [38]. Semantic MDR, developed in this study, is one of the first implementations of the DEX profile and plays the Metadata Source role [29]. Figure 2.4 shows the actors directly involved in the DEX Profile and the relevant transactions between them. The profile is XML based; designed with SOAP [39] web services and exposes an API depicted with a WSDL [40] definition.

2.5 Semantic Web & Linked Data

Sir Tim Berners-Lee - its creator - defines the Semantic Web as "a web of data that can be processed directly and indirectly by machines" [41]. It can be considered an extension to the available Web system with better methodologies

in order to express the meanings of the things by representing knowledge in standardized ways, i.e. by defining ontologies [42]. In the Semantic Web context, an ontology is a schema for the data in a domain; in other words, is the explicit formal specification of the terms and relations among them in a specific domain.

Linked Data is defined as "a term used to describe a recommended best practice for exposing, sharing, and connecting pieces of data, information, and knowledge on the Semantic Web using URIs and RDF" [43]. Linked Data is what machines can make most out of the Semantic Web technologies; large scale integration of and reasoning on data on the Web. The collection of Semantic Web technologies such as RDF [44], OWL [45], SKOS [6], SPARQL [46] etc. provides an environment where applications can query that data, hop over the links and draw inferences.

There are several different open data sets available on within the, so-called Linked Open Data cloud. Examples include Wikipedia, Wikibooks, Geonames, WordNet, the DBLP bibliography and many more that are published under permissive licenses. The goal of the W3C Linking Open Data community project is to extend the Web with a data commons by publishing various open data sets as RDF on the Web and by establishing appropriate links between data items from different data sources [47]. These links enable the Semantic Web applications to navigate from a data item within one data source to related data items within other sources.

As of September 2011, in the Linking Open Data initiative, there are 295 data sets consisting of over 31 billion RDF triples, which are interlinked by around 504 million RDF links, as displayed in the cloud diagram in Figure 2.5.

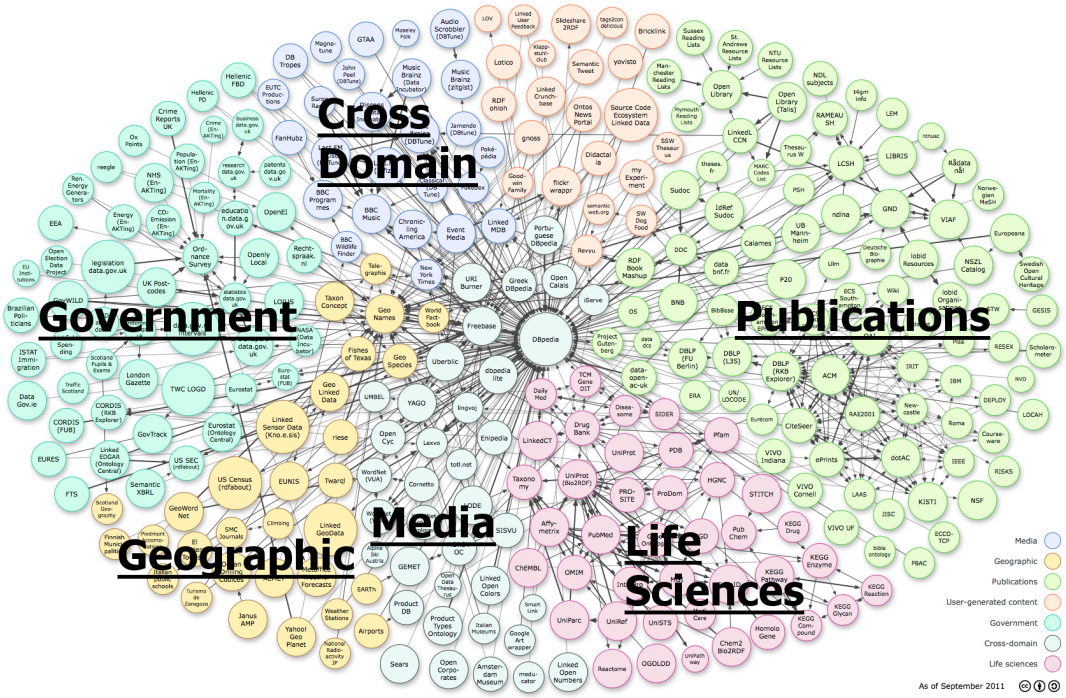


Figure 2.5: Linking Open Data cloud diagram as of September 2011

CHAPTER 3

SEMANTIC METADATA REGISTRY/REPOSITORY

The first challenge we would like to address is to maintain the definitions of common data elements (CDE) in a machine processable manner rather than keeping them in PDF documents or spread sheets so that it becomes possible to search, query and link to them. For this we have selected to adopt the ISO/IEC 11179 - Metadata Registries standard since it is becoming a standard for metadata management in eHealth.

There are numerous adoptions of ISO/IEC 11179 registries [33, 31, 35, 48, 49, 50] as also listed in 2.2 to address semantic interoperability, several of which are in healthcare domain. These central metadata registries are used to maintain a set of common data elements in the selected domain so that data sources and data requesters can agree on unambiguous semantics of the selected data elements in the chosen domain. To address the data interoperability at a larger scale, it should be possible to link and reuse the CDE definitions across application domains which can be greatly enabled by a semantically interlinked federated metadata registry (MDR) framework. Centralized metadata registries would not scale as it is not practical to manage the CDEs within different application domains in a single registry; each set of data elements can evolve in time, there should be a more flexible mechanism to manage and exploit the linked set of CDEs across domains.

3.1 Proposed extensions to ISO 11179 Standard to achieve federated metadata management for semantic interoperability

A federated MDR framework should enable the following basic functionalities:

- Searching the CDEs maintained by different MDRs
- Retrieving the standard description of a selected CDE from an MDR
- Reusing the CDEs maintained in a different MDR by referencing to the respective CDE

In order to facilitate the data interoperability more effectively across domains, a semantically linked federated MDR framework should support some additional functionality:

- It should be possible to link and semantically associate the CDEs across different MDRs in reference to well-accepted knowledge organization system (KOS) ontologies and terminology systems.
- It should be possible to easily query these semantic relationships within and across MDRs. We have chosen to apply Linked Open Data (LOD) principles as the basis of this semantically linked federated MDR framework. Linked Data is a recommended best practice for exposing, sharing and connecting pieces of data, information and knowledge on the Semantic Web using Uniform Resource Identifiers (URIs) and RDF. It provides a natural way to expose the CDEs maintained in different MDRs openly in the LOD cloud and interrelate them with each other as depicted in Figure 3.1.

3.1.1 MDRs in the LOD Cloud

In order to integrate the MDRs within the LOD cloud, the following principles are adopted:

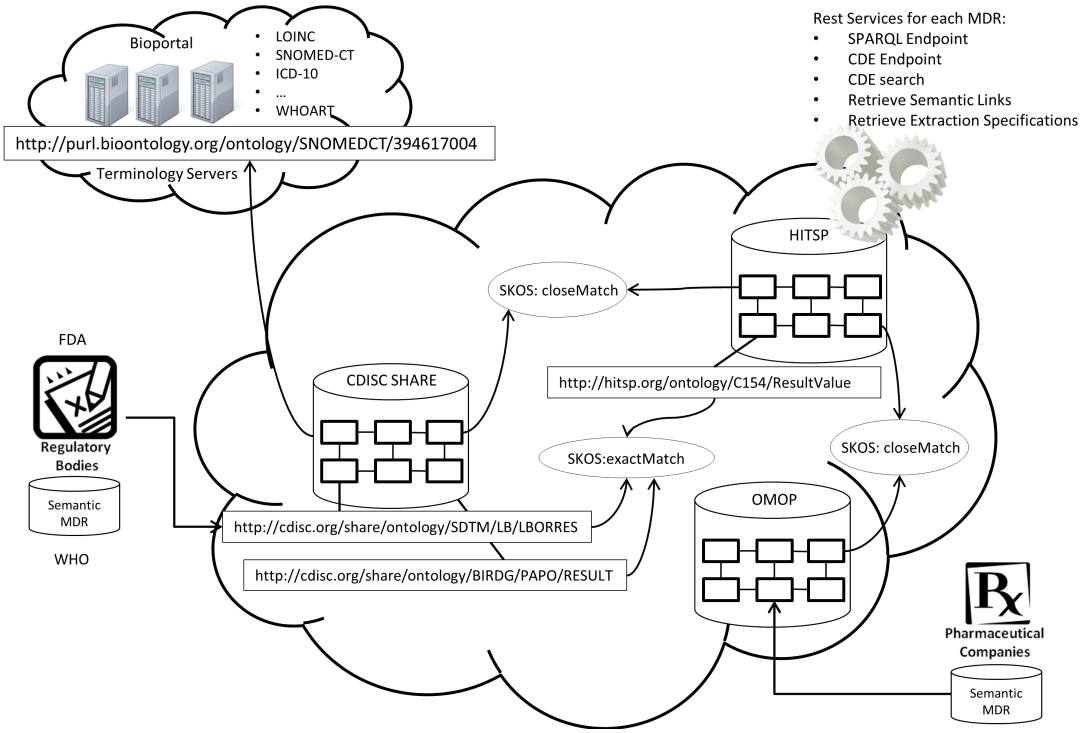


Figure 3.1: Federated semantic MDR framework. Within the LOD cloud, each MDR maintains a set of CDEs together with the corresponding components and relations. The CDEs are linked to CDEs of other MDRs through KOSs and annotated with terminology systems.

- Each CDE is uniquely identified by a URI
- Each CDE is dereferenceable, that is, MDRs provide the necessary HTTP-REST services for looking up the CDEs by using their unique URIs
- Each MDR provides semantic RDF descriptions of the CDEs, which are accessible through the provided HTTP services. When a CDE is looked up through its URI, the RDF description of the CDE is returned where all context of the CDE is presented in RDF: each RDF property is interpreted as a hyperlink to the other (possibly further linked) registry resources. This automatically opens up access to more data which is usually referred to as the "follow-your-nose principle". To enable this, we have created an OWL ontology from ISO/IEC 11179 metamodel. We designed the ontology with OWL-Lite which is the lightest sublanguage of OWL with highest simplicity and lowest complexity. When a CDE is looked up, its RDF description in conformance to the ISO/IEC 11179 metamodel is returned

which includes links to other related MDR resources like the *object class*, *property*, *value domain*, *enumerated value list*s, *context* and *classification scheme item*s that this CDE is related to. It should be noted that each of these resources are also maintained as uniquely identifiable LOD resources; hence, not only the CDEs but all objects within the ISO/IEC metamodel are readily available through the LOD principles: i.e. openly accessible with unique URIs with semantic descriptions attached.

3.1.2 Linking CDEs to Terminology Systems

In the Semantic MDR, it is possible to annotate the CDEs with external terminology systems. Inline with the LOD approach, in our implementation, links of the CDEs to the terminology system codes are also referred through their unique URIs in the LOD cloud. BioPortal [51] already provides a wide range of terminology resources through the LOD principles where each terminology code is uniquely identified with a URI. In the Semantic MDR, for each terminology system a *Classification Scheme (CS)* resource is created as shown in Figure 3.2. When an MDR resource is going to be related with a code from a terminology system, a *Classification Scheme Item (CSI)* resource is created under this CS resource and linked with the MDR resource. The unique URI of the terminology system code is recorded in the *value* property of CSI resource. In this way, all the CDEs across different MDRs annotated with the same terminology system code will be linked with the unique resource description created for the terminology system code, which directly provides a means to search and link the CDEs across domains through the LOD principles.

3.1.3 Linking CDEs to other CDEs

In our approach, it is possible to set other semantic links between the CDEs maintained in different MDRs as a part of semantic description of the CDE. For recording the semantic relationships between the CDEs across MDRs, the *Classification Scheme (CS)* constructs available in ISO/IEC 11179 model are

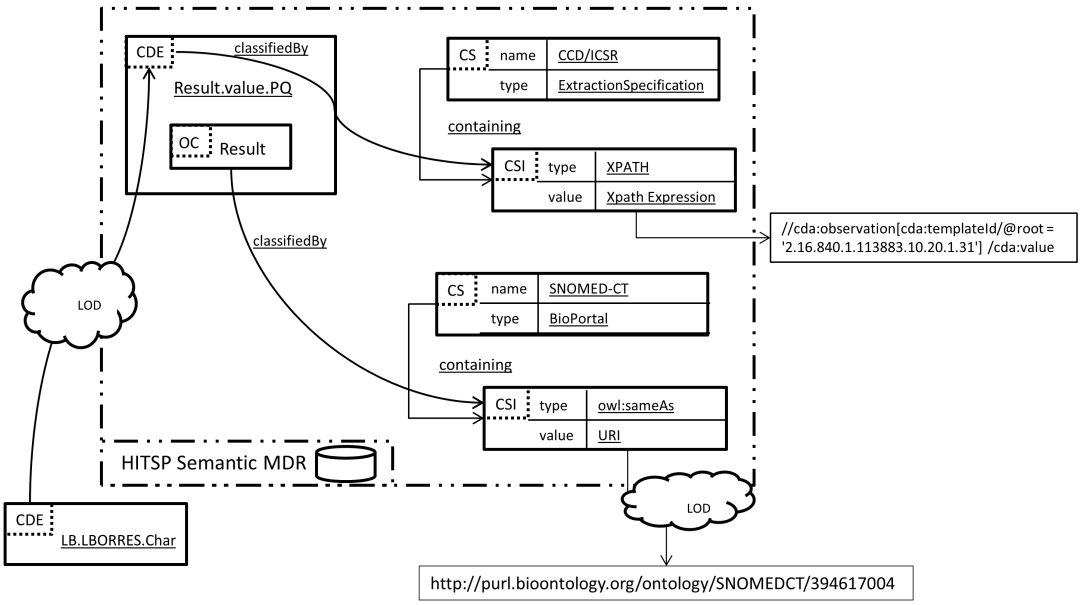


Figure 3.2: Annotations and links of a CDE and its Object Class (OC) inside a Semantic MDR. The OC is annotated with a concept (term) from SNOMED-CT which is maintained under BioPortal through owl:sameAs property. The CDE has an "Extraction Specification" which is an XPath expression pointing the exact place of the CDE in the HL7 CCD models. These annotations and links are modeled through Classification Scheme and Classification Scheme Item elements of the ISO/IEC 11179 metamodel.

utilized. For each external MDR, a CS resource is created as presented in Figure 3.3. Whenever a semantic relationship is to be created between CDEs, CSI resources are created and linked with the source CDE where the *type* property is set as the URI of the semantic relationship and *value* attribute is set as the unique URI of the target CDE. For identifying the semantic relationships, we are using upper KOS ontologies like SKOS [6]. In particular, SKOS **closeMatch** and **exactMatch** properties are exploited. By using such already existing semantic resource sets like SKOS, we ensure that the CDEs are properly interlinked with each other via the other well-known LOD resources.

3.1.4 Linking CDEs to Extraction Specifications

One of the additional functionalities we would like to enable through a federated MDR framework is retrieving "Extraction Specifications" for a CDE defined

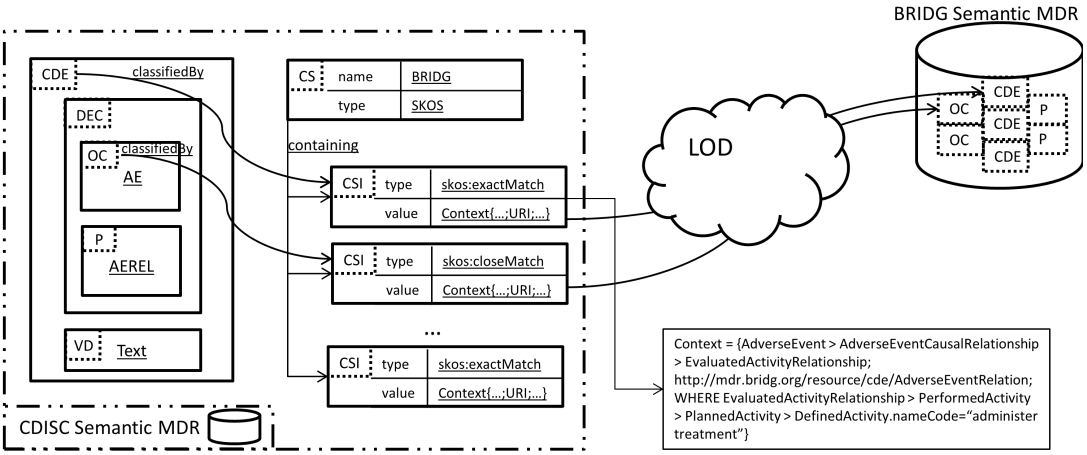


Figure 3.3: The components of a CDE together with their classifications for the LOD links to other CDEs and components residing in a different Semantic MDR. The CDE - AE.AEREL.Text - has a skos:exactMatch relation with the CDE - AdverseEventRelation - in the Semantic MDR which holds the BRIDG CDEs. If needed, CDE mappings to other CDEs can be given through a Context in which some pre-conditions and rules can be specified.

in a selected domain, from a content model in a different domain. CDEs are often abstract data element definitions, which are later used to annotate the actual data elements in implementation dependent models which carry clinical content. These implementation dependent models are called content models or information models. For example, HITSP C154 data elements are used to annotate parts of the CCD content models to indicate the unambiguous meaning of the CCD elements. Maintaining the links between these abstract CDEs with the implementation dependent content models through an MDR architecture would facilitate retrieving machine processable extraction specifications that can be used to enable dynamic interoperability across different domains. Through our federated MDR framework, it becomes automatically possible to extract the SDTM annotated data sets from a medical summary conforming to HITPS C32 content model specifications (annotated with C154 data elements).

In our approach, each MDR can maintain several different content models as the implementations of the abstract CDEs served over the MDR. For the supported content models, MDR can maintain extraction specifications for each CDE available in the registry as a machine-processable specification of accessing the corresponding part of the content model. Content models can be

XML documents, database schemes or RDF instances. Based on the type of the content model, different types of extraction specifications can be supported. An extraction specification is any script which can be executed on its associated content model. Current implementation of the Semantic MDR supports three types of extraction specifications:

1. **XPath** [10]: If the content model specification is based on XML Schema [11] and the data is serialized in XML, then it can be queried through XPath scripts. As shown in the example of Figure 3.2, the information pointed by a CDE can be extracted from HL7 CCD based patient summaries when there are XPath scripts in the extraction specifications of the CDEs.

2. **SPARQL** [46]: If the content model specification is based on RDF and the data is residing in RDF graphs, then SPARQL scripts can be executed on the graph to retrieve the pointed information. An example is shown in Table 3.1 which is a part of the *ExtractionSpecification* for the CDE – "Patient.Allergy.Severity".

3. **SQL**: If the content model is a relational model and data is residing in legacy relational databases, then SQL scripts can be executed to retrieve the associated information with the CDEs.

Table3.1: SPARQL script to retrieve the severity information for the Allergy on a Patient. The target content model is SALUS Common Information Model [1, 2] which can serve data through RDF graphs

```
SELECT ?severity
    WHERE {
        ?pt a salus:Patient.
        ?pt salus:allergy ?allergy.
        ?allergy salus:severity ?severity.
    }
```

In the Semantic MDR, a *Classification Scheme (CS)* resource is created for each content model. The *type* of this CS is set as **ExtractionSpecification**. For each extraction specification linked to the CDE resources, a *Classification Scheme Item (CSI)* resource is created where *type* property is set from the value set

{XPath, SQL, SPARQL} and *value* property contains the extraction expression as presented in Figure 3.2.

3.1.5 Federation through Linked Data Principles

Within the generic metamodel as shown in Figure 3.2 and Figure 3.3, all external relations are indicated through *Classification Scheme Item*s (CSI) which are grouped under the *Classification Scheme*s. Therefore modeling a link to another CDE is similar to a link to a term (concept) in an external terminology system (i.e. SNOMED-CT) or to an extraction specification pointing to an implementation dependent model. That is, from the perspective of the Semantic MDR, these external resources are all metadata, but they are expected to follow the Linked Data principles and adopt well-known semantic schemes (like the SKOS), ontologies (like the ISO/IEC 11179 ontology) or the standardized serializations (like the IHE DEX profile). The beauty behind the federated semantic MDR framework is that, it does not enforce the compliance to all of the mentioned specifications. It can use and deduce as much knowledge as it can acquire from the linked resources. For example, in the current implementation we make use of the REST endpoints of BioPortal for terminology annotations. BioPortal provides the RDF serializations of the terms through well-known knowledge organization systems such as SKOS; as a result, we can automatically process a number of attributes such as labels and unique identifiers. Hence, if two different CDEs from two different Semantic MDRs are classified by the same term coming from BioPortal, a search through the federated architecture with the identifier of or a keyword belonging to that term would successfully find the two CDEs. These links to the terminology systems can also be used for searching the CDEs from the federated MDR framework, as a next step the extraction specifications of the discovered CDEs from the selected content models can be retrieved.

Apart from the best practices, it is a known fact that most of the existing EHR systems do not use standard terminologies or groupers. Instead, they use their local, proprietary coding schemes and vocabularies for data annotation.

Making this existing legacy EHR data available for clinical research is a challenging task and there needs to be some additional effort in order to succeed the data interoperability with the existing systems. Our framework minimizes this effort because direct manipulation of the legacy data is not required. One needs to introduce the CDEs of the local coding system or content models used by the legacy systems to a local Semantic MDR and establish the appropriate mappings to the other standard based CDEs in the federated MDR framework. For example, in Figure 3.2, it can be assumed that if a local CDE is used to annotate the lab results in an EHR system and that these local CDE is linked with the HITSP C154 CDE – "Result.value.PQ", then from the mappings of "Result.value.PQ" different extraction specifications can be reached. In this example the XPath expression pointing to the exact location of the CDE in HL7/ASTM CCD model can be retrieved from the federated semantic MDR framework.

3.2 Design & Implementation of the Semantic MDR

The Semantic MDR provides the capabilities of a metadata registry and a metadata repository at the same time. While we utilize several services for the federated architecture of the semantic metadata registries, we also implement web based, easy-to-use graphical user interfaces for the management of the CDEs including browsing, searching, editing and automatic importing in order to meet the requirements of a metadata repository.

3.2.1 Ontology of the ISO/IEC 11179 Metamodel

ISO/IEC 11179 provides a relational model for the structure of the MDRs through its entity-relationship diagrams as introduced in section 2.2. To be able to add semantic capabilities such as handling inter-links between CDEs and handling external links to other repositories, terminology systems and classification schemes etc., the Semantic MDR has been built on top of a triple store which bases the knowledge on the ontological representation of the

ISO/IEC 11179 metamodel. While building the ISO/IEC 11179 ontology, we adopt OWL [45] such that an OWL resource for each metamodel construct is created according to the mappings given in Table 3.2. In addition to the direct mappings of the ISO/IEC 11179 constructs, all relationships (i.e. class hierarchies, class-to-class relations) have been reflected to the ontology in order to be fully compliant with the metamodel. Full version of the ontology can be found in [52].

Table3.2: Mapping of ISO/IEC 11179 metamodel constructs to OWL constructs

ISO/IEC 11179 metamodel construct	**OWL construct**
class	owl:Class
attribute	owl:DatatypeProperty
composite attribute	owl:ObjectProperty
class relationship	owl:ObjectProperty

3.2.2 MDR Knowledge Base

The Semantic MDR opens up several services in various layers which take root from its MDR Knowledge Base as shown in Figure 3.4. The goal is to enable the federated communication through Linked Data principles. RESTful part of these services and its use in succeeding the interoperability between the clinical research and patient care domains are introduced in the following sections.

Since powerful semantic capabilities following Linked Data approach require more sophisticated data management than the relational model, we base the data persistence on top of a Triple Store component as presented in Figure 3.4. Apache Jena [53] has been adopted as the RDF framework which also has native support for OWL ontologies. Apache Jena has a built-in triple store backend, Jena TDB [54]. Our Triple Store components can selectively use either Jena TDB or Virtuoso [55] which is another high performance triple store implementation. They provide native SPARQL support, and have pros and cons over each other according to the usage context [56]. That's why the

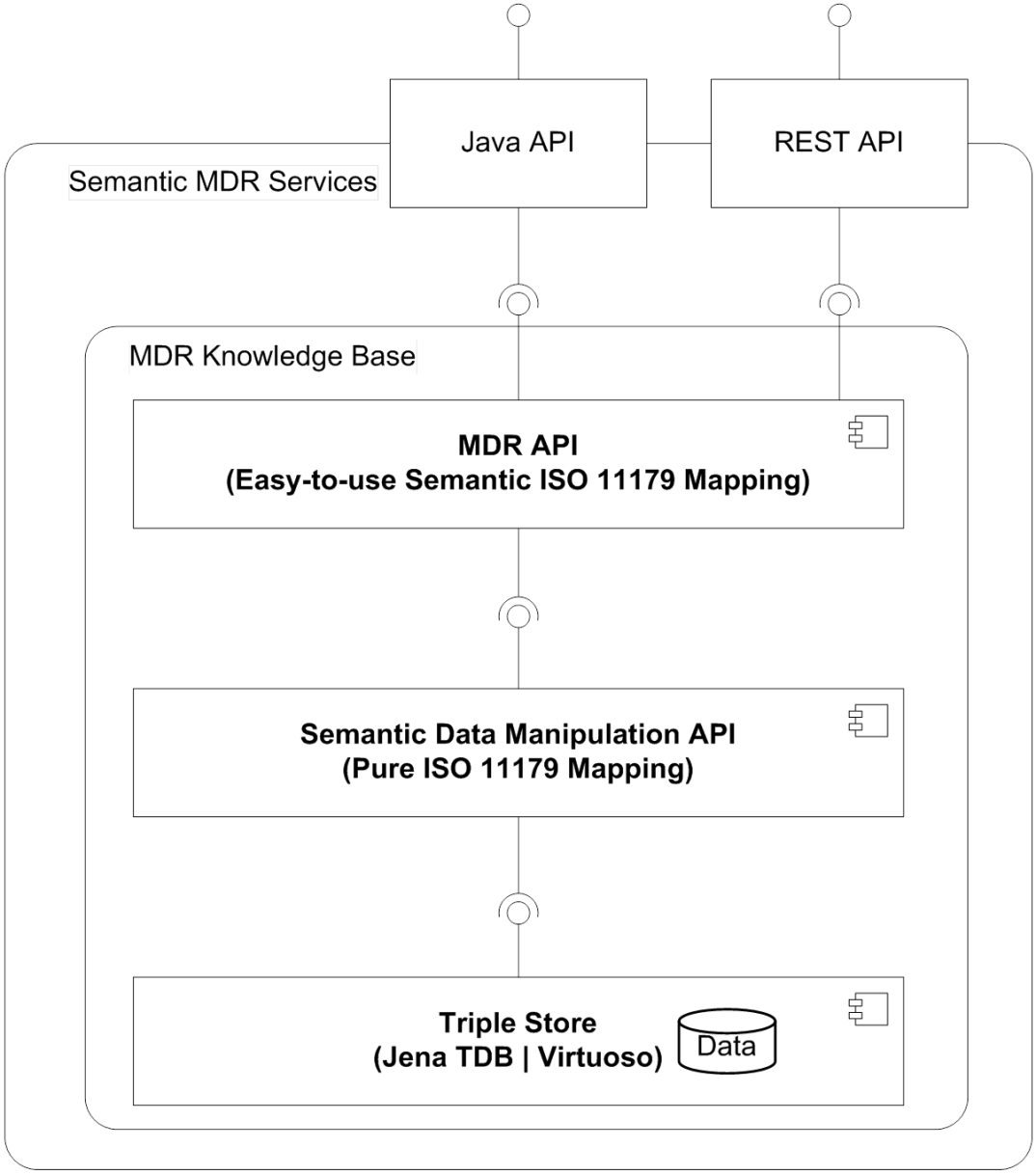


Figure 3.4: High-level view of the architecture of the Semantic MDR Service Layer. At the bottom, there is a triple store serving as a backend for the MDR Knowledge Base. Above the triple store, there is a 3 layered API to perform semantic operations on this Triple Store. Semantic Data Manipulation API is a direct implementation of the ISO/IEC 11179 metamodel which reflects the operations to the underlying RDF graph. MDR API an abstraction layer which hides the complex details of the ISO/IEC 11179 metamodel and provides easy-to-use methods for the data manipulation.

Semantic MDR provides a driver component which can automatically be integrated with both of the triple store implementations.

The Semantic MDR develops different types of importers in order to automatically populate the knowledge base with the CDEs of the widely used content models. The current implementation can import OMOP CDM v4.0, SDTM v1.3, CDASH v1.1, HITSP C154 v1.0 and part of the BRIDG model using different serialization formats including SQL, RDF, XML Schema and comma-separated values. Importers have domain dependent implementation since each content model comes with its own specification. Even if two content models are represented through XML schema definition; since the structure of the content is different, it would not be feasible to implement an XSD Importer to parse all the content models defined through the XML schema. Content model importers make use of the MDR API exposed on top of the MDR Knowledge Base. Since the MDR API enforces a pattern to create and manage the CDEs in the repository, importers follows this pattern while populating the knowledge base. First, a *Context* to represent the content model is created. Afterwards, other resources necessary to create the *DataElements* are created on the context such as *ObjectClass, Property, ValueDomain*, etc. As the last step, *DataElements* in the content model is imported into the Context. Importers exhibit a good example of the MDR API usage.

3.2.3 RESTful interface

Once all of the links to external terminology systems, CDEs in external MDRs and links to the content models become a part of the semantic description of a CDE, Semantic MDRs can open some simple REST services to ease the semantic query of the CDEs across MDRs. A full list of the proposed REST services is presented below. Through these services, it becomes possible to perform federated queries on the MDRs to retrieve semantic descriptions of the CDEs and process these for achieving semantic interoperability across domains.

- **SPARQL endpoint:** Native SPARQL support. Functionalities of all other REST services can be provided by the SPARQL endpoint. RDF and SPARQL aware systems can build many semantic applications by consuming the SPARQL endpoints of the Semantic MDRs.

- **CDE endpoint**: Given the URI, retrieve the full RDF description of the CDE from an MDR.

- **CDE search**: Parameterized search for the CDEs through the allowed properties defined in ISO/IEC 11179 meta-model. For example, query CDEs by *Object Class* and/or *Classification Scheme Item*. In this way it becomes possible to search CDEs annotated with a specific terminology system code.

- **Semantic links**: Retrieve all semantic links of the CDEs to the CDEs in other MDRs. Given the URI of a CDE (source), MDR returns the URIs of the other CDEs (target) interlinked with the source CDE, together with the URIs of the semantic relationships between these CDEs (e.g. skos:exactMatch). The requester can then directly lookup the full semantic description of the target CDEs, as unique URIs of the CDEs will already direct the user/application to the MDR where it is maintained in.

- **Extraction specification**: Retrieve "extraction specifications" for a CDE in a selected domain for a supported content model. Input is the URI of the CDE and URI of the content model. Note that the HL7 CCD content models provided by HITSP or IHE Patient Care Coordination Domain are already uniquely identifiable through Object Identifiers (OIDs).

Semantic MDR implementation is maintained as an open source project under GitHub [57] and referred by the associated work group of ISO as one of the vendor implementations of the ISO/IEC 11179 standard.

3.3 Exploiting linked metadata registries for semantic interoperability

In our scenario, which reflects one of the pilot application scenarios from SALUS project and implemented in Chapter 4, a study data manager in a pharmaceutical company aims to design the data collection set for a new trial. The objective is to prepare a properly annotated study design document so

that it can be automatically populated with patient data coming from HL7 CCD based content models through the information retrieval process of the Federated Query Service. The flow of the scenario is depicted in Figure 3.5 and the steps are described in the following:

1. The study manager searches the local MDR of her organization to retrieve the data elements together with their descriptions for the selected set of variables in the data collection set. The local MDR returns a list of data element descriptions, including the unique URIs of the matching SDTM CDEs maintained by the MDR managed by CDISC.

2. The study manager prepares the study protocol as a CDISC ODM document annotated with the SDTM CDEs and sends it to the Contract Research Organization (CRO).

3. The Electronic Data Capture (EDC) system of the CRO automatically processes the study protocol and tries to map the data items identified in the data collection set to the parts of HL7 CCD medical summary documents of the study patients it collects from the participating care organizations.

4. EDC queries the federated MDR framework for the extraction specifications of the SDTM CDEs from HL7 CCD format. If the CRO is using a Semantic MDR, then the federated search system is already embedded into the MDR. Otherwise, the federated query service end-point is invoked by the CRO's EDC. The service asks for the extraction specifications of each SDTM CDE to the registered MDRs through the RESTful interfaces.

5. None of the MDRs directly provide the extraction specification of the selected SDTM CDE (say LBORRES which stands for "results of a lab test") from HL7 CCD format. The query service asks for the Semantic Links of LBORRES to the registered MDRs. In our example scenario - a Semantic MDR maintaining BRIDG model data elements - provides a mapping between the LBORRES CDE in CDISC SDTM domain to the

"PerformedObservationResult.value.Any" CDE in BRIDG domain. It also maintains a mapping between "PerformedObservationResult.value.Any" CDE and the "Result.Value" CDE from HITSP domain. Hence, when the federated query service asks for Semantic Links of LBORRES, BRIDG MDR returns two URIs of

   a. "PerformedObservationResult.value.Any" from BRIDG

   b. "Result.Value" from HITSP

6. "Result.Value" CDE is served in a Semantic MDR hosted by HITSP which is linked with "PerformedObservationResult.value.Any" CDE through **skos:exactMatch** semantic relationship. The federated MDR search system now looks up to the HITSP MDR to retrieve the extraction specification of "Result.Value" CDE in RDF format and the extraction specification to the HL7 CCD content model is available as "cda:observation[cda:templateId/@root='2.16.840.1.113883.10.20.1.31']/-cda:value" as an XPath query.

7. In this way, the EDC is able to retrieve the required data elements in the data collection set from the HL7 CCD documents provided for each study visit by the participating organizations.

A similar flow can be achieved through retrieving the RDF descriptions of the CDEs by calling the CDE endpoints and by processing these RDF descriptions where semantic links and links to extraction specifications are already available.

As depicted in the example scenario, through the proposed federated MDR framework, it is possible to facilitate data interoperability across clinical research and care domains although different standards and different CDEs are in use. Similar to this scenario, another use case can be automatic population of the case safety reports to notify adverse drug events through Individual Case Safety Report documents [2].

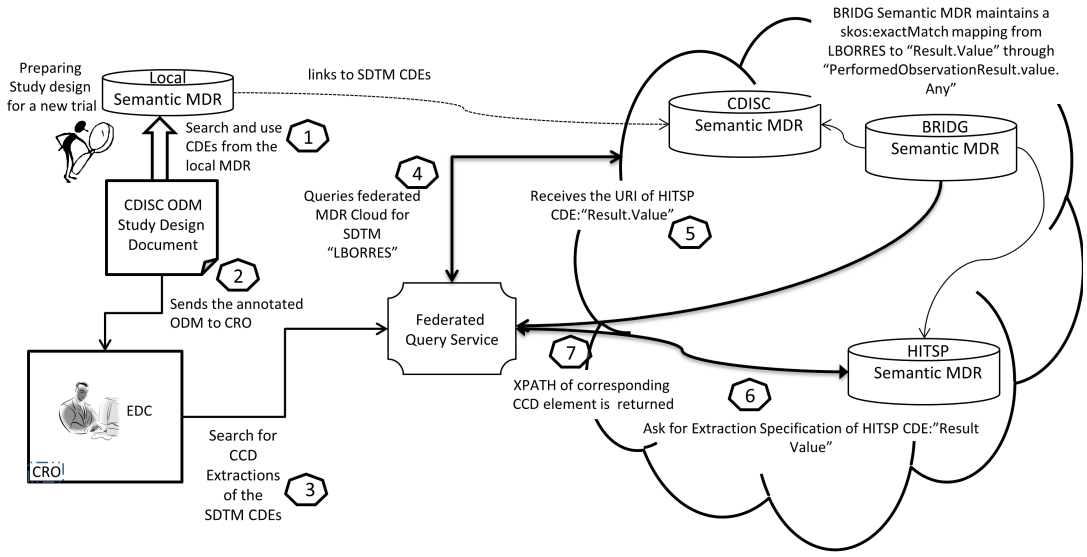


Figure 3.5: Step-by-step representation of the scenario in which the federated semantic MDR framework is used for the interoperability of clinical research and clinical care domains. A properly annotated study design document can be automatically populated through the information retrieval process of HL7 CCD based content models with the help of the Federated Query Service. The service makes use of the simple REST interfaces of the Semantic MDRs

CHAPTER 4

POST MARKETING SAFETY STUDY TOOL

It is a well-accepted fact that due to the limited size and duration of the clinical trials, drugs may still have serious side effects - Adverse Drug Reactions (ADRs) – after they are marketed. Postmarketing drug surveillance systems have been in place in order to analyze additional information about a drug's safety, efficacy and optimal use to capture such ADRs. Pharmacovigilance is the science of detection, assessment, understanding and prevention of ADRs [58]; and postmarketing surveillance is one of the fundamental activities within pharmacovigilance. During the last decades, postmarketing activities in pharmacovigilance have largely based on spontaneous case reports and still the majority of the activities depend on spontaneous reports. However, there are certain limitations on surveillance activities with spontaneous report data [59, 60, 61, 62].

Pharmcoepidemiology is another field about drug safety which studies the use and effects of drugs in large populations to bridge the gap between clinical trials phase and post market information of drugs. As in the case of pharmacovigilance, postmarketing surveillance is of vital importance for pharmacoepidemiology; especially for evidence development about effectiveness, safety and quality of drugs in terms of ADRs [63, 64].

4.1 Secondary Use of Electronic Health Records for Postmarketing Surveillance

At present, postmarketing drug surveillance is largely being carried out with traditional methods both for pharmacovigilance and pharmacoepidemiology. In pharmacovigilance, there is active research on data mining algorithms [65] on spontaneous report databases. On the other side, dedicated cohort and case-control studies are being performed within pharmacoepidemiological research. Although these traditional methods are currently dominant, a new research area is emerging which uses the already available electronic health data for clinical research purposes which is referred as the secondary use of Electronic Health Records (EHRs). EHRs provide a huge, but still under-utilized source of information on the real world use of drugs for observational studies. Although EHRs are primarily designed for patient care, they also contain a broad range of clinical information highly relevant for surveillance studies. EHR data available in clinical care systems can clearly complement and strengthen existing postmarketing safety studies [61, 2, 66]. Relative to spontaneous reports, EHRs cover extended parts of the underlying medical histories, include more complete information on potential risk factors and are not restricted to patients who have experienced an adverse drug reaction.

Successful utilization of available EHRs for clinical research in terms of access, management and analysis of patient data within and across different functional domains is a critical factor in terms of secondary reuse [66]. In line with this vision, there are important efforts for building large data pools from the EHRs to benefit from the available longitudinal observational data. The Sentinel Initiative from the U.S. Food and Drug Administration (FDA) aims to build a distributed network for active postmarketing surveillance for drug safety in the U.S [67, 68]. The Observational Medical Outcomes Partnership (OMOP) is another important initiative targeting a similar objective for improvements in post market drug monitoring [69]. There are several other pharmacoepidemiological databases such as the Clinical Practice Research

Datalink (CPRD) [70] which is based on the General Practice Research Database (GPRD) experience in the UK and The Health Improvement Network (THIN) database containing longitudinal medical data [71]. As a natural result, data mining on such national and international data pools appears as a new research area for signal detection and safety monitoring [61, 62].

The objective of the aforementioned initiatives is to use the available EHR data held by multiple different systems for clinical research purposes (mainly for postmarketing surveillance, comparative effectiveness research and evidence development). While some of them offer to hold central databases where participating EHR sources should transform and transfer EHR data to the central database, some prefer to keep distributed networks where EHR data resides in owner systems but data is transformed to an agreed information model and kept in a database conforming to that model. Although the debate is not yet fully resolved for some researchers, distributed systems expose clear advantages in terms of scalability and privacy [66, 72]. In addition to the distributed architecture of the Sentinel Initiative, recent research projects like SALUS (Scalable, Standard based Interoperability Framework for Sustainable Proactive Post Market Safety Studies) [2], TRANSFoRm (Translational Research and Patient Safety in Europe) [73] and EHR4CR (Electronic Health Records for Clinical Research) [74] address the different levels of the interoperability problem between the clinical research and patient care domains with a distributed perspective.

4.2 Objective of the Post Marketing Safety Study Tool (PMSST)

In this thesis, we address the heterogeneity problem among common data models for clinical researchers who work on EHR data for postmarketing surveillance studies. We demonstrate that this problem of interoperability can be solved in an upper level with the use of Common Data Element (CDE) phenomenon. If the applications share the machine processable definitions of the data elements and there are established links between data elements of

different domains (i.e. clinical research and patient care domains), this can be used to facilitate automatic access to data across different domains. Hence, in the context of postmarketing surveillance, uniform observational analysis methods can be designed and implemented independent from the underlying EHR database model, either the source is a pharmacoepidemiological database or directly a hospital information system.

In the light of the Common Data Element (CDE) based interoperability approach, we design and implement the Post Marketing Safety Study Tool (PMSST) which can extract any needed information from a patient record after it is retrieved as a result of an eligibility query or it is directly accessed from EHR database within a data mining routine. Our design is built upon the notion of CDEs and it makes use of a Semantic Metadata Registry (MDR) to retrieve data element definitions and use their extraction specifications to access data [75]. With the use of the extraction specifications, PMSST lets the researcher to be able to define what need to be extracted from the patient records with the help of the CDEs accessed from a Semantic MDR. With this dynamic behavior, the researcher writes her methods on the schema/template which will be created based on the data elements that she manipulates. With the help of the underlying interoperability framework, postmarketing surveillance methods do not have to be restricted to the data model of the EHR source. Figure 4.1 presents a schematic representation of the integrated components which enables the execution of the PMSST through the Semantic MDR based interoperability approach that we introduce in this thesis.

PMSST retrieves the CDE definitions from a Semantic MDR where any common data element model can be maintained according to the ISO/IEC 11179 metamodel. Study Data Tabulation Model (SDTM) [19] is a standard data model for the pharmaceutical companies while submitting information about clinical studies to FDA. Pharmaceutical companies like Roche use SDTM variables for data annotation during their postmarketing surveillance studies. In our implementation, the registry maintains the SDTM variables and SALUS Common Data Elements, and there are semantic links between SDTM and SALUS data elements as introduced in Chapter 3. SALUS

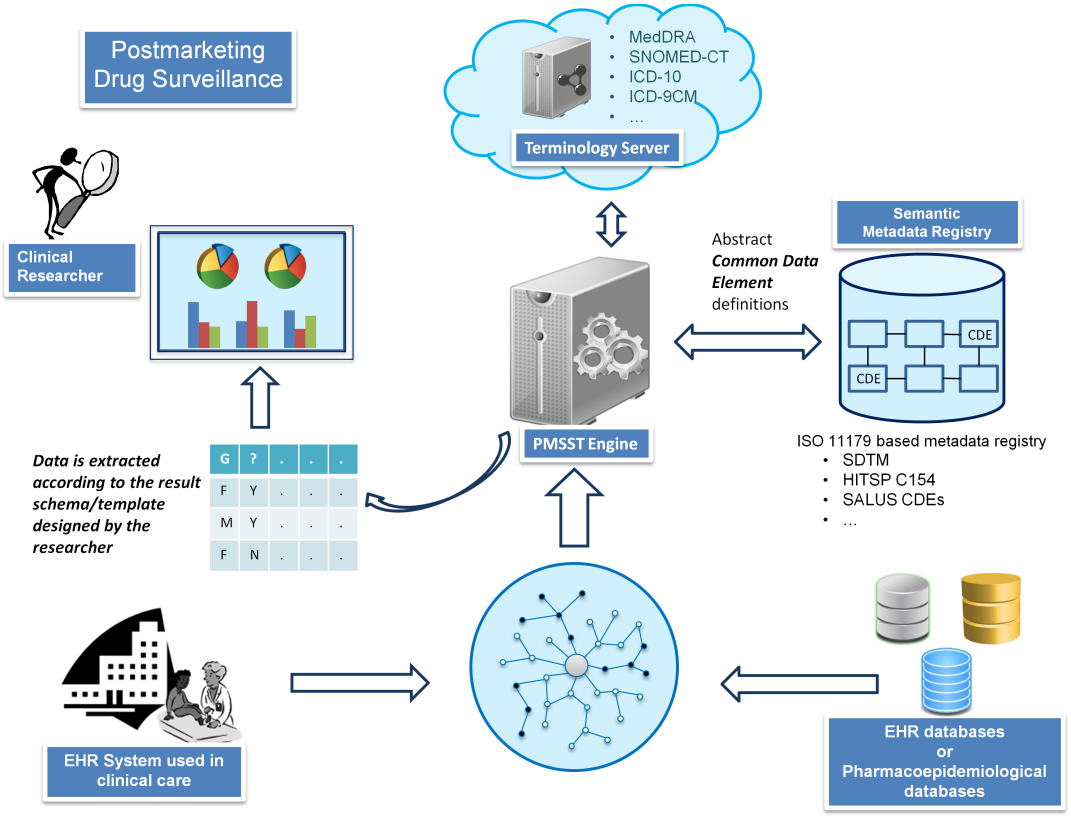


Figure 4.1: Overall architecture of the PMSST integrated with the Semantic MDR based interoperability approach.

interoperability framework exposes an RDF based semantic model, therefore the EHR data is retrieved in the form of the semantic model of SALUS. MDR maintains the abstract definitions of the CDEs which are not bound to any implementation model. The link between the CDE and the concrete model is the "extraction specification". When it is executed, it extracts the data from the data model pointed by the CDE.

Within PMSST, a clinical researcher designs a data schema (a template) by using SDTM variables on which she writes scripts (i.e. SAS [76]) for surveillance studies. The system knows how to extract information from the underlying EHR data by using the extraction specifications of the CDEs. Therefore, the researcher is not bound to the data model of the underlying database; it could be a system providing HL7 CCD based patient summaries or an OMOP database or any other EHR database as well as a pharmacoepidemiological database. As long as the appropriate extraction specifications (i.e. XPath scripts for HL7 CCD) are

available, PMSST can extract necessary information. The communication with the metadata registry is carried out through the IHE Data Element Exchange (DEX) profile in which PMSST implements the metadata consumer role while the Semantic MDR implements the metadata source.

4.3 PMSST Use Case

Background for the Use Case

Congestive Heart Failure (CHF) is a leading cause of hospitalization for patients aged 65 years and older. CHF is of particular concern in diabetic patients in whom incidence rates are two to five times greater than those in the general population. The United Kingdom Prospective Diabetes Study (UKPDS) estimated incidence rates of 2.3-11.9 cases per 1000 patient years in diabetic patients. Several risk factors of CHF in diabetic patients have been identified. These include, for example, duration of diabetes, history of ischemic heart disease, renal function, hypertension, diabetes treatments and HbA1C. However, the incidence of CHF in diabetic patients with a recent acute coronary event is not fully known. In particular, no estimates of CHF for different treatment regimens are available in these patients.

Roche is conducting clinical trials in both acute coronary syndrome (ACS) patients and in ACS patients with diabetes. Whilst the trials are blind, it is important to compare the observed overall incidence rate of an important adverse event like CHF in the trials with that in similar background populations. Such a comparison provides a context to the observed incidence and enables us to identify any potential safety concerns earlier on (e.g. if the observed incidence in the trial is greater than the background).

Objective

The objective of this use case is to estimate incidence rates of CHF in diabetic patients with a recent acute coronary syndrome (ACS) event considering other diabetic medications of patients such as type 2 diabetes (T2D) and related treatment regimes as well. The estimation results should be stratified based on

patient demographics such as age or gender.

Patient Selection

Identify all patients with a first ACS event defined by acute myocardial infarction or unstable angina during the period 2005 to 2011. Include only those patients who have a minimum of 1 year history prior the ACS. Exclude those patients who died within 30 days after the ACS event. Exclude those patients aged less than 18 at the time of ACS. The remaining patients define an ACS cohort of interest. For each patient, the STARTDATE is set to 30 days after the ACS event (so if a patient has an ACS on 5th July 2007, his start date is set to 4th August 2007). We allow a 30 day delay to ensure the ACS has stabilised. For each patient define the LASTDATE as the minimum of (date of death, the date patient transfers out of the system and can provide no more data, 31st Dec 2011)

Result Definition

For the ACS cohort described in the previous section, we identified the necessary result schema (like a common information model required for this surveillance study) composed of several schema items. This can be resembled to the columns of a relational database table. While some of the result schema items can be extracted from the EHR data using the extraction specification of a single SDTM data element, some of the result schema items require further calculations. For instance, whilst the start date of the ACS event can be extracted in a single operation, we should take the start date of the ACS event into account to be able to produce the result for "Average systolic blood pressure (BP) over 12 months before the start of ACS" result schema item; it requires querying of particular measurements within a particular timeframe and calculation of the mean value.

Table 4.1 shows the complete set of the schema items together with the corresponding SDTM data elements. In order to calculate the final results for the schema items, some of the data elements should be provided with specific MedDRA [77] codes to indicate the values according to the requested information which are also indicated in Table 4.1.

Table4.1: Result schema details for the PMSST use case.

Scheme Item Description	Data Elements of the Schema Item	Corresponding SDTM Data Element Name	MedDRA Code for MH.MHPTCD
Sex	Sex	DM.SEX	
Date of birth	Date of Birth	DM.BRTHDTC	
Date of Acute Coronary Syndrom (ACS) event	ACS event Start date of ACS event	MH.MHSTDTC MH.MHSTDTC	10051592
Date of Acute Myocardial Infarction	Acute Myocardial Infarction Start date of Acute Myocardial Infarction	MH.MHPTCD MH.MHSTDTC	10000891
Date of Unstable Angina	Unstable Angina Pectoris Start date of Unstable Angina Pectoris	MH.MHPTCD MH.MHSTDTC	10002388
History of type 2 diabetes (T2D) before start of ACS (Y/N)	T2D Start date of T2D	MH.MHPTCD MH.MHSTDTC	10067585
Date of the first T2D diagnosis ever	Start date of T2D	-	
Average HbA1C over the 12 months before start of ACS	Test name: HbA1C Test value Test unit Test time indicator	LB.LBTESTCD LB.LBORRES LB.LBORRESU LB.LBDTC	
Average systolic blood pressure (BP) over 12 months before start of ACS	Systolic BP measurement Systolic BP value Systolic BP unit Systolic BP time indicator	VS.VSTESTCD VS.VSORRES VS.VSORRESU VS.VSDTC	
Average diastolic BP over 12 months before start of ACS	Systolic BP measurement Systolic BP value Systolic BP unit Systolic BP time indicator	VS.VSTESTCD VS.VSORRES VS.VSORRESU VS.VSDTC	
History of hypertension before start of ACS (Y/N)	Hypertension Start date of hypertension	MH.MHPTCD MH.MHSTDTC	10020772
Last Body Mass Index (BMI) before start of ACS	BMI measurement BMI value BMI unit BMI time indicator	VS.VSTESTCD VS.VSORRES VS.VSORRESU VS.VSDTC	
Last weight before start of ACS	Weight measurement Weight value Weight unit Weight time indicator	VS.VSTESTCD VS.VSORRES VS.VSORRESU VS.VSDTC	
Last length before start of ACS	Length measurement Length value Length unit Length time indicator	VS.VSTESTCD VS.VSORRES VS.VSORRESU VS.VSDTC	
Ever smoked (Y/N)	Smoking	SU.SUCAT	
Smoked within the last 3 months (Y/N)	['Smoking'] Smoking time indicator	SU.SUENDTC	
Taken sulfonylurea anytime within 3 months before start of ACS (Y/N)	Sulfonylurea therapy Therapy time indicator	CM.CMDECOD CM.CMENDTC	
Taken metformin anytime within 3 months before start of ACS (Y/N)	Metformin therapy Therapy time indicator	CM.CMDECOD CM.CMENDTC	
Taken insulin anytime within 3 months before start of ACS (Y/N)	Insulin therapy Therapy time indicator	CM.CMDECOD CM.CMENDTC	

Table4.1: Result schema details for the PMSST use case (continued).

Taken Thiazolidinediones anytime within 3 months before start of ACS (Y/N)	Thiazolidinedione therapy Therapy time indicator	CM.CMDECOD CM.CMENDTC	
Taken other oral anti-diabetic drugs within 3 months before start of ACS (Y/N)	Other anti-diabetic drug therapy Route of drug Administration Therapy time indicator	CM.CMCLAS CM.CMROUTE CM.CMENDTC	
Had a Congestive Heart Failure (CHF) before start of ACS (Y/N)	Congestive Heart Failure Congestive Heart Failure time indicator	MH.MHPTCD MH.MHSTDTC	10007559
Had a CHF after start of ACS (Y/N)	Congestive Heart Failure Congestive Heart Failure time indicator	MH.MHPTCD MH.MHSTDTC	10007559
Date of CHF after start of ACS	Congestive Heart Failure Congestive Heart Failure time indicator	MH.MHPTCD MH.MHSTDTC	10007559
Patient died any time after start of ACS (Y/N)	Date of death	DM.DTHDTC	
Date of Death	Date of death	DM.DTHDTC	

How the Use Case Affected PMSST Design?

The patient selection phase is the execution of the eligibility criteria for retrieving the data of the defined cohort. For PMSST, this execution is handled through the semantic interoperability layer of SALUS. However, this could be any other system like Sentinel or Query Health [78] from which data is retrieved in the form of a content model. As long as the extraction specifications of the selected CDEs to that content model are available and reachable with appropriate links in the Semantic MDR, PMSST can perform the same execution to build data according to the result schema defined by the researcher by the help of the CDEs.

Analyzing the use-case presented in Section 4.3, we elicited the key requirements for PMSST and we based the design of the key functionalities of PMSST on these requirements. During the result schema definition process, values of particular result schema items might be used in defining other schema items. Therefore, PMSST provides a flexible variable definition mechanism. PMSST keeps track of the variable definitions and generates the queries to be applied on the EHR data and organizes their execution order.

As it can be seen in Table 4.1, some of the result schema items need further calculations such as the average value of blood pressure measurements, date of the first occurrence of T2D diagnosis and last weight value before the ACS event.

We design PMSST such that it would present different selection and calculation options automatically considering the value domain of the result schema item.

Value domains of the used CDEs may be referring to different terminology/coding systems. For example, while asking whether a patient has T2D or not, the researcher at Roche uses the MHPTCD element from the MH domain of SDTM. Since, this element requires a coded value from MedDRA [77], researcher should be easily assigning values to such data elements during her schema design. For this purpose, PMSST has been integrated with a terminology server so that it would recommend possible values based on the result schema item through a type-a-head search mechanism.

4.4 PMSST System Description

PMSST is a web based tool which can be used from modern web browsers. It has been implemented with the latest high performance web technologies incorporating HTML5 design principles and RESTful client-server communication. The tool is composed of an eligibility query execution and a data selection part. The former is out of the scope of this thesis: Upon the execution of an eligibility query, a cohort of patient data is retrieved in the form of a content model. We claim that the CDE based interoperability implementation of PMSST can make use of any content model as long as the appropriate extraction specifications are available for the abstract CDE definitions within the Semantic MDR framework.

Figure 4.2 presents the data selection phase of PMSST. The user can define a result schema at this phase by using the CDE definitions retrieved from the Semantic MDR. In our implementation, the registry maintains SDTM variables and SALUS CDE set according to ISO/IEC 11179 meta-model principles. When the user decides to use SDTM, the object classes (aka domains) are presented to the user to give a top-down browsing experience. When a domain is selected, the data elements created out of that object class is presented to the user. When a CDE is selected, it appears on the left hand side to create further calculations.

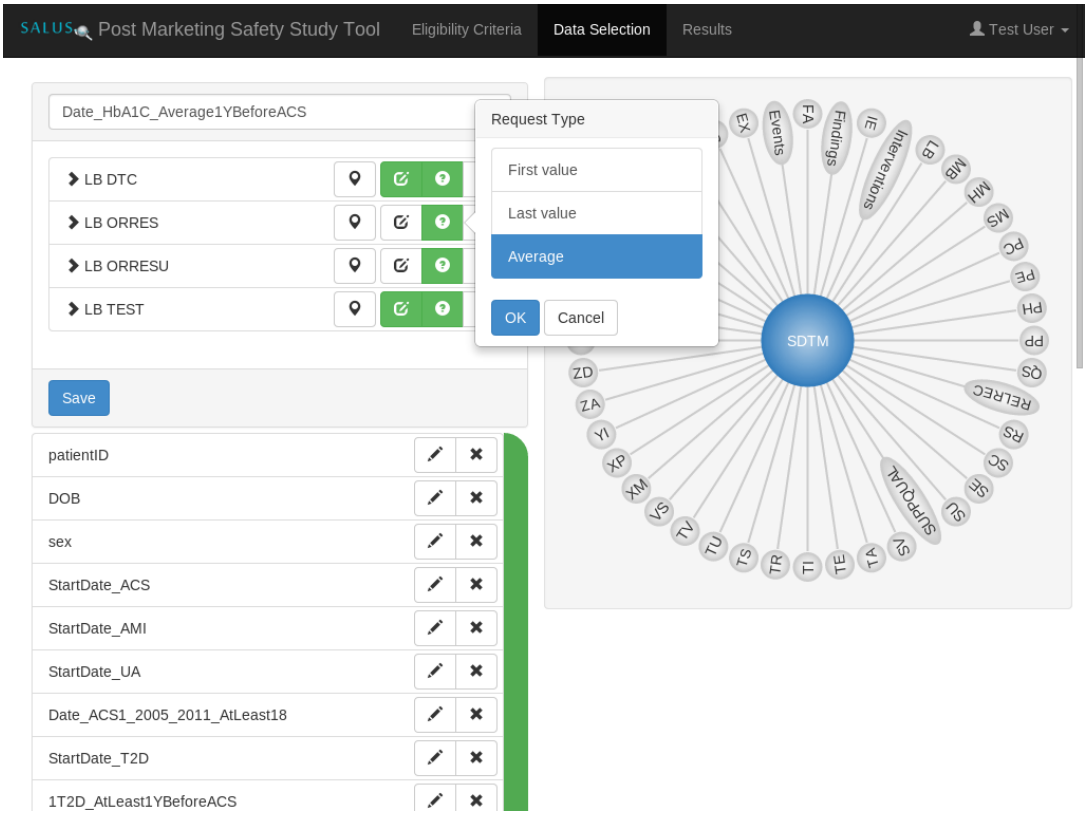


Figure 4.2: A snapshot of PMSST while the researcher defines a result schema. On the right hand side, domains of SDTM forms a circle; if selected, then CDEs of that domain forms the circle. On the left hand side, a schema item: Date_HbA1C_Average1YBeforeACS is created out of 4 SDTM elements. Below that, a list of other schema items are shown.

Once a schema item is designed, it is saved and schema design continues. The user can edit or delete an existing item anytime during the design phase.

4.4.1 Data flow between components

PMSST is composed of several different components among which a number of integration mechanisms exist. In Figure 4.3 the flow of data between those integrated components are depicted and the steps of the flow are described in the following.

Figure 4.3 shows the steps of the data flow during the execution of PMSST and the steps can be described as follows:

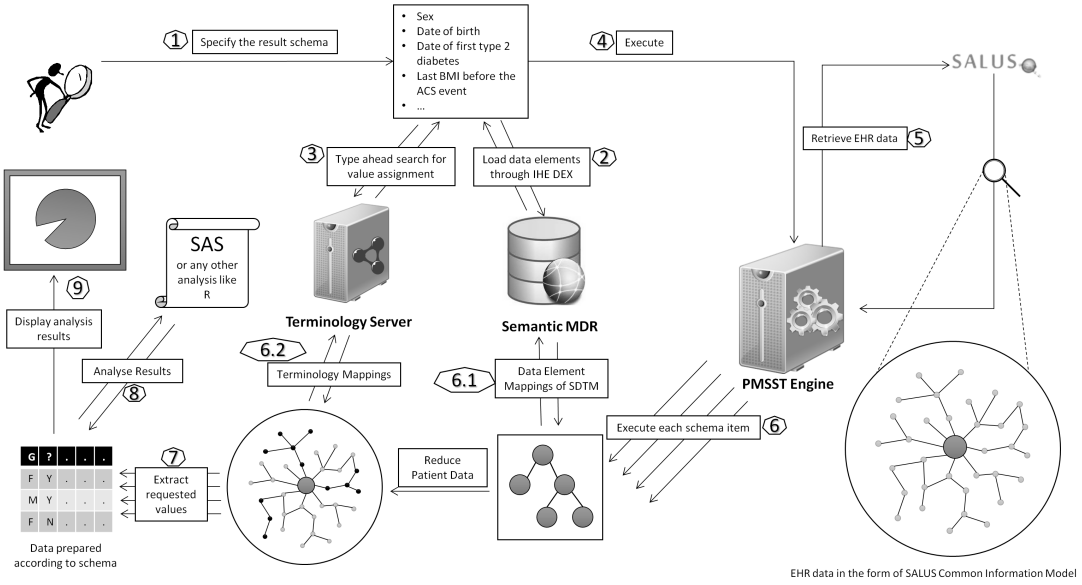


Figure 4.3: Step-by-step representation of the data flow between different components. A clinical researcher uses PMSST in order to define a result schema so that when patient data is retrieved from the underlying EHR source(s), data will be automatically transformed to that schema.

1. The researcher uses a web browser to define the result schema by using the CDEs. Roche uses SDTM variables in our deployment as identified in Table 4.1.

2. CDEs are maintained in the Semantic MDR and retrieved through the IHE DEX profile. The user browses the CDEs starting from the object classes in a top-down fashion.

3. If the user likes to restrict the value of a selected data element (i.e. set Acute Myocardial Infarction to MHPTCD element), possible values can automatically be searched from the terminology server. PMSST knows in which coding system to look for the term by analyzing the value domain of the CDE definition automatically.

4. After the user completes the result schema, i.e. defining each schema item by using the abstract CDE definitions, the schema definition is sent to the PMSST engine on the server side.

5. Eligibility query is sent to the SALUS system and EHR data is retrieved in the form of SALUS Common Information Model.

6. For each schema item definition, PMSST engine extracts information from the EHR data and performs necessary calculations to place into the appropriate location according to the schema definition.

    6.1. Result schema is defined by SDTM elements. Semantic MDR keeps the mappings between SDTM and SALUS CDEs as presented in Table 4.2. And, SALUS CDEs has the extraction specifications to access the necessary information from the EHR data. CDE definitions, mappings and extraction specifications are retrieved from the Semantic MDR in conformance to the IHE DEX profile. Since SALUS CIM is an RDF based model, the extraction specifications of the SALUS CDEs are SPARQL scripts.

    6.2. If the schema item definition includes a value in one of its defining CDEs, value analysis should be done. However, in our deployment, EHR data is coded with ICD9-CM system while SDTM elements refer to MedDRA or NCI terms. The terminology server includes mappings between these different coding systems and PMSST can do value matching with the help of this terminology server.

7. Data is produced conforming to the result schema defined by the researcher.

8. The user can write analysis methods on top of this schema independent from the underlying EHR source model. In our deployment, Roche implements SAS scripts to do the analysis.

9. Finally the analysis results are presented to the researcher.

4.4.2 CDE mappings

PMSST makes use of the abstract CDE definitions retrieved from the Semantic MDR. In order to enable the retrieval of the extraction specifications given the SDTM variables, we mapped the SDTM elements to the SALUS CDEs. We implemented an automatic content model importer on top of the open API of the Semantic MDR for importing the SDTM variables and their mappings to

SALUS CDEs. Since SALUS CDEs have also mappings to HITSP C154 Data Dictionary [35] elements, our work transitively created a link between SDTM variables and HITSP C154 elements. Table 4.2 lists the mappings used during the execution of our implementation.

Table4.2: Mappings of the Common Data Elements: SDTM – SALUS CDE set – HITSP C154 Data Dictionary

SDTM	SALUS CDE	HITSP C154
DM	Patient	Personal Information
DM.DMSEX	Patient.Gender.CD	1.06 Personal Information Gender
DM.DMBRTHDTC	Patient.DateOfBirth.Date	1.07 Personal Information Person Date of Birth
DM.DMDTHDTC	Patient.TimeOfDeath.Date	7.09 Conditions Time of Death
MH	Patient.Condition.Condition	Conditions
MH.MHPTCD	Condition.ProblemCode.CD	7.04 Conditions Problem Code
MH.MHSTDTC	Condition.TimeInterval.IVLTS	7.01 Conditions Problem Date
LB	Patient.Result.Result	Result
LB.LBTEST	Result.Type.String	15.03 Result Result Type
LB.LBTESTCD	Result.Type.CD	15.03 Result Result Type
LB.LBORRES	Result.Value.PQ	15.05 Result Result Value
LB.LBORRESU	Result.Value.PQ	15.05 Result Result Value
LB.LBDTC	Result.TimeInterval.IVLTS	15.02 Result Result Date/Time
VS	Patient.VitalSign.Result	Vital Sign
VS.VSTESTCD	Result.Type.CD	14.03 Vital Sign Vital Sign Result Type
VS.VSTEST	Result.Type.String	14.03 Vital Sign Vital Sign Result Type
VS.VSORRES	Result.Value.PQ	14.05 Vital Sign Vital Sign Result Value
VS.VSORRESU	Result.Value.PQ	14.05 Vital Sign Vital Sign Result Value
VS.VSDTC	Result.TimeInterval.IVLTS	14.02 Vital Sign Vital Sign Result Date/Time
SU	Patient.SocialHistory.SocialHistory	Social History
SU.SUCAT	SocialHistory.ObservationCode.CD	19.02 Social History Social History Type
SU.SUENDTC	SocialHistory.TimeInterval.IVLTS	19.01 Social History Social History Date
CM	Patient.Medication.Medication	Medication
CM.CMDECOD	MedicationInformation.ActiveIngredient.CD	8.13 Medication Coded Product Name
CM.CMENDTC	Medication.TimeInterval.IVLTS	-
CM.CMROUTE	Medication.Route.CD	8.07 Medication Route

Although the usage of the tool starts with defining an eligibility criteria and retrieving EHR data according to that query, our implementation is independent of the content model according to which the EHR data is shaped. For example, if the underlying EHR system can provide HL7 CCD based patient summaries, then PMSST can seamlessly process the data by using the corresponding extraction specification retrieved from the Semantic MDR. That is possible because HITSP C154 defines XPath expressions from its CDE definitions to HL7 CCD based documents and PMSST can retrieve the

extraction specifications through the HITSP C154 mappings. This time, the extraction specifications would be XPath expressions and the clinical researcher would not be aware of this. It means that PMSST can automatically communicate with an EHR system which is capable of exporting HL7 CCD based document summaries and make the data available for clinical research automatically.

CHAPTER 5

E-BUSINESS CASE STUDY

Federated Semantic Metadata Registry based interoperability approach can be applied to different domains in order to succeed data interoperability. Although our primary motivation comes from eHealth domain, in this chapter we present a case study where the introduced framework is used to extract data automatically from different electronic business documents conforming to different standards.

Electronic Business (eBusiness) has been one of the key players in the field of document interoperability even before the settlement of the modern Internet. Start of the related research goes back to late 1960s [79, 80]. From a general perspective, we can divide the approaches of the researchers to the interoperability problem into two. As described in Chapter 1, the early and practical approach is defining formal interfaces (i.e. document schemas) for information exchange between applications. This approach exhibits a top-down vision in which information models, entities and their interactions are strictly modeled according to the document schemas. Following this top-down strategy, today, there are several different eBusiness document standards in use which are developed by different standardization bodies for different application areas of eBusiness.

Starting with Electronic Data Interchange (EDI) [81] and continuing with XML based standards, early interoperability solutions for eBusiness documents followed a top-down strategy and document schemas have been defined. There are several industry specific standards which have been developed with this strategy such as the ones from the North American Automotive Industry

Action Group [82], Health Level 7 Standards Development Organization [83], the Petroleum Industry Data Exchange (PIDX) committee [84], the Chemical Industry Data Exchange (CIDX) organization [85], Open Travel Alliance [86], and RosettaNet Consortium [87].

Since there cannot be a single standard which fits to every requirement, having different document standards has created a new interoperability problem for eBusiness applications. An effort for the solution of this problem has been the UN/CEFACT Core Components Technical Specification (CCTS) [88] which follows the bottom-up strategy to define Core Components for the eBusiness domain. UN/CEFACT CCTS provides a methodology to identify a set of reusable building blocks, called Core Components to create electronic documents. Core Components represent the common data elements of everyday business documents such as Address, Amount, or Line Item. These reusable building blocks are then hierarchically assembled into business documents such as Order or Invoice by using the CCTS methodology. Core Components are defined to be context-independent so that they can later be restricted to different contexts such as a specific industry or a country. Many Core Components defined by UN/CEFACT are available to the business systems from UN/CEFACT Core Component Library [89].

Having UN/CEFACT CCTS in action, new standards have been released following the CCTS methodology and referred to the Core Component Library while designing their own artifacts. Although major electronic business document standards are based on CCTS at some level; this does not make them interoperable. The analysis on

- OASIS Universal Business Language (UBL) [90]
- OAGIS Business Object Document (BOD) [91]
- Global Standards One (GS1) XML [92]

reveals that there are considerable differences in their document design principles: the use of code lists and the XML namespaces, how they use the CCTS methodology and how they handle extensibility and customization [80].

Furthermore, the current accepted practice of storing the document artifacts in spreadsheets does not facilitate developing automated semantic interoperability support tools.

In addition to UN/CEFACT Core Component Library, standards are creating their own data element libraries or their completely independent information models. GS1 Global Data Dictionary [93] can be given as an example to the former while ENTSO Common Information Model [94] to the independent information model. These efforts try to ensure interoperability within the boundaries of the associated initiatives. When it comes to achieving a broader range of interoperability, there is a need to establish automatic data extraction architecture between different standards through an easily manageable framework.

In this chapter, we show the use of our Semantic Metadata Registry [9] based solution for the interoperability problem between different electronic business document standards. We base our solution on the advantages of the bottom-up and the top-down approaches with a unified view on top of the semantic web technologies. We present that instead of document translations between independently evolving standards, it is more manageable and cost-effective to address automatic data extraction from the documents through abstract Core Component definitions and their implementation specific extraction specifications.

We model the Core Components as abstract common data element definitions. The Semantic Metadata Registry maintains these abstract data element definitions which are not bound to any standard's implementation model – the schema. The link between the abstract data element and the schema is the extraction specification; when it is executed, it extracts data from the document instance from the location pointed by the data element definition. For example, in our case study context, we can model the e-mail address of the sales contact with an abstract data element definition and an XPath expression – which can extract the e-mail address from a UBL Invoice document instance – serves as the extraction specification of that data element

for the UBL schema. The same data element can have different extraction specifications (XPath expressions) for different document schema standards.

It is an experienced fact that standards evolve in time and enterprises are forced to follow the changes by developing new adapters for document processing. Figure 5.1 shows a schematic representation of the interoperability problem between different eBusiness document standards. Each standard ensures document interoperability within the associated boundaries. We observe that some standards (i.e. UBL) are customized further which leads to sub-boundaries where interoperability issues arise. Using semantic technologies in line with the already available methodologies brings a new opportunity for easy management of standards development from the standardization body point of view and easy adaptation to new standards and updates from the enterprise point of view.

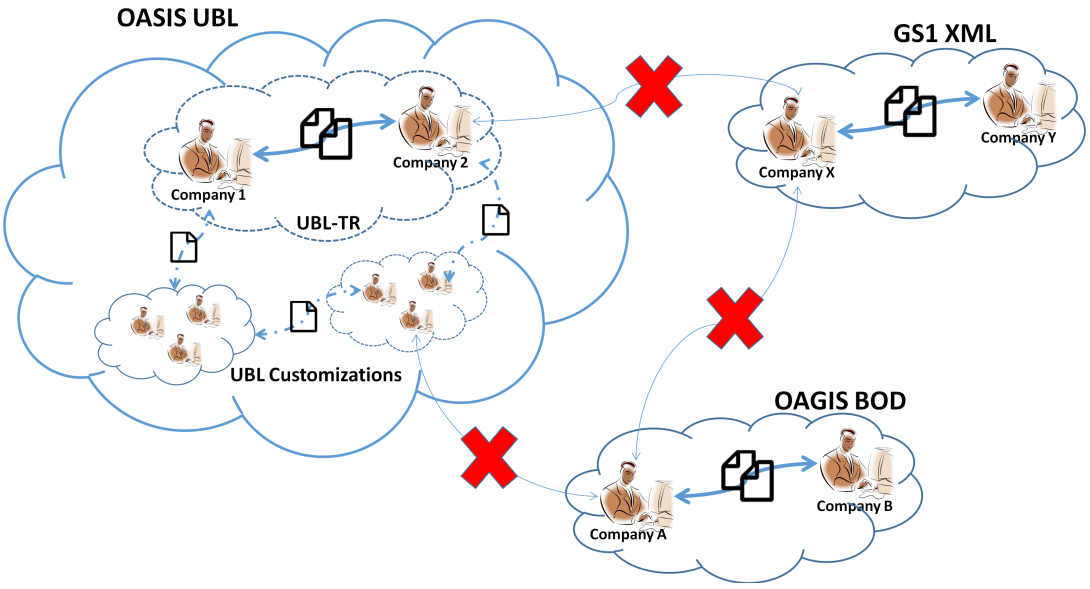


Figure 5.1: Interplay of the major eBusiness document standards. Electronic document exchange is problematic across boundaries.

In this case study of the eBusiness domain, we present the implementation results of the introduced solution with the use of a single semantic metadata registry which maintains common data elements defined by UN/CEFACT Core Component Library. These data elements have been imported to the knowledge base of the semantic metadata registry and extraction specifications have been defined. In our context, since the three standards – OASIS UBL,

OAGIS BOD and GS1 XML – are XML based specifications and have their XML Schema Definitions, extraction specifications are XPath expressions. With this solution, any standard can evolve in its own direction by ensuring the correct XPath definitions for the Common Data Elements maintained in the Semantic Metadata Registry. There is no need to invest on mappings and translations between different document standards. Each update on the standards causes serious updates on these translations. Our solution increases the decoupling between the interoperating applications from the syntax and semantics of the underlying document specifications.

5.1 Semantic Representation of the Core Components

UN/CEFACT's aim in developing the Core Component Library through CCTS methodology is to let document schemas be designed from standard, reusable building blocks. This bottom-up approach tries to support the top-down schema definitions by providing a common baseline for all electronic business documents. The Core Components have gained widespread adoption by dominant electronic document standardization bodies. OASIS UBL, OAGIS BOD, GS1 XML, CIDX and many other standards have taken up the CCTS methodology. However, existing standards have well-established document schemas which are already in use and radical schema modifications for conforming to CCTS cause backward incompatibility problems. Therefore, they apply the CCTS methodology selectively and more importantly do not always base their document artifacts on the core components defined in the UN/CEFACT Core Component Library. This resistance to the adoption of the new standards can also be observed from the fact that electronic business interoperability is still achieved heavily through EDI based messages mostly due to the existing infrastructure investments.

In this study with the use of the Semantic Metadata Registry and extraction specifications for the Core Component definitions, we increase the decoupling between data extraction routines from structured documents and the underlying application logic. Therefore, this solution can easily be integrated with the

already existing electronic business document standards.

UN/CEFACT Core Component Library is published as a spreadsheet. Table 5.1 shows a number of Core Components from the lately released CCL sheet [89]. According to the CCTS methodology, electronic business document standards are expected to base their artifacts on the published Core Components. However, this is not the case. First of all, published Core Components are not machine-processable. Our solution proposes to maintain these common data elements in an ISO/IEC 11179 based Semantic Metadata Registry and define extraction specifications for the associated standards and establish appropriate links among the data elements. Compared to the spreadsheets, this architecture exposes the abstract Core Component definitions in a structured form through a standard meta-model. Moreover, the semantic representation of the Core Components enhanced with commonly used knowledge organization systems and Linked Data principles makes them searchable, accessible, dereferanceable and processable through standard and federated mechanisms.

Table5.1: A part of the UN/CEFACT Core Component Library. Core Components are published through spreadsheets.

UN00000011	Address. Identification. Identifier	A unique identifier for this address.
UN00000012	Address. Format. Code	The code specifying the format of this address.
UN00000014	Address. Postcode. Code	A code specifying the postcode of the address.
UN00000032	Address. Post Office Box. Text	The unique identifier, expressed as text, of a container commonly referred to as a box, in a post office or other postal service location, assigned to a person or organization, where postal items may be kept for this address.
UN00000019	Address. Block Name. Text	The block name, expressed as text, for an area surrounded by streets and usually containing several buildings for this address.
UN00000020	Address. Building Number. Text	The number, expressed as text, of a building or house on a street at this address.
UN00000021	Address. Building Name. Text	The name, expressed as text, of a building, a house or other structure on a street at this address.

UN/CEFACT CCTS methodology is based on the ISO/IEC 11179 standard. That is, data elements of the UN/CEFACT Core Component Library are

based on the meta-model exposed by the ISO/IEC 11179 standard. Figure 5.2 shows a decomposition of the "Address.Postcode.Code" element according to the modeling constructs of the ISO/IEC 11179. This kind of decomposition increases the reuse of building blocks across different data elements, which leads to an increased data interoperability. Although the grounding of the CCTS is to stimulate the reuse of the common data elements across different standards, the document schemas of the standards have not evolved in this direction. We believe that one of the important reasons behind this diversion is the lack of the machine processable definitions of the published Core Components. By importing the Core Components into the knowledge base of the Semantic Metadata Registry, the Core Components are represented in a structured and standard way fully conforming to the ISO/EIC 11179 decomposition.

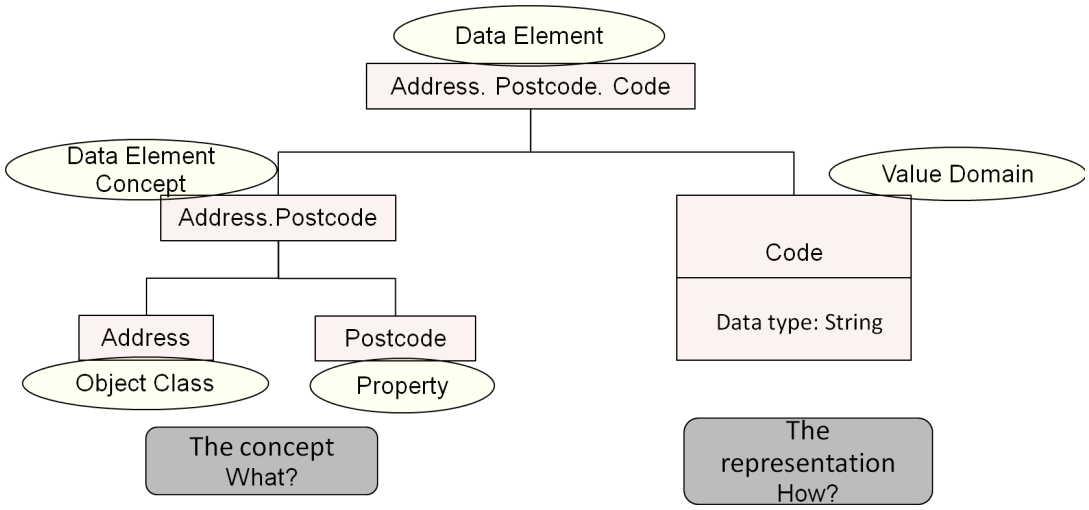


Figure 5.2: Decomposition of a Core Component – Address. Postcode. Code – according to the Object Class, Property, and Value Domain constructs of the ISO/IEC 11179 meta-model.

Table 5.2 shows a portion of the semantic description for the "Party.PostalAddress" element. We implemented an automatic importer which reads the Core Components from the published spreadsheets and creates the corresponding registry structures in the Semantic Metadata Registry. Since the spreadsheets of UN/CEFACT are aligned with the CCTS methodology which follows the ISO/IEC 11179 rules, creating the correspondences depicted in Figure0 5.2 was straightforward in terms of processing the input files. The

semantic representation of a Core Component is conformant to the ISO/IEC 11179 ontology, which is an OWL ontology directly implements the ISO/IEC 11179 meta-model. This ontology is the information schema of the metadata registry in a top-down fashion. On the other hand, the internal mechanics of the Semantic Metadata Registry refers to the well known knowledge organization systems like SKOS and depicts mappings to the other data elements or components of the data elements. Since every component – a data element, an object class or a data element concept – is uniquely identified and dereferenceable, semantic applications can hop over the links and find the abstract definition of the data element in question. By retrieving the extraction specification of the data element, the application can extract data from a document instance automatically.

Table5.2: A part of the N3 representation of the Party. Postal Address element exported from the Semantic Metadata Registry

```
:DE_5539c300-a872-4fba-8973-779f3561beda
        a owl:Class;
        rdfs:subClassOf :DataElement;
        :administeredBy :STEW_66165477-3532-4e43-8860-865281cf6e1e;
        :classifiedBy :CSI_618cac6b-5f6c-4675-b84d-da21ca681653;
        :dataElementAdministrationRecord :AR_0a701528-375b-4e63-b7ba-377840;
        :expressingDataElementConceptExpression :DEC_f0982a9b-513-b40b-89162;
        :having :AIC_8b9f7cff-28b3-47eb-99b8-84cac1a661f4;
        :registeredBy :RA_eu.salusproject.tr.com.srdc.mdr_null;
        :representedByDataElementRepresentation :NEVD_178bf-3d1f-4fa2-ac82-c3656;
        :submittedBy :SUB_9ee85362-d3f6-4d3c-8322-2750ace809fb.

:CSI_618cac6b-5f6c-4675-b84d-da21ca681653
        a owl:Class;
        rdfs:subClassOf :ClassificationSchemeItem;
        :classificationSchemeItemTypeName "XPATH"^^xsd:string;
        :classificationSchemeItemValue "Party/cac:PostalAddress"^^xsd:string;
        :containedIn :CS_c9cf804f-9fc9-4768-b66f-f3e7a465be64.

:CS_c9cf804f-9fc9-4768-b66f-f3e7a465be64
        a owl:Class;
        rdfs:subClassOf :ClassificationScheme;
        :administeredBy :STEW_38d27178-3715-4b0c-91c9-3f1c42ff9965;
        :classificationSchemeAdministrationRecord :AR_3fa3-40ca-ba34f223162a6d;
        :classificationSchemeTypeName  "UBL OID"^^xsd:string;
        :containing :CSI_618cac6b-5f6c-4675-b84d-da21ca681653;
        :having :AIC_90c4cdfd-d50e-43d4-9bfc-cb332b20a54e;
        :registeredBy :RA_eu.salusproject.tr.com.srdc.mdr_null;
        :submittedBy :SUB_85754d70-fa18-4b9a-b4f7-2f4a2b8432d8.
```

In Table 5.2, it can be observed that the Core Component: “Party. Postal Address” is a data element which is represented with “rdfs:subClassOf :DataElement” relation by referring to the constructs of the ISO/IEC 11179 OWL ontology. Mappings to other Core Components or representation of the

extraction specifications are handled through the classification scheme items of the ISO/IEC 11179 meta-model. Second row of Table 5.2 is the serialization of a classification scheme item whose type is XPATH and value is the XPath expression to extract the postal address from a Party component in a business document.

In this study, we implement the introduced methodology by importing the Core Components of UN/CEFACT into a single installation of the Semantic Metadata Registry. However, a federated framework can be utilized to make the most advantage out of the introduced methodology. There are several different document standards in eBusiness domain where major ones are based on UN/CEFACT Core Component Library. Whether based on UN/CEFACT CCL or not, standards continue to evolve in their own directions and they build their own common data elements as in the case of GS1 GDD. In addition to our proposed architecture for automatic data extraction through extraction specifications, we propose that these efforts can be coordinated with the help of the semantic web technologies. Each organization can maintain its own data elements in a Semantic Metadata Registry and they can be linked with each other within the Linked Open Data Cloud. We briefly elaborate on this with a scenario in the next section.

5.2 Semantic MDR in eBusiness

In this case study, we have implemented a client which simulates an arbitrary Enterprise Resource Planning (ERP) component for information extraction from electronic business documents. The software makes use of the Semantic Metadata Registry to retrieve abstract Core Component definitions and access the extraction specifications for schema that the document instance conforms to. In Figure 5.3, we present the data flow in a use case scenario where we suppose that our client implementation is a component of any arbitrary ERP software. Our claim is that this interoperability approach can be used within a coordinated environment instead of message translations or individual adapters for different document standards.

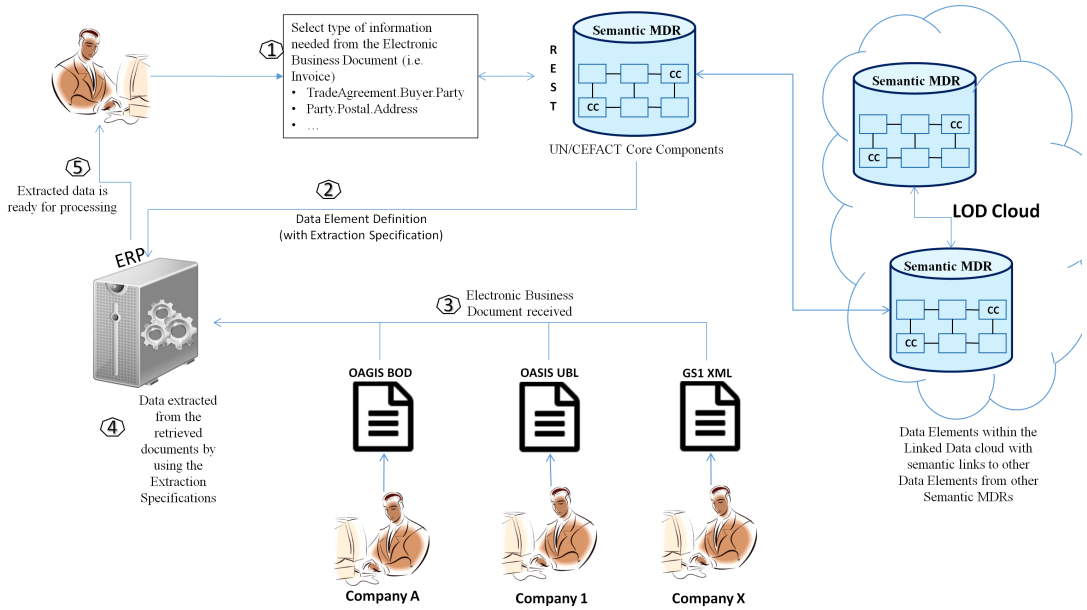


Figure 5.3: The hypothetical ERP (Enterprise Resource Planning) component which implements the introduced automatic data extraction mechanism. This architecture can be extended within the LOD cloud by using several linked Semantic Metadata Registries.

The steps of the scenario in Figure 3 can be described as follows:

1. The client software (which is a component of the hypothetical ERP system in our scenario) has the list of the Core Components - with unique identifiers – for which the values should be extracted from the business document retrieved from a business party.

   (a) Core Components can be searched through the exposed interface of the Semantic Metadata Registry.

   (b) In ordinary business logic, ERP software already knows what to look for in the received electronic business document. In any case, the system has the unique, dereferenceable links of the Core Components.

2. Through the exposed HTTP-REST interface, the links (URIs) are requested from the Semantic Metadata Registry. For example, the semantic representations of TradeAgreement.Buyer.Party and Party.Postal.Address are retrieved which are in the similar form shown in Table 5.2. Federated query service of the Semantic Metadata Registry

can follow the links between data element mappings if exists between different Semantic Metadata Registry installations. However; in our implementation, we utilize a single registry installation. In this single Semantic Metadata Registry, UN/CEFACT Core Component Library is maintained where the Core Components are annotated with three extraction specifications to OAGIS BOD, OASIS UBL and GS1 XML document schema standards.

3. ERP receives an electronic business document in one of the standards addressed in the implementation of this paper: OASIS UBL, OAGIS BOD or GS1 XML. For example, a UBL Order document is received from the ERP of a remote business party who wants to order some goods from this party.

4. The client software can automatically extract data from the received business document (UBL Order document) through the XPath expressions for the Core Components. In the semantic description of the Core Components, XPath expressions (the extraction specifications) are given through the ISO/IEC 11179 ClassificationSchemeItems. This is exemplified in Table 5.2.

5. The ERP can use the extracted information for further processing. In our implementation, the client software shows the extracted values in order to validate the interoperability framework.

Semantic Metadata Registry provides a highly user-friendly web based interface for the Core Component management in the meta-data level. Create, update and delete operations for the common data element definitions and their components like object classes and properties can be performed from the web based interface. Mappings between different data elements can be established and extraction specifications can be defined. As shown in Figure 5.4, data elements are presented through the ISO/IEC 11179 meta-model constructs together with the extraction specifications and semantic mappings. Since a persistent triple store exists in the background and the interface of the Semantic Metadata Registry ensures compatibility to the ISO/IEC 11179

ontology, semantic descriptions of the Core Components are convenient for easy consumption of electronic business applications.

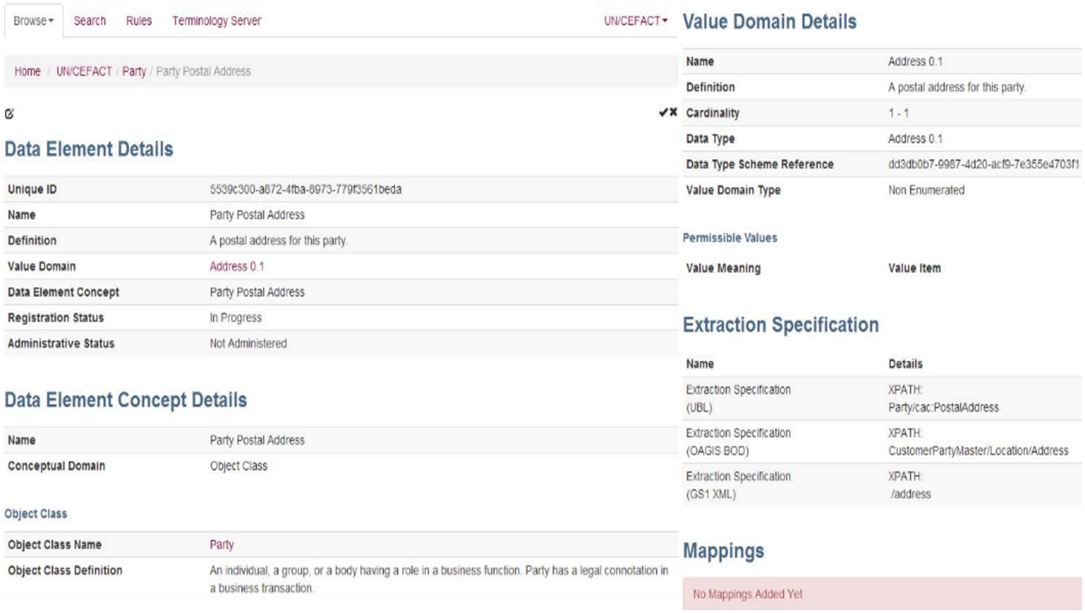


Figure 5.4: Web based graphical user interface of the Semantic Metadata Registry. Core Components can be managed according to the ISO/IEC 11179 metamodel through web-based actions.

During the implementation of the introduced solution, UN/CEFACT Core Component Library – which includes 6299 Core Components – has been imported to the knowledge base of a Semantic Metadata Registry with an automatic content model importer. We have utilized a single installation of the Semantic Metadata Registry which holds the semantic descriptions of the UN/CEFACT Core Components. We defined the extraction specifications – the XPath expressions – by comparing the Core Components with the artifacts of the standards in question, namely OASIS UBL, OAGIS BOD and GS1 XML. Since these standards are based on UN/CEFACT CCTS, it was possible to find a corresponding artifact and generate the XPath expression for the Core Components of UN/CEFACT from the mentioned standards. This means we performed naïve mapping between the artifacts of these standards by using UN/CEFACT Core Components as a hub. It is important to note that our intention is to implement and demonstrate that the introduced solution in this thesis can be used for electronic business document interoperability through automatic data extraction. We evaluated the implementation with a dummy

client software implementation which can be imagined as a part of an ERP system. We describe a usage scenario with a step-by-step use case implementation.

CHAPTER 6

RELATED WORK

There is extensive research on data interoperability in literature in several different application domains. As introduced in Chapter 1, a high level perspective can classify the interoperability approaches as top-down and bottom-up. We observe that as the technology advances (i.e. Semantic Web technologies and Linked Data principles), this high level classification persists.

In Chapter 2, a background and brief introduction to some fundamental examples for the related work are presented. Metadata registry based strategies for the implementation of the bottom-up approach and lately the Linked Open Data paradigm which also implements the bottom-up data interoperability approach are introduced. ISO/IEC 11179 standard has been introduced as the backbone of the metadata registry approach. On the other hand, we see that, in different domains, there are efforts to build common information models which lead to formally defined interfaces. Either semantic (i.e.ontology based) or not, such top-down approaches require the interoperating systems to translate the data to the indicated common information model.

Research on ISO/IEC 11179 continues as it gets to its newer versions with the updates according to the requirements of several domains in which metadata should be managed through controlled mechanisms. Electronic health is one of the important domains in this target. Ngouongo et al. [28] discusses whether ISO/IEC 11179 covers healthcare standards in empirical research or not, and presents the analysis results. It is claimed that ISO/IEC 11179 is a strong

candidate to become a norm for the management of healthcare standards such as [95].

There are several efforts trying to address the interoperability between the clinical research and patient care domains. One major approach to the problem of semantic interoperability is to build a common data model where the interoperating systems are required to interact through this well-defined data model. The research behind OMOP CDM [96], FDA Mini-Sentinel [97], I2B2 [98], STRIDE [99] and EU-ADR [100] are among some of the efforts that adopt this top-down strategy to reuse existing EHR data for the clinical research purposes. In addition to these projects, Laleci et al. [1] builds a semantic common information model to exchange data between clinical research and clinical care systems. Weber et al. [101] presents a prototype of a federated query tool for clinical data repositories through a common information model.

Another major approach is to identify the CDEs of the content models of the interoperating systems and provide direct mappings between them. Apart from the strict relations within a content model, this approach attaches more importance to the elicitation of the data elements. Fadly et al. [102] presents mapping algorithms to identify semantic coherence between clinical care and clinical research data elements in order to pre-populate electronic Case Report Forms. Jiang et al. [103] presents a prototype implementation for CDISC SHARE using already available semantic tools where they try to provide an environment for CDE management. Kunz et al. [104] utilize a repository of the CDEs to help developers reuse appropriate elements to enable interoperability of their systems. Pathak et al. [105] analyses the effects of adoption of the CDEs in large-scale epidemiological and genome-wide studies on cross-study analyses.

In eHealth domain, there are also several standardization efforts addressing the problem of semantic interoperability in question. The IHE Drug Safety Content (DSC) [106] and Clinical Research Data Capture (CRD) [107] profiles are two efforts to address pre-filling of safety reports and case report forms (CRF) by retrieving the data from medical summaries expressed in HL7 CCD format.

However, both of these profiles propose static Extensible Stylesheet Language (XSL) [108] mappings between a predefined medical summary template in CCD and a generic CRF form. This approach is not flexible and extensible; these XSL mappings are only valid for the given pre-population data formats; once these pre-population data templates are modified due to emerging requirements, new mappings are needed [109, 110]. The new IHE Data Exchange (DEX) [75] profile proposal in IHE QRPH domain addresses the shortcomings of IHE CRD and DSC profiles. A metadata registry is envisioned to maintain the research and healthcare CDEs, and the exact correspondences between them. In this thesis, we extend this idea by providing a semantically linked federated MDR framework to show how the DEX idea can scale in the presence of disparate CDE definition efforts by different organizations.

Our proposal tries to unify the top-down and bottom-up approaches with an analogy to the unification of old, well-established methods with newly emerging, latest technologies. Tao et al. [111] already shows the value in using OWL to represent relational meta-models, including ISO/IEC 11179. Shukair et al. also makes use of the semantic web technologies to address the semantic interoperability of metadata repositories [112]. They create a custom mechanism in order to share the data element definitions across different metadata repositories. They do not address the data exchange part which comes after metadata exchange. To the best of our knowledge, our work is the first attempt to apply a comprehensive set of semantic web technologies with the commonly adopted MDR standard – ISO/IEC 11179 – through the Linked Data principles and applying the extraction specification aspect on top of the abstract common data element definitions.

Current research on postmarketing surveillance for pharmacovigilance and pharmacoepidemiology tries to unify the available EHR data on a common information model. Most of the time, this forces the EHR systems to implement the necessary adapters for transforming data into the defined common model and persist in a separate database, either central or distributed. On the other hand, some approaches transform the query to the native data model at each transaction. However, data and processing

requirements of different areas of clinical research change in time while the quality, quantity and availability of EHR data on patient care side increases. This forces the researchers to update their information models accordingly or come up with the new ones. The literature exemplifies this situation clearly.

Vaccine Safety Datalink [113] is an early initiative for transforming EHR data for post marketing safety surveillance of vaccines. FDA's Sentinel initiative and the Mini-Sentinel pilot system [24, 97] is one of the latest and important efforts for post marketing surveillance, built on the experiences of Vaccine Safety Datalink. Mini-Sentinel builds a distributed system to answer safety queries of clinical researchers through a common information model. OMOP introduces its own Common Data Model (CDM) [23] to transform EHR data. Informatics for Integrating Biology and the Bedside (i2b2) [27, 98] is another parallel effort with similar objectives and exposes its own common information mode. CPRD [70] is a European example of the latest pharmacoepidemiological databases and there are several ongoing projects supported by European Medicines Agency and European Commission using a common information model for surveillance activities. The fact is that those common information models are not so "common"; they are only used within the boundaries of the associated initiatives and projects.

In this thesis, we introduce the Post Marketing Safety Study Tool in Chapter 4 which utilizes a different interoperability architecture than existing common information model based efforts. Our architecture makes use of the interoperability approach that we introduce in this thesis in which the abstract CDE definitions are bound to the implementation specific content models through the extraction specifications. This lets the researcher use any set of abstract CDEs and design its study based on a model depicted by those CDEs. The prerequisite is that the CDEs should be imported to the knowledge base of the Semantic MDR and appropriate links between different CDE sets should be established. In our work, we use automatic content model importers for CDE acquisition and establish the necessary links between SDTM variables and SALUS CDEs. We know that there are several initiatives defining abstract CDE sets and map to existing sets and content models. Hence; we believe

these different initiatives can contribute to the postmarketing surveillance activities and to the field of clinical research informatics in a general sense, if considered like a network of different common information models.

Regarding the case study that we perform in electronic business domain, UN/CEFACT Core Component Technical Specification [89] is the major effort addressing the document interoperability in eBusiness. This effort tries to create a common basis for the artifacts of different document standards through the use of Core Components [90]. One major drawback is that Core Components are served through spreadsheets and they are not machine-processable. In addition, as described in [80], adoption of UN/CEFACT CCTS requires schema changes in the existing standards, which breaks backward compatibility. In the eBusiness case study, we apply the theory of common data elements maintained in Semantic Metadata Registries to the electronic business domain inline with UN/CEFACT CCTS methodology. We create machine-processable semantic descriptions of the Core Components as the eBusiness common data elements to the knowledgebase of a Semantic Metadata Registry and serve for the interoperating applications.

Translation of the electronic business documents is another major effort in the literature. Translating a document instance from one standard format into another one is generally achieved by means of XSL transformations [108] using schema matching techniques as described in [114]. It is important to note that this process is manual and needs manual updates in the XSL mappings upon any change in the document standards in question. Our solution eliminates the problem of document translation by introducing the extraction specifications. If an update is required to a document schema of a standard, the responsible standardization body needs to update the associated extraction specifications. This is an independent process from the other interoperating applications and has no effect on the implementations compared to the document translation techniques. Applications, by using the semantic web technologies, can interact with the Semantic Metadata Registry and perform automatic data extraction from different electronic document standards.

There are efforts which benefit from semantic web technologies for eBusiness document interoperability. Kabak [115] defined OWL ontologies for the three standards mentioned in the eBusiness case study of this thesis in addition to a CCTS upper ontology and defined mappings between these ontologies. Document instances then translated according to these semantic mappings. This is another top-down methodology which takes the UN/CEFACT CCTS into account for defining mappings. Biggest limitation of this method is that any change in the standards cause major updates in the ontologies and mappings among them like earlier top-down methodologies and XSL transformations. We eliminate this by focusing on automatic data extraction instead of whole document transformation. Similar efforts [116, 117] exist in the literature all aiming to achieve document translations for the document interoperability in different levels and suffer from similar limitations.

CHAPTER 7

CONCLUSION

In this thesis, we introduce a federated semantic metadata registry framework where machine processable definitions of the common data elements (CDEs) across domains can be shared, reused and semantically interlinked with each other through Linked Data principles. We demonstrate how such a framework can be utilized to address the semantic interoperability challenge across application domains.

There are already several adoptions of metadata registry/repository (MDR) systems [33, 31, 35, 48, 49, 50]. Some of them are maintained in a single organization like Roche GDSR [48], some are at project level in a specific domain like caDSR [33], some are at national level for eHealth domain like METeOR [4] in Australia, some are at national level but not restricted to a specific domain like NIEM US Federal metadata registry [3], and some are at global level like CDISC SHARE [110] addressing data interoperability across domains but covering a selected set of data sets. On top of these, there are efforts to define core set of data elements through spreadsheets, PDF documents or UML models like HITSP C154 [12] and FHIM [15] respectively. Our approach is not a disruptive effort; instead it builds upon and complements all of these as follows: through a semantically linked federated MDR framework, we believe these efforts can be linked with each other, multiplying their potential for semantic interoperability to a greater extent.

In the context of the SALUS project, Post Marketing Safety Study Tool (PMSST) and all related components have been implemented and deployed on

top of the SALUS system on the central EHR database of the Lombardy Region, Italy. This regional database includes electronic health records of ∼16 million patients with over 10 years longitudinal data in the average. Clinical researches in Roche are validating PMSST by using it in real life use cases, one of which is presented in Section 4.3. Figure 7.1 shows a part of the SAS analysis result that we produced on top of a simulated data set. Until the real deployment within SALUS project, we worked with the simulated data to collect further requirements from clinical researchers and improve the capabilities of PMSST.

The automatic importers for various content models including SDTM, SALUS and HITCP C154 led to the systematic and semantic representation of these models as common data elements formulated within the Semantic MDR. The mappings between these content models set up the CDE based interoperability architecture which is different from the majority of the existing efforts. This approach enables a very loosely coupled system addressing the heterogeneity problem among common data models for clinical researchers who work on EHR data for postmarketing surveillance studies.

PMSST enables clinical researches to define result models on which the postmarketing safety studies will be conducted without being aware of the structure of the underlying data sources. The main benefit of the utilization of CDE based interoperability architecture is the ability of developing surveillance methods which do not have to be restricted to the data model of the EHR source. Moreover, all CDE based interactions with the ISO/IEC 11179 based Semantic MDR has been implemented through the IHE DEX profile [75]. This increases the level of interoperability while promoting the potential in CDE sharing between different application domains.

In electronic business domain, we present that there are numerous different standards and having different standards is inevitable since no standard can contain all of the data needed in every environment. Each standard has its own document schema and the standards continue to evolve in their own directions. On the other hand, we know that industries are reluctant to change and the

Covariance Matrix for the Log-Rank Statistics				
ageband	18-49	50-59	60-69	70-79
18-49	4.43041	-2.05019	-2.15323	-0.22699
50-59	-2.05019	3.97490	-1.73745	-0.18726
60-69	-2.15323	-1.73745	4.10320	-0.21252
70-79	-0.22699	-0.18726	-0.21252	0.62677

Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	18.2039	3	0.0004

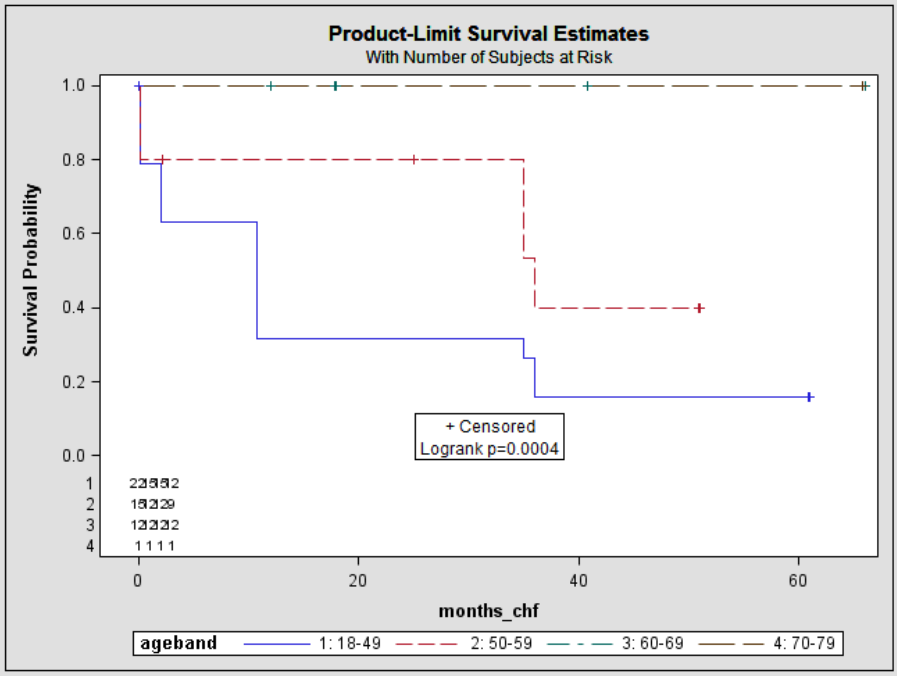


Figure 7.1: A part of the SAS result executed on the simulated data through PMSST

biggest reason is not to throw away the existing infrastructure investments. EDI is an example for this. Therefore, it is vital to build frameworks which can work with existing systems.

Semantic web technologies and Linked Data principles lead to undiscovered opportunities while building solutions for the electronic document interoperability. We believe that a federated Semantic Metadata Registry

framework can add the power of semantic web conveniently to the interoperability problem of electronic business documents as well. And, this is inline with the requirements of the eBusiness domain which allows different standards to evolve independently, but stay connected with the common data elements of other standards through Linked Data principles.

7.1 Discussion

Introduced interoperability framework does not depend on message translation or data transformation which are the dominant mechanisms of data interoperability approaches in the literature, either top-down or bottom-up. eHealth domain includes several implementations of such data interoperability frameworks such as Mini-Sentinel, OMOP or i2B2. They all exhibit a common information model, and data sources (i.e. EHR systems) implement adaptors in order to transform data into the dictated information models. These interoperability approaches suffer from the fact that requirements of the exhibited common information models change in time and this results in development efforts of the adaptors for each data source. Moreover, adding new data sources to such interoperability frameworks requires to implement new adaptors. Using abstract CDEs and extraction specifications integrated with the Linked Data principles eliminates those limitations of the available approaches. Two main aspects with respect to which we can make a comparison with the available data interoperability frameworks are:

- **Easy integration**: The vision behind the federated semantic metadata registry approach is that a Semantic MDR can be plugged into the linked datasets within the Linked Data cloud similar to the easy integration of the Bioportal services, thanks to the semantic web technologies. Compared to developing different adapters for each interoperating system, available data interoperability approaches do not exhibit a comparable easy integration mechanism.

- **Adaptability**: Any specification is subject to change in time because of

the ever-changing requirements and advancing technology. An adaptable interoperability framework can adapt itself to the changes in the data requirements easily. For the available frameworks, this means updates on the data adaptors. However; for the introduced data interoperability framework, the available implementation always continues working and the updates can be reflected to the framework with independent touches such as updating the links between the common data elements or creating/updating the extraction specifications based on the changes on the target content models. Hence, we claim that our model is more adaptable than the available interoperability approaches in the literature.

7.1.1 Limitations

We have a visionary objective about the use of the Semantic MDRs for data interoperability within the Linked Data cloud. Apart from validating the correctness and applicability which is shown with the PMSST implementation; we claim that in our data interoperability framework

- systems can be easily plugged/unplugged without affecting the rest of the system,
- there are well-established semantic links between data elements and
- the required extraction specifications are defined.

However; the main limitation to achieve this goal is the expectation that the coordinated approach that we propose is going to be adopted by the large-enough data element dictionaries and associated standardization initiatives. Current trend is to publish text-based documents to describe the data standards which need to be changed so that the specification can also be machine-processable.

Another limitation is that although we provide automatic content model importers; for proprietary information models, a manual effort is required to define the corresponding extraction specifications for the common data elements. For the commonly used standards such as HL7 CCD, the extraction

specifications can be imported to the knowledge base with the associated content model importer, however for the models like SALUS Common Information Model or Mini-Sentinel or OMOP, extraction specifications should be written manually if they are not already available. This limitation can be overcome with the proposed coordinated approach in which we say that the CDEs should stay as abstract definitions (i.e. metadata) and should be bound to the implementation with the extraction specifications.

7.2 Future Work

Validation activities of the SALUS project on real EHR databases is the means of the evaluation of the introduced interoperability framework and the Post Marketing Safety Study Tool (PMSST) developed on top of that. PMSST is addressing a specific use case for the diabetic patients experienced a type of acute coronary syndrome. As a future work, we plan to implement other use cases within the postmarketing surveillance studies of the pharmaceutical companies. Being a partner in the SALUS consortium, Roche provides real world use cases for this purpose. On the other hand, our theory on the advancement of the available ISO/IEC 11179 standard is in parallel to the development of the 3rd edition of the ISO/IEC 11179 standard which is trying to cover semantic web technologies with new constructs in the metamodel. Improvements on the Semantic MDR implementation according to the updated metamodel of the ISO/IEC 11179 standard 3rd edition is a planned future work in terms of tool development.

REFERENCES

[1] G. B. Laleci, M. Yuksel, and A. Dogac, "Providing semantic interoperability between clinical care and clinical research domains," *Biomedical and Health Informatics, IEEE Journal of*, vol. 17, no. 2, pp. 356–369, 2013.

[2] SALUS: Scalable, Standard based Interoperability Framework for Sustainable Proactive Post Market Safety Studies. [Online]. Available: http://www.salusproject.eu Last visited on July 2014.

[3] National Information Exchange Model. [Online]. Available: http://www.niem.gov Last visited on July 2014.

[4] Metadata Online Registry. Australian Institute of Health and Welfare. [Online]. Available: http://meteor.aihw.gov.au Last visited on July 2014.

[5] *ISO/IEC 11179: Information Technology – Metadata Registries (MDR) Parts 1–6 (2nd edition)*, International Organization for Standardization (ISO) / International Electrotechnical Commission (IEC) Std.

[6] Simple Knowledge Organization System (SKOS). World Wide Web Consortium (W3C). [Online]. Available: http://www.w3.org/2004/02/skos/ Last visited on July 2014.

[7] Friend of a Friend (FOAF). The Friend of a Friend Project. [Online]. Available: http://www.foaf-project.org/ Last visited on July 2014.

[8] schema.org. [Online]. Available: http://schema.org/docs/meddocs.html Last visited on July 2014.

[9] A. A. Sinaci and G. B. Laleci Erturkmen, "A federated semantic metadata registry framework for enabling interoperability across clinical research and care domains," *Journal of biomedical informatics*, vol. 46, no. 5, pp. 784–794, 2013.

[10] *XML Path Language (XPath)*, World Wide Web Consortium (W3C) Std. [Online]. Available: http://www.w3.org/TR/xpath/ Last visited on July 2014.

[11] *XML Schema*, World Wide Web Consortium (W3C) Std. [Online]. Available: http://www.w3.org/XML/Schema Last visited on July 2014.

[12] C 154: HITSP Data Dictionary. Healthcare Information Technology Standards Panel (HITSP). [Online]. Available: http://www.hitsp.org/ConstructSet_Details.aspx?&PrefixAlpha=4&PrefixNumeric=154 Last visited on July 2014.

[13] C 32: HITSP Summary Documents Using HL7 Continuity of Care Document (CCD) Component. Healthcare Information Technology Standards Panel (HITSP). [Online]. Available: http://www.hitsp.org/ConstructSet_Details.aspx?&PrefixAlpha=4&PrefixNumeric=32 Last visited on July 2014.

[14] Continuity of Care Document (CCD). Health Level 7 (HL7) / American Society for Testing and Materials (ASTM). [Online]. Available: http://www.hl7.org/documentcenter/public_temp_DC68F8CB-1C23-BA17-0CB6B9727B87B502/pressreleases/20070212.pdf Last visited on July 2014.

[15] Federal Health Information Model (FHIM). Federal Health Interoperability Modeling Initiative. [Online]. Available: http://www.fhims.org/content/420A62FD03B6_root.html Last visited on July 2014.

[16] Transitions of Care Initiative (ToC). Standards & Interoperability Framework. [Online]. Available: http://wiki.siframework.org/Transitions+of+Care+(ToC)+Initiative Last visited on July 2014.

[17] Clinical Element Data Dictionary (CEDD). Standards & Interoperability Framework. [Online]. Available: http://wiki.siframework.org/S%26I+Clinical+Element+Data+Dictionary+WG Last visited on July 2014.

[18] Query Health. Standards & Interoperability Framework. [Online]. Available: http://wiki.siframework.org/Query+Health Last visited on July 2014.

[19] Study Data Tabulation Model (SDTM). Clinical Data Interchange Standards Consortium (CDISC). [Online]. Available: http://www.cdisc.org/sdtm Last visited on July 2014.

[20] Clinical Data Acquisition Standards Harmonization (CDASH). Clinical Data Interchange Standards Consortium (CDISC). [Online]. Available: http://www.cdisc.org/cdash Last visited on July 2014.

[21] BRIDG Model. The Biomedical Research Integrated Domain Group (BRIDG). [Online]. Available: http://www.bridgmodel.org Last visited on July 2014.

[22] Reference Information Model (RIM). Health Level 7 (HL7). [Online]. Available: http://www.hl7.org/implement/standards/rim.cfm Last visited on July 2014.

[23] Common Data Model. Observational Medical Outcomes Project (OMOP). [Online]. Available: http://omop.org/ Last visited on July 2014.

[24] Sentinel Initiative – Mini-Sentinel. US Food and Drug Administration (FDA). [Online]. Available: http://mini-sentinel.org/ Last visited on July 2014.

[25] Clinical Element Models (CEM). GE/Intermountain Healthcare. [Online]. Available: http://www.clinicalelement.com/ Last visited on July 2014.

[26] Detailed Clinical Models. The National E-Health Transition Authority (NEHTA). [Online]. Available: http://www.nehta.gov.au/our-work/clinical-documents Last visited on July 2014.

[27] i2b2 Star Schema. Informatics for Integrating Biology and the Bedside (i2b2). [Online]. Available: http://www.i2b2.org Last visited on July 2014.

[28] S. Ngouongo, M. Löbe, and J. Stausberg, "The iso/iec 11179 norm for metadata registries: Does it cover healthcare standards in empirical research?" *Journal of biomedical informatics*, vol. 46, no. 2, pp. 318–327, 2013.

[29] Metadata Standards. ISO/IEC JTC1 SC32 WG2. [Online]. Available: http://metadata-standards.org/ Last visited on July 2014.

[30] CIHI Data Dictionary. Canadian Institute for Health Information. [Online]. Available: http://www.cihi.ca/CIHI-ext-portal/internet/EN/TabbedContent/standards+and+data+submission/standards/data+architecture/cihi010692 Last visited on July 2014.

[31] Cancer Grid Metadata Registry. UK Cancer Grid. [Online]. Available: http://www.cancergrid.eu/ Last visited on July 2014.

[32] P. M. Nadkarni and C. A. Brandt, "The common data elements for cancer research: remarks on functions and structure," *Methods of information in medicine*, vol. 45, no. 6, p. 594, 2006.

[33] G. A. Komatsoulis, D. B. Warzel, F. W. Hartel, K. Shanbhag, R. Chilukuri, G. Fragoso, S. d. Coronado, D. M. Reeves, J. B. Hadfield, C. Ludet, *et al.*, "cacore version 3: Implementation of a model driven, service-oriented architecture for semantic interoperability," *Journal of biomedical informatics*, vol. 41, no. 1, pp. 106–123, 2008.

[34] Environmental Data Registry. US Environmental Protection Agency. [Online]. Available: http://www.epa.gov/edr/ Last visited on July 2014.

[35] The United States Health Information Knowledgebase (USHIK). Agency for Healthcare Research and Quality (AHRQ). [Online]. Available: http://ushik.ahrq.gov/ Last visited on July 2014.

[36] Global Justice XML Data Model (GJXDM). US Department of Justice. [Online]. Available: http://www.it.ojp.gov/topic.jsp?topic_id=43 Last visited on July 2014.

[37] Report on Pharmacovigilance. The European Parliament. [Online]. Available: http://www.europarl.europa.eu/sides/getDoc.do?type=REPORT&reference=A7-2010-0159&language=EN Last visited on July 2014.

[38] Quality, Research and Public Health (QRPH). Integrating the Healthcare Enterprise (IHE). [Online]. Available: http://wiki.ihe.net/index.php?title=Quality,_Research_and_Public_Health Last visited on July 2014.

[39] *Simple Object Access Protocol (SOAP), organization=World Wide Web Consortium (W3C), url=http://www.w3.org/TR/soap/, note=last visited on July 2014*, Std.

[40] *Web Services Description Language (WSDL), organization=World Wide Web Consortium (W3C), url=http://www.w3.org/TR/wsdl, note=last visited on July 2014*, Std.

[41] T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web," *Scientific American*, vol. 284, no. 5, pp. 34–43, May 2001.

[42] N. Shadbolt, T. Berners-Lee, and W. Hall, "The Semantic Web Revisited," *IEEE Intelligent Systems*, vol. 21, no. 3, pp. 96–101, May-Jun 2006.

[43] Linked Data - Connect Distributed Data across the Web. [Online]. Available: http://linkeddata.org/ Last visited on July 2014.

[44] *Resource Description Framework (RDF)*, World Wide Web Consortium (W3C) Std. [Online]. Available: http://www.w3.org/RDF/ Last visited on July 2014.

[45] *Web Ontology Language (OWL)*, World Wide Web Consortium (W3C) Std. [Online]. Available: http://www.w3.org/2001/sw/wiki/OWL/ Last visited on July 2014.

[46] *SPARQL Query Language for RDF*, World Wide Web Consortium (W3C) Std. [Online]. Available: http://www.w3.org/TR/rdf-sparql-query/ Last visited on July 2014.

[47] Linking Open Data Community Project. World Wide Web Consortium (W3C). [Online]. Available: http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData Last visited on July 2014.

[48] F. K and M. F, "Semantic models for cdisc based standard and metadata management," in *CDISC Interchange Europe, Brussels*, 2012.

[49] S. J *et al.*, "A national metadata repository for empirical research in germany," in *15th International Open Forum on Metadata Registries, Berlin*, 2012.

[50] R. B and H. P, "Implementation of an iso/iec 11179 based metadata registry to foster interoperability of health telematics applications," in *15th International Open Forum on Metadata Registries, Berlin*, 2012.

[51] P. L. Whetzel, N. F. Noy, N. H. Shah, P. R. Alexander, C. Nyulas, T. Tudorache, and M. A. Musen, "Bioportal: enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications," *Nucleic acids research*, vol. 39, no. suppl 2, pp. W541–W545, 2011.

[52] OWL Ontology for the ISO/IEC 11179 Metamodel. [Online]. Available: https://github.com/srdc/semanticMDR/blob/master/core/src/main/resources/model/salus.mdr.owl Last visited on July 2014.

[53] Jena. Apache Software Foundation. [Online]. Available: http://jena.apache.org/ Last visited on July 2014.

[54] Jena TDB. Apache Software Foundation. [Online]. Available: http://jena.apache.org/documentation/tdb/ Last visited on July 2014.

[55] Virtuoso Universal Server. Open Link Software. [Online]. Available: http://virtuoso.openlinksw.com/ Last visited on July 2014.

[56] C. Bizer and A. Schultz, "The berlin sparql benchmark," *International Journal on Semantic Web and Information Systems (IJSWIS)*, vol. 5, no. 2, pp. 1–24, 2009.

[57] Open source Semantic MDR implementation. [Online]. Available: https://github.com/srdc/semanticMDR Last visited on July 2014.

[58] Pharmacovigilance. World Health Organization (WHO). [Online]. Available: http://www.who.int/medicines/areas/quality_safety/safety_efficacy/pharmvigi/en/ Last visited on July 2014.

[59] M. Hauben, V. Patadia, C. Gerrits, L. Walsh, and L. Reich, "Data mining in pharmacovigilance," *Drug safety*, vol. 28, no. 10, pp. 835–842, 2005.

[60] G. N. Norén, R. Orre, A. Bate, and I. R. Edwards, "Duplicate detection in adverse drug reaction surveillance," *Data Mining and Knowledge Discovery*, vol. 14, no. 3, pp. 305–328, 2007.

[61] M. Suling and I. Pigeot, "Signal detection and monitoring based on longitudinal healthcare data," *Pharmaceutics*, vol. 4, no. 4, pp. 607–640, 2012.

[62] J. C. Nelson, A. J. Cook, O. Yu, S. Zhao, L. A. Jackson, and B. M. Psaty, "Methods for observational post-licensure medical product safety surveillance," *Statistical methods in medical research*, p. 0962280211413452, 2011.

[63] B. L. Strom, S. E. Kimmel, and S. Hennessy, "The future of pharmacoepidemiology," *Textbook of Pharmacoepidemiology*, pp. 447–454, 2013.

[64] R. Platt, M. Wilson, K. A. Chan, J. S. Benner, J. Marchibroda, and M. McClellan, "The new sentinel network—improving the evidence of medical-product safety," *New England Journal of Medicine*, vol. 361, no. 7, pp. 645–647, 2009.

[65] M. Hauben and A. Bate, "Decision support methods for the detection of adverse events in post-marketing data," *Drug discovery today*, vol. 14, no. 7, pp. 343–357, 2009.

[66] P. Coorevits, M. Sundgren, G. O. Klein, A. Bahr, B. Claerhout, C. Daniel, M. Dugas, D. Dupont, A. Schmidt, P. Singleton, *et al.*, "Electronic health records: new opportunities for clinical research," *Journal of internal medicine*, vol. 274, no. 6, pp. 547–560, 2013.

[67] R. E. Behrman, J. S. Benner, J. S. Brown, M. McClellan, J. Woodcock, and R. Platt, "Developing the sentinel system—a national resource for evidence development," *New England Journal of Medicine*, vol. 364, no. 6, pp. 498–499, 2011.

[68] M. A. Robb, J. A. Racoosin, R. E. Sherman, T. P. Gross, R. Ball, M. E. Reichman, K. Midthun, and J. Woodcock, "The us food and drug administration's sentinel initiative: expanding the horizons of medical product safety," *Pharmacoepidemiology and drug safety*, vol. 21, no. S1, pp. 9–11, 2012.

[69] Observational Medical Outcomes Project (OMOP). Foundation for National Institutes of Health. [Online]. Available: http://omop.org/ Last visited on July 2014.

[70] Clinical Practice Research Datalink (CPRD). National Institute for Health Research. [Online]. Available: http://www.cprd.com/ Last visited on July 2014.

[71] The Health Improvement Network (THIN). University College London. [Online]. Available: http://www.ucl.ac.uk/pcph/research-groups-themes/thin-pub Last visited on July 2014.

[72] J. S. Brown, J. H. Holmes, K. Shah, K. Hall, R. Lazarus, and R. Platt, "Distributed health data networks: a practical and preferred approach to multi-institutional evaluations of comparative effectiveness, safety, and quality of care," *Medical care*, vol. 48, no. 6, pp. S45–S51, 2010.

[73] TRANSFoRm: Translational Research and Patient Safety in Europe. [Online]. Available: http://www.transformproject.eu Last visited on July 2014.

[74] EHR4CR: Electronic Health Record Systems for Clinical Research. [Online]. Available: http://www.ehr4cr.eu Last visited on July 2014.

[75] *Data Element Exchange (DEX) Profile*, Integrating the Healthcare Enterprise (IHE) Std. [Online]. Available: http://www.ihe.net/uploadedFiles/Documents/QRPH/IHE_QRPH_Suppl_DEX.pdf Last visited on July 2014.

[76] (2000-2004) SAS 9.1.3 Help and Documentation. SAS Institute Inc.

[77] Medical Dictionary for Regulatory Activities (MedDRA). International Conference on Harmonisation (ICH). [Online]. Available: http://www.meddra.org/ Last visited on July 2014.

[78] J. G. Klann, M. D. Buck, J. Brown, M. Hadley, R. Elmore, G. M. Weber, and S. N. Murphy, "Query health: standards-based, cross-platform population health surveillance," *Journal of the American Medical Informatics Association*, pp. amiajnl–2014, 2014.

[79] F. Lampathaki, S. Mouzakitis, G. Gionis, Y. Charalabidis, and D. Askounis, "Business to business interoperability: A current review of xml data integration standards," *Computer Standards & Interfaces*, vol. 31, no. 6, pp. 1045–1055, 2009.

[80] Y. Kabak and A. Dogac, "A survey and analysis of electronic business document standards," *ACM Computing Surveys (CSUR)*, vol. 42, no. 3, p. 11, 2010.

[81] *Electronic Data Interchange (EDI)*, US National Institute of Standards and Technology (NIST) Std.

[82] Automotive Industry Action Group. [Online]. Available: http://www.aiag.org/ Last visited on July 2014.

[83] Health Level 7 (HL7). [Online]. Available: http://www.hl7.org/ Last visited on July 2014.

[84] Petroleum Industry Data Exchange (PIDX). [Online]. Available: http://www.pidx.org/ Last visited on July 2014.

[85] Chemical Industry Data Exchange (CIDX). [Online]. Available: http://www.cidx.org/ Last visited on July 2014.

[86] OpenTravel Alliance (OTA). [Online]. Available: http://www.opentravel.org/ Last visited on July 2014.

[87] RosettaNet. [Online]. Available: http://www.rosettanet.org/ Last visited on July 2014.

[88] *Core Components Technical Specification (CCTS)*, UN Centre for Trade Facilitation and E-business (UN/CEFACT) Std. [Online]. Available: http://www.unece.org/cefact/codesfortrade/CCTS/CCTS-Version3.pdf Last visited on July 2014.

[89] Core Component Library. United Nations. [Online]. Available: http://www.unece.org/fileadmin/DAM/cefact/codesfortrade/unccl/CCL13A.zip Last visited on July 2014.

[90] *Universal Business Language (UBL)*, Advancing Open Standards for Information Society (OASIS) Std. [Online]. Available: https://www.oasis-open.org/committees/ubl Last visited on July 2014.

[91] *OAGIS Business Object Document (BOD)*, Open Applications Group Std. [Online]. Available: http://www.oagi.org/oagis/9.0/Documentation/Architecture.html Last visited on July 2014.

[92] *GS1 XML*, Global Standards One (GS1) Std. [Online]. Available: http://www.gs1.org/ecom/xml Last visited on July 2014.

[93] *GS1 Global Data Dictionary*, Global Standards One (GS1) Std. [Online]. Available: http://gddold.gs1.org/gdd/public/default.asp Last visited on July 2014.

[94] Common Information Model for Energy Markets. European Network of Transmission System Operators for Electricity. [Online]. Available: https://www.entsoe.eu/major-projects/common-information-model-cim/cim-for-energy-markets/ Last visited on July 2014.

[95] Operational Data Model. Clinical Data Interchange Standards Consortium (CDISC). [Online]. Available: www.cdisc.org/odm Last visited on July 2014.

[96] S. J. Reisinger, P. B. Ryan, D. J. O'Hara, G. E. Powell, J. L. Painter, E. N. Pattishall, and J. A. Morris, "Development and evaluation of a common data model enabling active drug safety surveillance using disparate healthcare databases," *Journal of the American Medical Informatics Association*, vol. 17, no. 6, pp. 652–662, 2010.

[97] L. H. Curtis, M. G. Weiner, D. M. Boudreau, W. O. Cooper, G. W. Daniel, V. P. Nair, M. A. Raebel, N. U. Beaulieu, R. Rosofsky, T. S. Woodworth, *et al.*, "Design considerations, architecture, and use of the mini-sentinel distributed data system," *Pharmacoepidemiology and drug safety*, vol. 21, no. S1, pp. 23–31, 2012.

[98] I. S. Kohane, S. E. Churchill, and S. N. Murphy, "A translational engine at the national scale: informatics for integrating biology and the bedside," *Journal of the American Medical Informatics Association*, vol. 19, no. 2, pp. 181–185, 2012.

[99] H. J. Lowe, T. A. Ferris, P. M. Hernandez, S. C. Weber, *et al.*, "Stride–an integrated standards-based translational research informatics platform," in *AMIA Annu Symp Proc*, vol. 14, 2009, pp. 391–395.

[100] P. Avillach, J.-C. Dufour, G. Diallo, F. Salvo, M. Joubert, F. Thiessard, F. Mougin, G. Trifirò, A. Fourrier-Réglat, A. Pariente, *et al.*, "Design and validation of an automated method to detect known adverse drug reactions in medline: a contribution from the eu–adr project," *Journal of the American Medical Informatics Association*, vol. 20, no. 3, pp. 446–452, 2013.

[101] G. M. Weber, S. N. Murphy, A. J. McMurry, D. MacFadden, D. J. Nigrin, S. Churchill, and 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*, vol. 16, no. 5, pp. 624–630, 2009.

[102] A. El Fadly, B. Rance, N. Lucas, C. Mead, G. Chatellier, P.-Y. Lastic, M.-C. Jaulent, and C. Daniel, "Integrating clinical research with the healthcare enterprise: from the re-use project to the ehr4cr platform," *Journal of biomedical informatics*, vol. 44, pp. S94–S102, 2011.

[103] G. Jiang, H. R. Solbrig, D. Iberson-Hurst, R. D. Kush, and C. G. Chute, "A collaborative framework for representation and harmonization of clinical study data elements using semantic mediawiki," *AMIA Summits on Translational Science Proceedings*, vol. 2010, p. 11, 2010.

[104] I. Kunz, M.-C. Lin, and L. Frey, "Metadata mapping and reuse in cabig^{TM}," *BMC bioinformatics*, vol. 10, no. Suppl 2, p. S4, 2009.

[105] J. Pathak, J. Wang, S. Kashyap, M. Basford, R. Li, D. R. Masys, and C. G. Chute, "Mapping clinical phenotype data elements to standardized metadata repositories and controlled terminologies: the emerge network experience," *Journal of the American Medical Informatics Association*, vol. 18, no. 4, pp. 376–386, 2011.

[106] *Drug Safety Content Profile (DSC)*, Integrating the Healthcare Enterprise (IHE) Std. [Online]. Available: http://www.ihe.net/Technical_Framework/upload/IHE_QRPH_TF_Supplement_Drug_Safety_Content_DSC_TI_2009-08-10.pdf Last visited on July 2014.

[107] *Clinical Research Data Capture Profile (CRD)*, Integrating the Healthcare Enterprise (IHE) Std. [Online]. Available: http://wiki.ihe.net/index.php?title=Clinical_Research_Data_Capture_-_(CRD) Last visited on July 2014.

[108] *The Extensible Stylesheet Language Family (XSL)*, World Wide Web Consortium (W3C) Std. [Online]. Available: http://www.w3.org/Style/XSL/ Last visited on July 2014.

[109] B. L, E. J, and L. PY, "Mapping ehr data to a research case report form: How a metadata repository, cdisc's share, can improve the ihe profile clinical research data (crd)," in *15th International Open Forum on Metadata Registries, Berlin*, 2012.

[110] CDISC SHARE. Clinical Data Interchange Standards Consortium (CDISC). [Online]. Available: http://www.cdisc.org/cdisc-share Last visited on July 2014.

[111] C. Tao, G. Jiang, W. Wei, H. R. Solbrig, and C. G. Chute, "Towards semantic-web based representation and harmonization of standard metadata models for clinical studies," *AMIA Summits on Translational Science Proceedings*, vol. 2011, p. 59, 2011.

[112] G. Shukair, N. Loutas, V. Peristeras, and S. Sklarß, "Towards semantically interoperable metadata repositories: The asset description metadata schema," *Computers in Industry*, vol. 64, no. 1, pp. 10–18, 2013.

[113] R. T. Chen, J. W. Glasser, P. H. Rhodes, R. L. Davis, W. E. Barlow, R. S. Thompson, J. P. Mullooly, S. B. Black, H. R. Shinefield, C. M. Vadheim, *et al.*, "Vaccine safety datalink project: a new tool for improving vaccine safety monitoring in the united states," *Pediatrics*, vol. 99, no. 6, pp. 765–773, 1997.

[114] E. Rahm and P. A. Bernstein, "A survey of approaches to automatic schema matching," *the VLDB Journal*, vol. 10, no. 4, pp. 334–350, 2001.

[115] Y. KABAK, "Semantic interoperability of the un/cefact ccts based electronic business document standards," 2009.

[116] N. Aničić and N. Ivezić, "Semantic web technologies for enterprise application integration," *Computer Science and Information Systems*, vol. 2, no. 1, pp. 119–144, 2005.

[117] Y. Yarimagan and A. Dogac, "A semantic-based solution for ubl schema interoperability," *Internet Computing, IEEE*, vol. 13, no. 3, pp. 64–71, 2009.

CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name: Sınacı, Ali Anıl
Nationality: Turkish (TC)
Date and Place of Birth: 2 February 1985, Diyarbakır
Marital Status: Married
Phone: +90 312 210 1763
Emil: anil@ceng.metu.edu.tr

EDUCATION

Degree	Institution	Year of Graduation
M.S.	Computer Engineering Department, METU	2009
B.S.	Computer Engineering Department, METU	2007
High School	İçel Anatolian High School	2003

PROFESSIONAL EXPERIENCE

Year	Place	Enrollment
2011-present	SRDC Yazılım Araştırma Geliştirme ve Danışmanlık Ltd. Şti.	Senior Researcher / Software Engineer
2010-2011	CENGSOFT Yazılım Araştırma Geliş. Eğitim ve Danışmanlık Ltd. Şti.	Consultant
2009-2010	AGMLab Bilişim Teknolojileri Ltd. Şti.	Consultant
2009-2011	Computer Engineering Department, METU	Teaching/Research Assistant
2007-2009	Software Research and Development Center, METU	Researcher / Software Engineer
2006-2007	Software Research and Development Center, METU	Part-time Software Developer

PUBLICATIONS

1. A. A. Sinaci, G. B. Laleci, S. Gonul, B. Thakrar, H. A. Cinar, N. K. Cicekli, "Post Marketing Safety Study Tool: A web based, dynamic and interoperable system for postmarketing drug surveillance studies", *Computer Methods and Programs in Biomedicine*, submitted for publication.

2. A. A. Sinaci, G. B. Laleci, A. Pacaci, Y. Kabak, N. K. Cicekli, A. Dogac, "Automatic Data Extraction from Electronic Business Documents with the use of Semantic Metadata Registries", *International Journal of Metadata, Semantics and Ontologies*, submitted for publication.

3. G. B. Laleci, L. Bain, A. A. Sinaci, F. Malfait, G. Low, "keyCRF: Using Semantic Metadata Registries to Populate an eCRF with EHR Data", *International Semantic Web Conference (ISWC 2014)*, submitted for publication.

4. S. Hussain, H. Sun, A. A. Sinaci, G. B. Laleci, C. Mead, A. J. G. Gray, D. L. McGuiness, E. Prud'hommeaux, C. Daniel, K. Forsberg, "A

Framework for Evaluating and Utilizing Medical Terminology Mappings", accepted as a full research paper in *European Medical Informatics Conference*, Istanbul, September 2014.

5. T. Krahn, M. Eichelberg, F. Muller, S. Gonul, G. B. Laleci, A. A. Sinaci, H. Jurgen-Appelrath, "Adverse Drug Event Notification on a Semantic Interoperability Framework", accepted as a full research paper in *European Medical Informatics Conference*, Istanbul, September 2014.

6. A. A. Sinaci, G. B. Laleci, A. Pacaci, "Clinical Research Data Collection from Medical Summaries through Semantic Metadata Registries", to be presented in *Workshop: The semantic interoperability challenge to exploit EHRs for enabling better care,clinical research and public health studies. European Medical Informatics Conference*, Istanbul, September 2014.

7. C. Daniel, A. A. Sinaci, D. Ouagne, E. Sadou, G. Declerck, D. Kalra, J. Charlet, K. Forsberg, L. Bain, C. Mead, S. Hussain, G. B. Laleci, "Standard-based EHR-enabled applications for clinical research and patient safety: CDISC – IHE QRPH – EHR4CR and SALUS collaboration" in *AMIA 2014 Joint Summits on Translational Science*, San Francisco, April 2014.

8. M. Yuksel, S. Gonul, G. B. Laleci, A. A. Sinaci, K. Depraetere, J. De Roo, T. Bergval, "Demonstration of the SALUS Semantic Interoperability Framework for Case Series Characterization Studies" in *International SWAT4LS Workshop*, Edinburgh, December 2013.

9. A. A. Sinaci, G. B. Laleci, S. Gonul, H. A. Cinar, A. Kaya, "Patient History Navigation with the Use of Common Data Elements" in *International SWAT4LS Workshop*, Edinburgh, December 2013.

10. L. Bain, G. B. Laleci, C. Daniel, A. A. Sinaci, "Data Element Exchange (DEX) Profile", Trial Implementation as a part of *IHE Quality, Research and Public Health Domain (QRPH) Technical Framework*, September 2013.

11. A. A. Sinaci, G. B. Laleci, "A Federated Semantic Metadata Registry

Framework for Enabling Interoperability across Clinical Research and Care Domains", *Journal of Biomedical Informatics*, Vol. 46, no. 5, pp.784-794, June 2013.

12. C. Daniel, G. B. Laleci, A. A. Sinaci, B. C. Delaney, V. Curcin, L. Bain, "Standard-based integration profiles for clinical research and patient safety", in *AMIA 2013 Joint Summits on Clinical Research Informatics*, San Francisco, March 2013.

13. G. Declerck, S. Hussain, Y. Pares, C. Daniel, M. Yuksel, A. A. Sinaci, G. B. Laleci, and M. C. Jaulent, "Semantic-sensitive extraction of EHR data to support adverse drug event reporting," in *International SWAT4LS Workshop*, Paris, Nov 2012.

14. F. Gigante, P. Crespi, A. A. Sinaci, R. V. Basar, "Analysis of the Information Resources for the Furniture Industry in BIVEE" in *NGEBIS Workshop, CAiSE Conference*, Valencia, June 2013.

15. R. V. Basar, A. A. Sinaci, F. Smith, F. Taglino, "Semantic UBL-like documents for innovation" in *NGEBIS Workshop, CAiSE Conference*, Valencia, June 2013.

16. A. A. Sinaci, "Jena based Implementation of a ISO 11179 Metadata Registry" in *ApacheCon EU*, Sinsheim, November 2012.

17. S. Gonul and A. A. Sinaci, "Semantic Content Management and Integration with JCR/CMIS Compliant Content Repositories" in *I-Semantics Conference*, Graz, September 2012.

18. A. A. Sinaci, M. Piersantelli, C. Cristalli, F. Gigante, G. B. Laleci and R. V. Basar, "A Document Centric Approach for User Requirements in BIVEE" in *NGEBIS Workshop, CAiSE Conference*, Gdansk, June 2012.

19. G. B. Laleci, A. Dogac, M. Yuksel, S. Hussain, G. Declerck, C. Daniel, H. Sun, K. Depraetere, D. Colaert , J. Devlies, T. Krahn, B. Thakrar, G. Freriks, T.Bergvall, and A. A. Sinaci, "Building the Semantic Interoperability Architecture Enabling Sustainable Proactive Post

Market Safety Studies," in *SIMI Wokshop - European Semantic Web Conference (ESWC)*, Crete, May 2012.

20. A. A. Sinaci and S. Gonul, "Semantic Content Management with Apache Stanbol" in *European Semantic Web Conference (ESWC)*, Crete, May 2012.

21. G. B. Laleci, G. Aluc, A. Dogac, A. Sinaci, O. Kilic and F. Tuncer, "A Semantic Backend System to Support Content Management Systems", *Knowledge-Based Systems Journal*, Vol. 23, pp.832-843, December 2010.

22. A. A. Sinaci, O. T. Sehitoglu, M. T. Yondem, G. Fidan, and I. Tatli, "SEMbySEM in Action: Domain Name Registry Service Through a Semantic Middleware" in eChallenges Conference, Warsaw, October 2010.

23. A. Dogac, G. B. Laleci, G. Aluc, A. A. Sinaci, W. Behrendt, B. Delacretaz, J. M. Pittet, "A semantically Enriched Persistence Mechanism for Interactive Knowledge Stack" in *eChallenges Conference*, Istanbul, October 2009.

24. T. Namli, A. Dogac, A. A. Sinaci, G. Aluc, "Testing the Interoperability and Conformance of UBL/NES based Applications" in *eChallenges Conference*, Istanbul, October 2009.

25. T. Namli, G. Aluc G., A. Sinaci, I. Kose, N. Akpinar, M. Gurel, Y. Arslan, H. Ozer, N. Yurt, S. Kirici, E. Sabur, A. Ozcam, A. Dogac, "Testing the Conformance and Interoperability of NHIS to Turkey's HL7 Profile" in *9th International HL7 Interoperability Conference (IHIC)*, Crete, October 2008.