Integrating Terminologies and Services using Semantic Web Standards

By
Sidra Shahbaz
NUST201260772MSEECS60012F

Supervisor
Dr. Khalid Latif
Department of Computing

A thesis submitted in partial fulfillment of the requirements for the degree of Masters of Science in Information Technology (MS IT)

In
School of Electrical Engineering and Computer Science,
National University of Sciences and Technology (NUST),
Islamabad, Pakistan.

(July 2015)

Approval

It is certified that the contents and form of the thesis entitled "**Integrating Terminologies and Services using Semantic Web Standards**" submitted by **Sidra Shahbaz** have been found satisfactory for the requirement of the degree.

Advisor: **Dr. Khalid Latif**

Signature: ________________

Date: ________________

Committee Member 1: **Dr. Hamid Mukhtar**

Signature: ________________

Date: ________________

Committee Member 2: **Dr. Sarah Shafiq Khan**

Signature: ________________

Date: ________________

Committee Member 3: **Dr. Farooq Ahmad**

Signature: ________________

Date: ________________

Dedication

With affection and gratitude, I would like to dedicate this thesis to my parents and teachers who have been a continuous source of inspiration and motivation for me and supported me all the way throughout my education.

Certificate of Originality

I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any degree or diploma at NUST SEECS or at any other educational institute, except where due acknowledgement has been made in the thesis. Any contribution made to the research by others, with whom I have worked at NUST SEECS or elsewhere, is explicitly acknowledged in the thesis.

I also declare that the intellectual content of this thesis is the product of my own work, except for the assistance from others in the project's design and conception or in style, presentation and linguistics which has been acknowledged.

Author Name: **Sidra Shahbaz**

Signature: ____________________

Acknowledgment

In the name of Allah the beneficent and merciful, on whom we all are dependent for eventual support and guidance.

I would like to express my immense gratitude to my supervisor Dr. Khalid Latif, who has guided and supported me throughout my thesis work and allow me to work in my own way and polish my skills. I highly appreciate his help in technical writing. His mentorship was predominant in sustaining versatile experience in my long term career goals.

I would also like to thank my committee members, Dr. Hamid Mukhtar, Dr. Sarah Shafiq Khan and Dr. Farooq Ahmad who continuously guided me in my thesis, and provide their valuable suggestions and encouragement.

Finally I would like to thank my siblings, friends and colleagues who have helped me in order to complete my thesis.

Table of Contents

List of Tables

List of Figures

Abstract

Many healthcare terminology standards and classifications such as ICD, SNOMED, CPT have emerged in last few years and are widely used in various healthcare applications to encode and represent patient records. However, almost all terminologies are represented using different data models and involve incompatible formats. For consistent understanding of patient records, terminologies should be represented using a canonical format. This thesis proposes a semantic model, build using HL7-CTS2 and SKOS to bridge the gap between the terminologies. The semantic model accommodates heterogeneous medical terminologies as a federated terminology distribution service. A RESTful framework is also developed over the semantic model to process the terminological contents for a proof of concept. The service allows browsing and navigation for searching terminology contents in an integrated platform. Moreover, the proposed system is expected to enhance the semantic interoperability beyond the traditional health care applications.

Chapter 1

Introduction

This chapter gives the basic idea of the concepts involved in this research. It also presents the background and motivation for this study. Moreover, it provides an idea of expected results, and methodology to get and evaluate the results. Finally, it presents the structure of this thesis document.

1.1 Introduction

In the current healthcare industry, the demand of Electronic Health Record (EHR) systems has emerged to provide convenient access and improved healthcare. Computer based patient data needs to be represented in standard form to support applications of EHR such as patient charting, order entry, decision support, reporting, diagnosis and data aggregation [2] and let applications exchange information with other EHR systems. In this regard, many healthcare standards have emerged in last few years including HL7 [3], IHE [4], and ISO-25964 [5] to represent patient records. Most of these standards reckon medical terminologies such as Systemized Nomenclature of Medicine (SNOMED), International Classification of Diseases (ICD), Current Procedural Terminology (CPT) and Logical Observation Identifiers Names and Codes (LOINC), as the most critical building block for encoding medical or health data.

EHR systems strive for accurate, precise, and unambiguous communication which requires collective and consistent understanding of the medical terminologies used to describe health records. Terminologies are generally represented using different knowledge models. For consistent understanding, terminologies should be represented using a canonical format and should be accessible through a single platform. HL7 Common Terminology Service v2 (CTS2) provides a common platform to accommodate diverse structure of different medical terminologies [6] but very few terminologies have been completely mapped to CTS2 because of its complexity. Moreover, CTS2 is limited to a common structural model for terminology behavior specific to healthcare applications, independent from any specific terminology implementation. Apelon for instance provides Distributed Terminology Service (DTS) for the management, deployment and extension to health care terminologies by implementing CTS2 model [7]. Many terminology providers are opting for Semantic Web based knowledge representation standards such as RDF [8], OWL [9], or SKOS [1] for encoding terminology resources. For instance, ICD-11 has been implemented using SKOS [10] . This presents us with a need and challenge of implementing CTS2 services using Semantic Web standards, more specifically using SKOS. As a modeling approach based on RDF, SKOS supports integration of terminologies with data on the web (or inside an application), otherwise not directly supported by CTS2.

For integrating medical terminologies, a knowledge model is designed by following the semantic web standards and CTS2 model. Terminology providers can continue their use of semantic web or other knowledge representation standards such as Topic Maps but would additionally provide schema mappings in a canonical format. This approach integrates medical terminologies to a common semantic terminology structure. The semantic structure helps in reasoning and will allow CTS2 compliant systems to read terminology data in CTS2 supported format. Moreover, the proposed model is expected to enhance the semantic interoperability beyond the traditional health care applications.

1.1.1 CTS2

CTS, an acronym of Common Terminology Service version 2, is a joint project of HL7 and OMG [11]. The intent of CTS2 was to provide a generic yet a standard model for the management and usage of medical terminologies, independent of any specific implementation [6]. CTS2 offers common, modular and deployable behaviors, which contribute in dealing with set of available medical terminologies in different deployment environments. The service contributes in achieving interoperability of terminologies across different environments by specifying a standard interface for access and also by specifying the queries for subsumption and inference.

1.2 Motivation

CTS2 provides a common platform to accommodate diverse structure of different medical terminologies but very few terminologies have been completely mapped to CTS2 because of its complexity. Moreover, CTS2 is limited to a common structural model for terminology behavior specific to healthcare applications, independent from any specific terminology implementation. Many terminology providers, on the other hand, are opting for semantic web based knowledge representation standards such as RDF, OWL, or SKOS for encoding terminology resources. This presents us with a need and challenge of implementing CTS2 services using semantic web standards, more specifically using SKOS. As an application of RDF, SKOS supports integration of terminology with data on the web, otherwise not directly supported by CTS.

1.3 Objective

The main purpose of this research is the integration of standard medical terminologies using semantic web standards and CTS2. The CTS2 conceptual model is mapped to SKOS and then standard medical terminologies are mapped to the semantic model of CTS2. This approach integrates medical terminologies to a common semantic terminology structure. The semantic structure of CTS2 model helps CTS2 compliant system to retrieve terminological content in CTS2 supported format.Moreover, the proposed model is expected to enhance the semantic interoperability beyond the traditional health care applications.

1.4 Problem Statement

Standard medical terminologies are developed with specific purposes in mind, are structured very differently. The format of the terminology may range from a simple flat list of concepts, to more complex poly-hierarchies. CTS2 provide the way for accommodating hybrid terminologies but is limited to a common structural model for terminology behavior. The representation of standard medical terminologies either follows semantic web structure or some other structure require a unified data model, So the content could be accessed via one unified interface.

1.5 Contribution

For the integration of terminologies, currently there is no semantic model of CTS2. HL7 and OMG have only published the conceptual model of CTS2. Major contributions of this work include:

- A unified terminology model represented using semantic web standards and based on CTS2 conceptual model.
- Representation of selected standard medical terminologies to the unified terminology model.
- Translation of selected medical terminologies according to the unified model.
- A RESTful framework developed over the unified model to process the terminological contents.
- Integration of CTS2 services to Semantic Electronic Medical Record (SEMR) to demonstrate its application in medical domain.

1.6 Evaluation

The evaluation is performed based on three perspectives.

- CTS2 services conformance to query profiles defined in CTS2 specifications.
- To ensure the correctness of representation of medical terminologies to the semantic unified terminology model, reverse translation is performed.
- CTS2 service load time is evaluated for each terminology.
- CTS2 service performance and throughput is evaluated against number of users.

1.7 Methodology

As per proposed methodology, firstly the selected medical terminological content is parsed, mapped over the unified model, and is stored in to the semantic store. Secondly the RESTful services are developed to access the terminological content from one platform.

1.8 Expected Results

As per proposed methodology, the selected medical terminologies should be parse and mapped to UTS model and finally stored in the semantic data store. The correctness of stored data should be checked by the reverse translation process. After the correctness check, terminological content should be accessible via RESTful services.

1.9 Structure

Rest of the thesis is structured as follows:

- Chapter 2: Background and structure information of Standard Medical Terminologies.
- Chapter 3: Description of Common Terminology Service version 2.
- Chapter 4: Description of Knowledge Representation Standards.
- Chapter 5: Literature Review explaining information about existing solutions.
- Chapter 6: Description of the proposed unified model and its mapping with CTS2 model.
- Chapter 7: System Architecture shows the overall model of the system and flow of information.
- Chapter 8: Evaluation explains the consistency, conformance and performance of the CTS2 system.
- Chapter 9: provides results and conclusion.

Chapter 2

Standard Medical Terminologies

2.1 SNOMED-CT

The Systematized Nomenclature of Medicine (SNOMED) is a controlled and structured collection of medical terms [12]. It is a scientifically validated and comprehensive resource and is considered as a very essential component for EHR. It covers many health domains such as diseases, anatomy, procedures, findings, substances and microorganisms. SNOMED Clinical Terms (known as SNOMED-CT) was evolved in 1999, by the merging, restructuring and expansion of two large-scale known terminologies i.e. SNOMED Reference Terminology (SNOMED-RT) and Clinical Terms version 3 (CTV3). SNOMED-CT is now known as a standard terminology instead of SNOMED and it allows mapping of entities with other internationally accepted standard terminologies. Its basic structure includes:

1. **Concept Codes** are numerical codes used for clinical terms and are organized as hierarchy. There are 300,000 active concepts in SNOMED-CT.
2. Text based **descriptions** of Concept Codes.
3. **Relationships** between Concept Codes of same or different code systems.

The complexity of SNOMED-CT lies in, the way contents are structured. There are three core files i.e. Concepts, Descriptions and Associations provided by SNOMED-CT. The problem here is to link up all three files content

Table 2.1: SNOMED Concept Example

CONCEPT ID	**STATUS**	**FULLY SPECIFIED NAME**
149885003	0	Entire conjunctival vein (body structure)
101414002	10	PREDBIOTIC (product)

into one common structure. Table 2.1 reflects the structure of SNOMED Concept using an example.

Table 2.2 reflects the structure of SNOMED Concept descriptions using an example.

Table 2.2: SNOMED Concept Description Example

ID	**STATUS**	**CONCEPT ID**	**TERM**	**TYPE**	**LANG**
529603014	0	149885003	Entire conjunctival vein (body structure)	3	en
234135012	0	149885003	Entire vein	2	en
234134011	0	149885003	Entire conjunctival vein	1	en
164557015	8	101414002	PREDBIOTIC	1	en
529874017	0	101414002	PREDBIOTIC (substance)	3	en

2.2 LOINC

LOINC [13] was the initiative by Regenstrief Institute. LOINC codes are used to represent laboratory and other medical observations that facilitate pooling and exchange of observations. LOINC is of flat grid-style structure of concepts and properties. Accommodating specialized properties of LOINC entities in the common structure is a challenging task because of its variety and structure. Table 2.3 reflects the structure of LOINC using an example.

Table 2.3: LOINC Example

Name	**Code**	**Component**
Nordoxepin	3862-0	Nordoxepin:MCnc:Pt:SerPlas:Qn
Haloperidol	3669-9	Haloperidol:MCnc:Pt:SerPlas:Qn:

2.3 CPT

CPT is a medical code set maintained by American Medical Association (AMA). CPT coding scheme describes surgical, medical and diagnostic services for uniform communication of information about procedures and medical services among the coders, patients, organization and physicians [14]. CPT is of flat kind structure and has one core file containing all the concepts and properties. Table 2.4 reflects the structure of CPT using an example.

Table 2.4: CPT Example

Code	**Long Description**	**Short Description**
10021	FINE NEEDLE ASPIRATION; WITHOUT IMAGING GUIDANCE	FNA W/O IMAGE
10022	FINE NEEDLE ASPIRATION; WITH IMAGING GUIDANCE	FNA W/IMAGE

2.4 ICD

ICD was developed in Europe [15]. ICD structure follows XML standard. It contains information about disease codes, alphabetical index of disease entities and classification for therapeutic, surgical and diagnostic procedures. ICD has multiple revisions and localized variations such as ICD-10-CM (Clinical Modification) applicable in the US. These codes are widely used in Billing.

```
<Class code="K11.7" kind="category">
    <Meta name="MortBCode" value="182"/>
        <Meta name="MortL4Code" value="4-024"/>
        <Meta name="MortL3Code" value="3-035"/>
        <Meta name="MortL2Code" value="2-072"/>
```

```
        <Meta name="MortL1Code" value="1-081"/>
        <SuperClass code="K11"/>
        <Rubric id="D0008593" kind="preferred">
                <Label xml:lang="en" xml:space="default">
        Disturbances of salivary secretion</Label>
        </Rubric>
        <Rubric id="D0009151" kind="inclusion">
                <Label xml:lang="en" xml:space="default">
        Hypoptyalism</Label>
        </Rubric>
        <Rubric id="D0009152" kind="inclusion">
                <Label xml:lang="en" xml:space="default">
        Ptyalism</Label>
        </Rubric>
</Class>
```

2.5 DRG

The Diagnosis Related Group [16] is used for classification of hospital scenarios into one of the available 467 groups. This classification system was developed in Yale School of Public Health. Like CPT and LOINC, DRG is also of flat kind structure and has one core file containing all the concepts and properties. Table 2.5 reflects the structure of DRG using an example.

Table 2.5: DRG Example

Code	**MDC**	**Type**	**DRG Title**
1	Pre	SURG	HEART TRANSPLANT OR IMPLANT OF HEART ASSIST SYSTEM W MCC
52	Pre	MED	SPINAL DISORDERS & INJURIES W CC/MCC

Chapter 3

Common Terminology Service

Common terminology is a joint project of HL7 and OMG [11]. The intent of CTS version 2 (referred as CTS2) was to provide a generic yet a standard model for the management and usage of medical terminologies, independent of any specific implementation [6]. CTS2 offers common, modular and deployable behaviors, which contribute in dealing with set of available medical terminologies in different deployment environments (c.f. Figure 3.1). The service contributes in achieving interoperability of terminologies across different environments by specifying a standard interface for access and also by specifying the queries for subsumption and inference. Medical terminologies provide fine granularity of shared semantics. Terminologies are developed considering a specific purpose in mind such as LOINC that covers lab domain. The expressivity of terminologies varies from flat list of concepts to complex hierarchical structure. CTS2 specifies a basic concept-level terminology model for accommodating different structured terminologies.

3.1 CTS Entities

Some of the important entities of CTS are discussed below

3.1.1 CodeSystem

A code system is an entity to collect information about the medical terminologies. It contains the meta data information of the medical terminology and is described by unique concepts along with their designations, associations and valuesets.

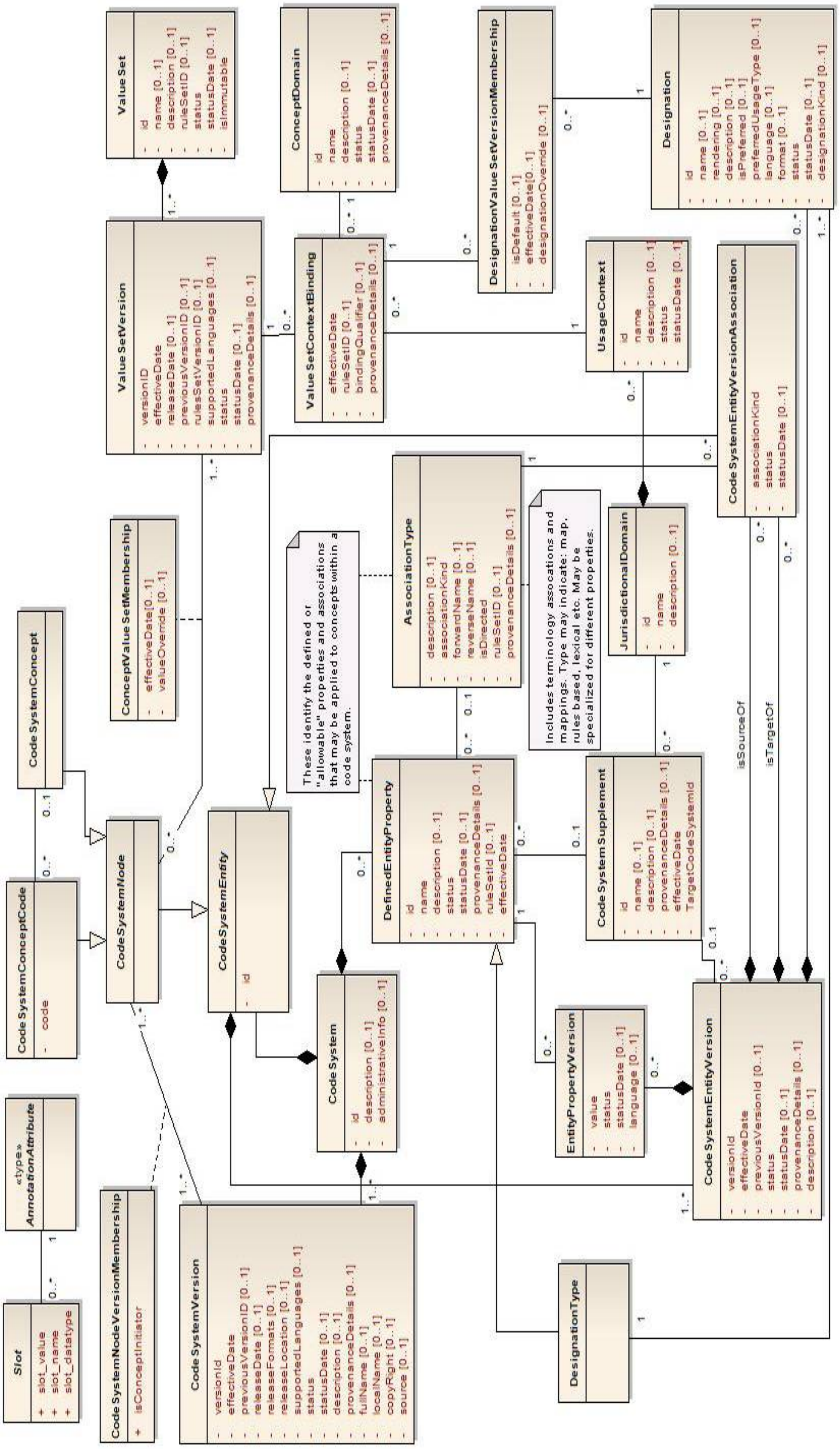

Figure 3.1: CTS 2 Conceptual Model

3.1.2 CodeSystemEntity

A code system entity class represents either a CodeSystemNode (node) or an CodeSystemEntityVersionAssociation (association) within the whole terminological content.

3.1.3 CodeSystemNode

A code system node is a individual node in a Code System. It represents either a CodeSystemConcept or a CodeSystemConceptCode.

3.1.4 CodeSystemConcept

A code system concept represents a abstract thing in context of a single Code System. It contains the meta data information about the concept. Each concept is unique within a code system and a Code System contains list of concepts. An example of code system concept is ”fever”, which is medical concept of SNOMED terminology.

3.1.5 CodeSystemEntityVersionAssociation

A code system entity version association represents and define relationships between the concepts within the code system. For example, the concept ”lung consolidation” of SNOMED-CT has an ”is-a” relationship with the concept ”disorder of lung”.

3.1.6 AssociationType

A association type class represents allowable association types of the medical terminologies. Its not part of code system but a separate entity to categorize associations.

3.1.7 Designation

A designation is a representation of concept. Each concept can have multiple representations. Designations are uniquely identified by their ids. For example, in SNOMED-CT, the concept ”fever” has designation ”fever (finding), ”febrile” and ”pyrexia”.

3.1.8 Valueset

A valueset represents set of unique concepts grouped to meet a specific purpose. Its complexity ranges from flat list of concepts from single code system to poly-hierarchical set of concepts from multiple code systems. Concepts are linked to valueset using ConceptValusetMembership class.

3.2 CTS Functional Profiles

The CTS2 functional profile defines set of operations within the CTS2 specification. The defined functions must be supported by the developed system in order to conform to the profile.

3.2.1 CTS Query Profile

The CTS2 query profile covers the query and searching of terminology content. Table 3.1 reflects the structure of SNOMED Concept using an example.

Table 3.1: CTS Query Profile

Member Operation	**Operation Profile**
List Code Systems	Provide list of code systems that meets input criteria
Return Code System Details	Reterive all code system information including meta data, concepts and associations
List Code System Concepts	Provide list of code system concepts that meets input criteria
Return Concept Details	Reterive all concept information including meta data, designations and associations
List Valuesets	Provide list of valuesets that meets input criteria
Return Valueset Details	Reterive all valueset information including meta data, designations and associations
List Valueset Contents	Provide list of valueset concepts that meets input criteria
List Associations	Provide list of associations that meets specified search criteria
Return Association Details	Reterive meta data of association
List Associations Types	Provide list of association types that meets specified search criteria
Return Association Details	Reterive meta data of association type

Chapter 4

Semantic Web Representation Standards

Semantic data is represented in various forms. Some of the known standards are explained below.

4.1 OWL

The Web Ontology Language is a W3C designed semantic web language for the representation of things, their relationships and group of things. Its a logic based language. Web Ontology Language (OWL) uses RDF/S for its syntax. Its instances/individuals, Range and Domain restrictions are described as in RDF/S.

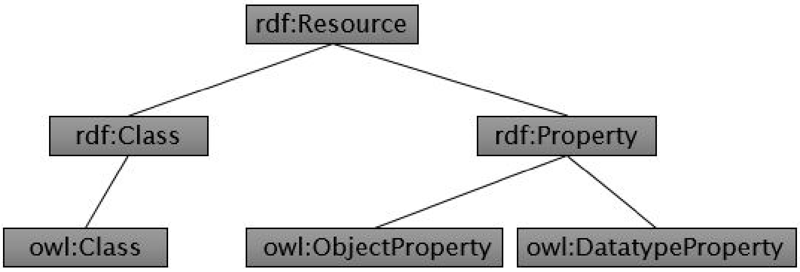

Figure 4.1: OWL

4.2 SKOS

Simple Knowledge Organization System (SKOS) was developed in the European project SWAD-Europe and was adopted as a standard by W3C [1]. It supports the representation, interoperability and use of a variety of structured vocabularies like taxonomies, thesauri or other classification schemes. It provides a way to represent Knowledge Organization Systems (KOS) using RDF to ensure the interoperability for information exchange and offers low migration cost for existing organization systems to semantic web environment. It can be seen as middle ware technology that bridges the gap between un-structured systems and semantic web. The essential components of SKOS are ConceptSchemes, Concepts, SemanticRelation, and DocumentaryNotes. Figure 4.2 shows the conceptual model of SKOS. Using SKOS, Concepts are identified by URIs, labelled using lexical strings, assigned lexical codes (notations), documented with various types of note, linked to concepts and organized into associations networks, aggregated to concept-schemes, grouped into collections and mapped to concepts of other concept-schemes.

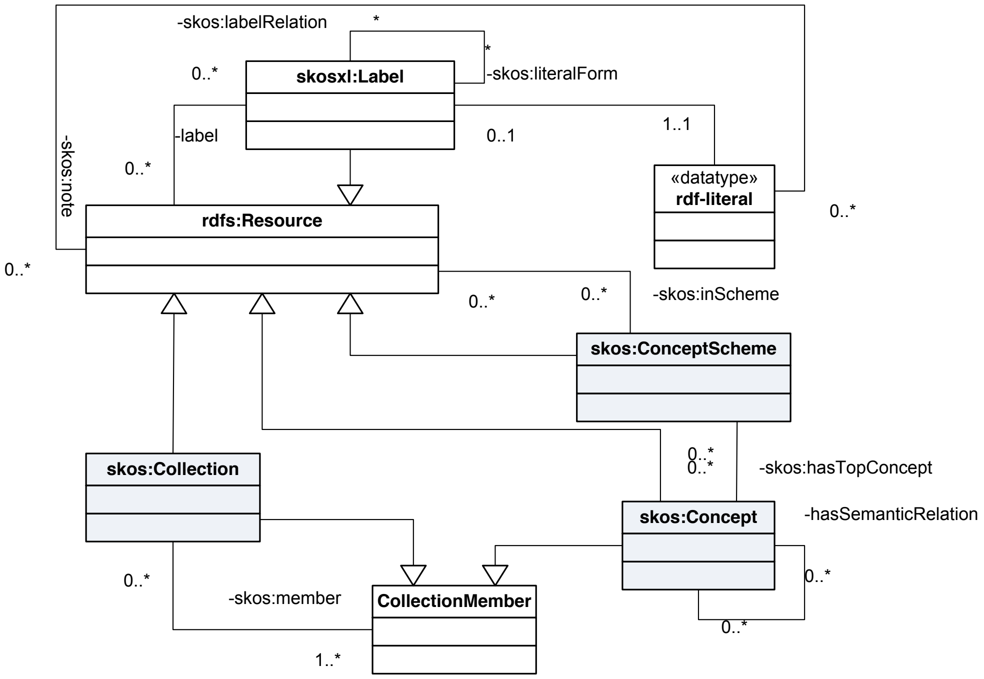

Figure 4.2: SKOS Conceptual Model [1]

Chapter 5

Related Work

Different research efforts have been carried out in past to cope with heterogenous medical terminology contents. Most of the solutions are based on limited number of terminologies and use relational databases for content storage. OpenGalen project, for instance, provides access to the contents of GALEN terminology [17]. GALEN is a coding system for representing medicines. GALEN reference model is built using GRAIL (GALEN Representation and Integration Language) and is delivered via GALEN Terminology Server. The GALEN reference model supports authoring, viewing, validating, and maintaining the GALEN terminology.

The LexGrid project [18] contributed ontologies and coding schemes of different disciplines through common model, tools and services for enabling the integration among otherwise heterogenous terminologies. Its goal was to made terminology resource available in real-time. LexGrid followed the bottom-up approach; they gathered information from multiple sources and designed a management tool named LexGrid editor. The editor supports authoring, viewing, validating, maintaining and extending terminologies that are defined in LexGrid common terminology model. The Lexicon Query Service (LQS) is similar to LexGrid in a way that it also specifies set of interfaces for accessing and querying medical terminologies [19]. The interfaces are developed using OMG Interface Description Language (IDL) standard. Guidelines for LQS method invocation via different languages and specification of the outcome of the method are provided in sufficient details. LQS has attempted to incorporate range of terminologies such as SNOMED and ICD.

A solution to integration terminology problem could also make a very good business case. Apelon, for example, developed Distributed Terminology System (DTS) as a commercial product. Apelon-DTS is an integrated solution for terminology services in distributed environment [7]. It supports deployment and maintenance of vocabularies within applications such as order entry system, patient charting and decision support system. The Apelon system normalizes data via lexical analysis, performs code translation by mapping medical data to coding system, and also performs semantic navigation, i.e., traversal of hierarchal relations between concepts.

Researchers have also investigated code system binding with EHR [20]. Such a work can augment the relationship between model of meaning (ontology) and model of codes (code system) grounded on the hypothesis that information models i.e. models of meaning and models of code are actually models of data structures. A code binding interface can further constrain and specify how codes should be used in data structures and also to restrict the attribute values to specific codes at different levels of abstraction. For Example, an attribute in the clinical data model may be restricted to specific codes for hypertension or of sub-categories of hypertension. As highlighted earlier, such work mainly caters singular terminology (SNOMED in many cases) and addition of further terminologies requires substantial changes in the knowledge model.

Contrary to integration, the work presented in [10] proposes mapping of ICD-11 with SKOS model. ICD-11 content structure is normalized to avoid semantic mismatch between both models. Interestingly it also suggests a RESTful service framework to support lookup and terminology authoring and uses the URI scheme proposed by WHO. To accommodate multiple terminologies, not just ICD-11 into single semantic structure, there is a need for a unified knowledge model based on semantic web standards such as SKOS. Most existing solutions either map single terminology to the presented model or only focus on integrated access of the terminology resources instead of semantically fusing multiple terminologies into a unified model. In the subsequent sections we present our approach towards a unified terminology model and a mechanism for integrated access of terminologies.

Chapter 6

Proposed Unified Terminology Model

The general idea of the proposed solution is given in Figure 6.1. The unified model is designed using SKOS and named as Unified Terminology Service (UTS). UTS model in combination with SKOS provides a generic knowledge model for representation and management of heterogenous terminologies. We have also mapped different terminologies to UTS model to demonstrate its application.

6.1 Challenges

The major challenge is the realization of non-semantic CTS2 conceptual model to semantic based SKOS model.

- Some mappings are already implicitly specified in CTS2 specifications, e.g., CodeSystem in CTS2 can be implicitly mapped with ConceptScheme in SKOS.
- Some mappings will be partly mapped e.g. lists in CTS2 can be partly mapped with Collections in SKOS.
- Some mappings do not exist in SKOS model; such mappings need to be extended using OWL in a logically consistent manner.

The first and second challenges are dealt by direct and partial mappings of CTS2 to SKOS and is shown in Table 6.1. The prefix used in the ontology model for CTS is “uts”. The third challenge is coped by making explicit entities and properties using OWL.

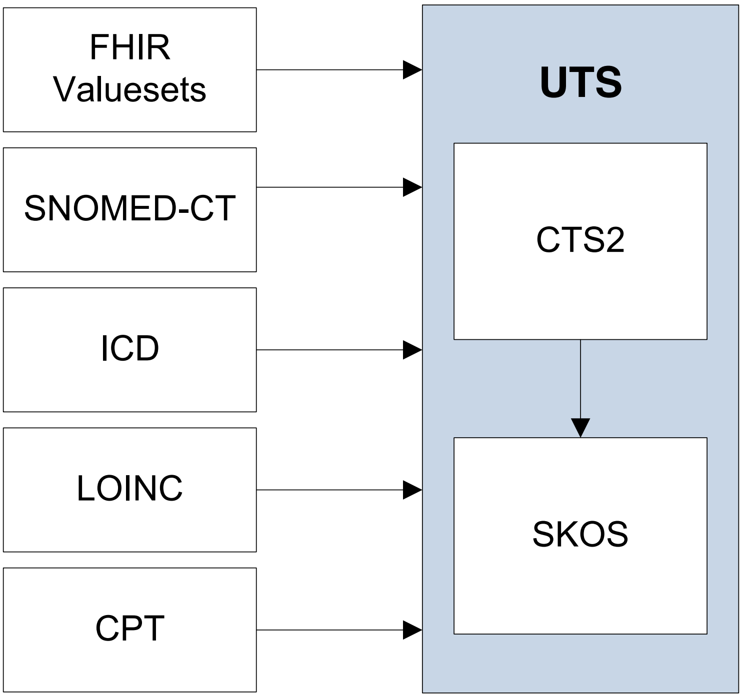

Figure 6.1: UTS as a combination of medical terminologies and CTS

6.2 Conceptual Model

After dealing with the challenges and keeping in mind the fact of accommodating heterogenous terminologies into minimal common structure, we have designed Unified Terminology Model (UTS) shown in Figure 6.2. Most of the UTS key concepts and properties are inherited from SKOS and are discussed below.

- **uts:Concept** is inherited from the skos:Concept. This entity is used for incorporating concepts of different medical terminologies having their own concept structure.

- **uts:CodeSystem** is inherited from skos:ConceptScheme. This entity is used for incorporating the metadata of the different medical terminologies.

- **uts:Valuset** is inherited from skos:Collection. This entity is used for incorporating the subsets, collections, valuesets of different medical terminologies.

Table 6.1: CTS2 to SKOS Mapping

CTS2 Entity	**SKOS Entity**
CodeSystem	ConceptScheme
Concept	Concept
ValueSet	Collection
DesignationType	prefLabel altLabel hiddenLabel
Designation	literal
Description	note
CodeSystemEntity VersionAssociation	semanticRelation
CodeSystemEnity	rdfs:resource
ConceptValueSet MemberShip	member
AssociationType	mappingRelation related broader broaderTransitive narrower narrowerTransitive
Name	rdfs:label

- **uts:EntityVersionAssociation** is inherited from rdf:Statement. The properties of rdf:Statement allows to define the subject, predicate and object separately, also there are some specialized properties to incorporate associations of different medical terminologies.

- **uts:CodeSystemEntity** defines the basic entities associated with the particular code system. Its specialized entities are uts:Concept, uts:Valueset and uts:EntityVersionAssociation.

- **uts:EntityStatus** is used for managing the statuses of the entities. The defined statuses are : reinstate, active, inactive, remove, cancel and suspend.

- **uts:CodeSystemStatus** is used for managing the statuses of the code systems. The defined statuses are: active and inactive.

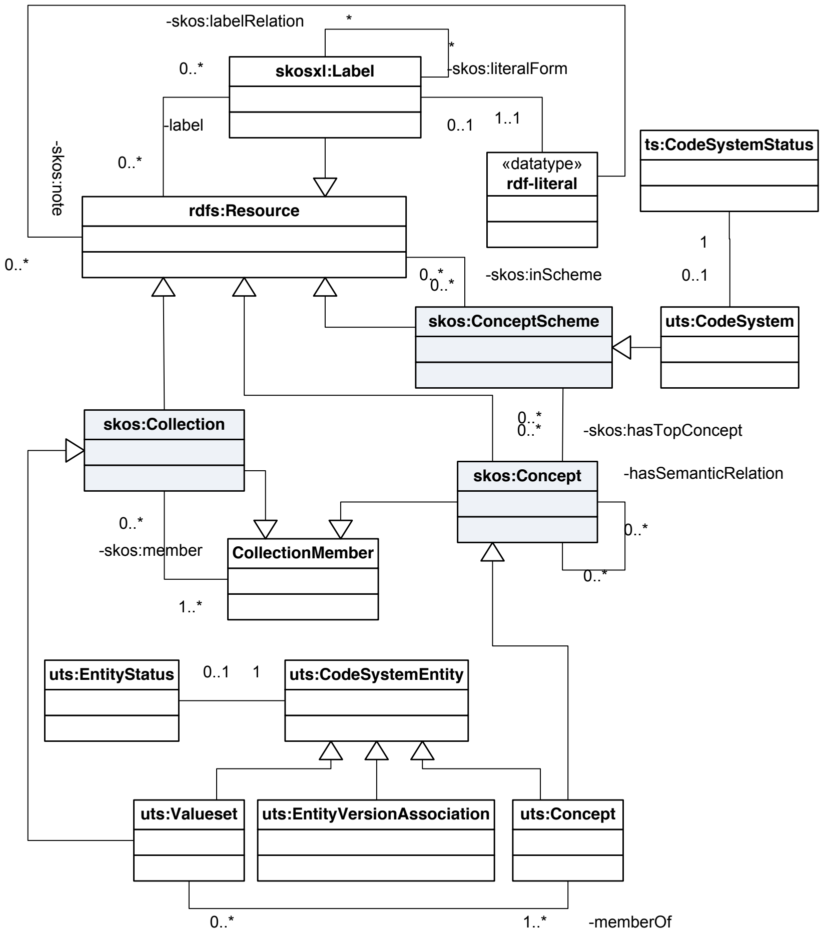

Figure 6.2: UTS Conceptual Model

6.3 Specialized Concepts and Properties

We have mapped the the CTS2 model to SKOS and came up with the UTS model that is a minimal common structural model to accommodate basic set of information about the heterogenous medical terminologies. Keeping in mind the fact, that terminology content vary in structure in terms of format, concepts and properties, there is a need to add some specialized concepts and properties in the UTS model to fully incorporate the terminology content i.e. without loss of any significant information. It will allow to store and retrieved content from unified repository.

The specialized concepts are inherited from skos:Concept and their associated properties are inherited from skos:semanticRelation. Namespace of ICD, LOINC and SNOMED-CT is http://hl7.org/fhir/sid/icd-10/ http://loinc.org/v2.46/

and http://snomed.info/sct/ respectively. Table 6.3 shows some of the concepts and their semantic properties.

Table 6.2: Specialized Concepts and Properties

Terminology	**skos:Concept**	**skos:SemanticRelation**	**rdfs:label**
ICD		icd:differentFrom	Is different from
LOINC	loinc:ChangeType	loinc:changeType	change type
	loinc:Class	loinc:class	class
	loinc:ScaleType	loinc:scaleType	Scale type
	loinc:MethodType	loinc:methodType	Method type
	loinc:OrderObs	loinc:orderObs	order or observation
	loinc:Property	loinc:property	property
	loinc:Source	loinc:source	source
	loinc:System	loinc:system	system
	loinc:TimeAspect	loinc:timeAspect	time aspect
SNOMED		snomed:sno.246061005	attribute
		snomed:sno.410662002	concept model attribute
		snomed:sno.260507000	Access
		snomed:sno.246090004	associated finding

6.4 Terminology Mappings

After mapping the CTS2 to SKOS based semantic model, next step is to map heterogeneous terminologies to the UTS model. Each terminology has its own structure and attributes, so there is a need to normalize their structure accordingly. SNOMED is one of the known comprehensive standard medical terminology and is widely used across the world. The major entities of SNOMED are concepts, descriptions, associations and subsets which are mapped to uts:concept, skos-xl:Label, skos:semanticRelation and uts:Valueset respectively. LOINC is known for laboratory orders and is one of the important terminologies for achieving interoperability among the laboratory reports of medical institutes round the globe. LOINC entities include Loinc_Num, Component, Relatednames (Fsn), Short_Name, Long_Name, and System. These entities are mapped to uts:conceptCode, uts:Concept, skos:Definition, rdfs:label and skos:altLabel respectively. ICD is a classification of diseases. We have opted the ICD version 10 and mapped it over UTS model. Rest of the mappings of SNOMED, LOINC, ICD, CPT, FHIR and DRG are shown in Table 6.3.

Table 6.3: Terminology Mappings with Unified Model

UTS Model	SNOMED	LOINC	ICD	FHIR	CPT	DRG
uts:codeSystemID	snomed:uid	loinc:uid	icd:uid			
uts:Concept	snomed:Concept	loinc:Component	icd:Chapter	fhir:Concept	cpt:Concept	drg:Concept
	snomed:Relationship Concept1	loinc:System	icd:Category			
		loinc:MethodType	icd:Block			
	snomed:Relationship Concept2	loinc:ScaleType	icd:code			
		loinc:ChangeType				
		loinc:Class				
		loinc:OrderObs				
uts:Association	snomed:Relationship		icd:Association			
uts:Valueset	snomed:Subset	loinc:ClassTypes		fhir:Valueset		drg:MDC drg:Type
uts:CodeSystem Status						
uts:EntityStatus	snomed:Concept Status	loinc:Status		fhir:Status		
	snomed:SubsetStatus					
	snomed:Description Status					
uts:conceptCode	snomed:ConceptCode	loinc:LoincNum		fhir:Code	cpt:Code	drg:Code
uts:EntityID	snomed:ValuesetID		icd:BlockCode	fhir:Identifier		
			icd:Category Code			
			icd:ChapterCode			
			icd:ModifierClass			
uts:statusDate		loinc:DateLast Changed		fhir:Date		
uts:releaseDate						
uts:previous VersionID		loinc:MapTo				
uts:versionID			icd:Version	fhir:Version		
uts:copyright				fhir:Copyright		
uts:source				fhir:Publisher		

UTS Model	SNOMED	LOINC	ICD	FHIR	CPT	DRG
uts:ConceptValueSetMembership	snomed:Valueset Member	loinc:ClassTypes Member				drg:MDC Member
						drg:Type Member
uts:CodeSystem EntityVersion Association	snomed:Relationship Type					
uts:Association-Type			icd:SuperClass	fhir:contains		
			icd:SubClass			
			icd:Exclusion			
uts:Description			icd:Text	fhir:Description		
			icd:Note			
			icd:ExclusionNote			
			icd:ModifierLink			
			icd:FootNote			
uts:Designation-Type	snomed:Synonym	loinc:LongName	icd:PrefLong			
	snomed:Pref Description	loinc:shortName	icd:PrefShort			
			icd:Inclusion			
			icd:Preffered			
uts:Designation	snomed:Term					
uts:name	snomed:ValuesetName		icd:Name	fhir:Name	cpt:Short Description	drg:Title
				fhir:Display		
rdfs:comment		loinc:Comment				
skos:changeNote		loinc:StatusReason				
skos:scopeNote			icd:Introduction			
skos:example			icd:CodingHint			
skos:definition	snomed:FullySpecified Name	loinc:RelatedNames	icd:Definition	fhir:Definition	cpt:long Description	
skos-xl:Label	snomed:Description					

Chapter 7

Proposed System Architecture and Implementation

7.1 System Architecture

To cope with the need of heterogeneous terminologies, only knowledge model does not solve the problem. To come up with real solution some application needs to be developed to prove that knowledge model is actually accommodating different terminologies. Figure 7.1 shows the proposed system architecture for building standardized RESTful services for heterogeneous medical terminologies. The architecture is designed keeping in mind that:

- Multiple clients could access terminology service
- Services should be light weight
- Access to the heterogeneous content should be efficient

The components which are involved in the architecture are described below.

7.1.1 Terminology Authoring Application

This component provides web based solution for the authoring of code system, concept, valueset, association and association type. The input of this component is JSON-LD of the respective entity and after successful storage of data, the component responds with the generated RDF. The component allows multiple clients to access the terminology service.

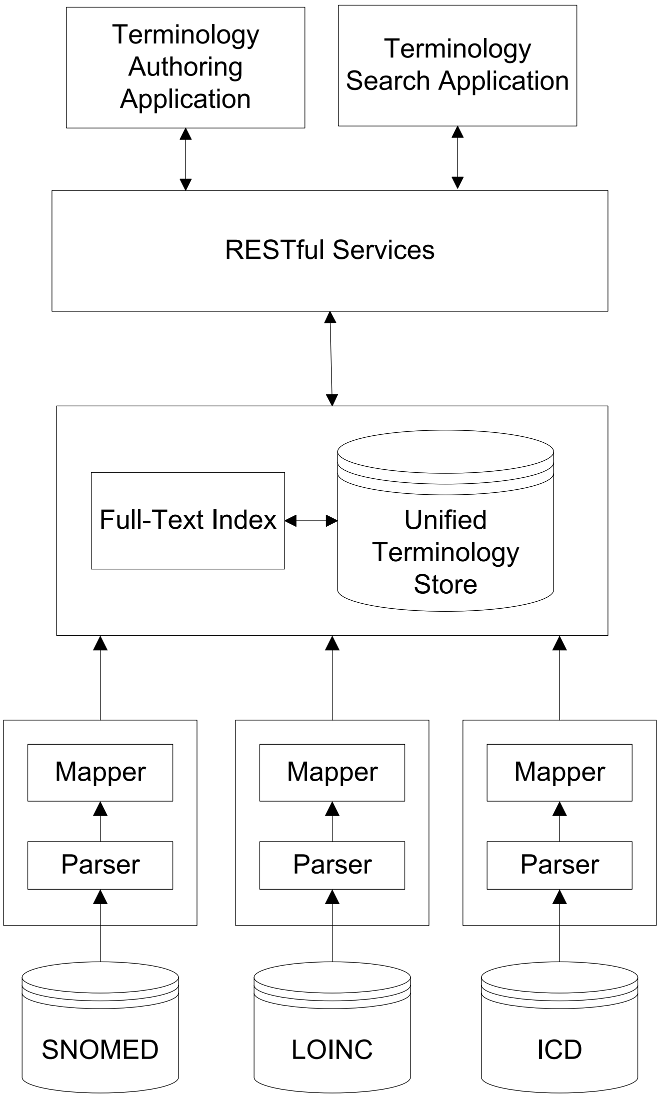

Figure 7.1: UTS Architecture

7.1.2 Terminology Search Application

This component provides web based solution to search and navigate among the content of various code systems. The input of this component is filtering criteria and id of the particular entity. The input is transformed into the respective SPARQL query and component responds with the result-set of the executed query.

7.1.3 UTS API as RESTful Service

This component is responsible for building RESTful services that will actually process the request received from the client application. The services will respond in multiple formats of semantic web including JSON-LD, RDF and Turtle. We have defined the URI templates to which a UTS service should respond. Table 7.1.3 contains the generic service paths and their descriptions that should be followed for access.

Table 7.1: REST Resource URI's

/entity	The service with such URI pattern provides with the ability to create and list code systems, associations, association types and valuesets.
/entity/{id}	The service with such URI pattern provides with the ability to edit, delete and details of a code system, association, association type and valueset.
/entity/{id}/concepts	The service with such URI pattern provides with the ability create and list code system's and valueset's concepts.
/entity/{id}/concepts/{cid}	The service with such URI pattern provides with the ability to edit, delete and get details of a concept, of a particular valueset or code system.
/entity?[search]	The service with such URI pattern provides with the ability to list code systems, concepts, associations, association types and valuesets that meets up the input search criteria.

The search criteria parameters are different for each entity i.e. code system, concept, valueset, association and association type, parameters are defined by following CTS2 specifications. Table 7.1.3 shows the search parameters for each entity.

Table 7.2: Entity & Search Parameters

codeSystem	?criteria
concepts	?criteria&matchType
associations	?subject&predicate&object
associationTypes	?criteria
valuesets	?criteria

7.1.4 Lucene Index

This component is responsible for efficient access of terminology content. Lucene indexes are used for storing indexes and to ensure efficient search and retrieval of the content.

7.1.5 SKOS based Terminologies

This component is responsible for storing the RDF content using the knowledge model of UTS. Terminology content is searched from datastore, if the queried content is not available in the indexes.

7.1.6 Translator

This component is responsible for translating the multiple terminologies content using defined UTS semantic model. The existing terminologies content is sent to parser, which parses the content based on format and sends the parsed content to mapper, mapper then generates RDF following the UTS model and RDF content is then stored in the Data store. The challenges that need to dealt while carrying out the translation process are:

- Parsing heterogeneous formats
- URI Schemes for each terminology
- Mapper of each terminology

7.2 Implementation

In the implementation, we focused on normalizing multiple terminologies including SNOMED, ICD, LOINC, CPT and DRG using SKOS based CTS2 owl model and building the CTS2 compliant semantic web RESTful services.

The major interface is resource from that all other interfaces are inherited (c.f. Figure 7.2). Another important interface is terminology reader which is responsible for taking the input stream along with the format of the content and send that stream to specific parser. The parser is specialized for each terminology as they are in different formats including JSON, CSV, XML, cLaml and RDF. The parsed content is sent to specific mapper. The reason for specialized mapper for each terminology is diversity in the structure and some specialized fields other than core that are mapped to UTS model. Terminology mapper is responsible for managing URI schemes and generating RDF. The URIs of terminologies are taken from the FHIR recommendations [21], but URI for each concept is defined by taking the terminology URI as a base name space and it's appended with the version of the terminology and the concept code. We have used apache wink implementation of JAX-RS framework for service implementation.

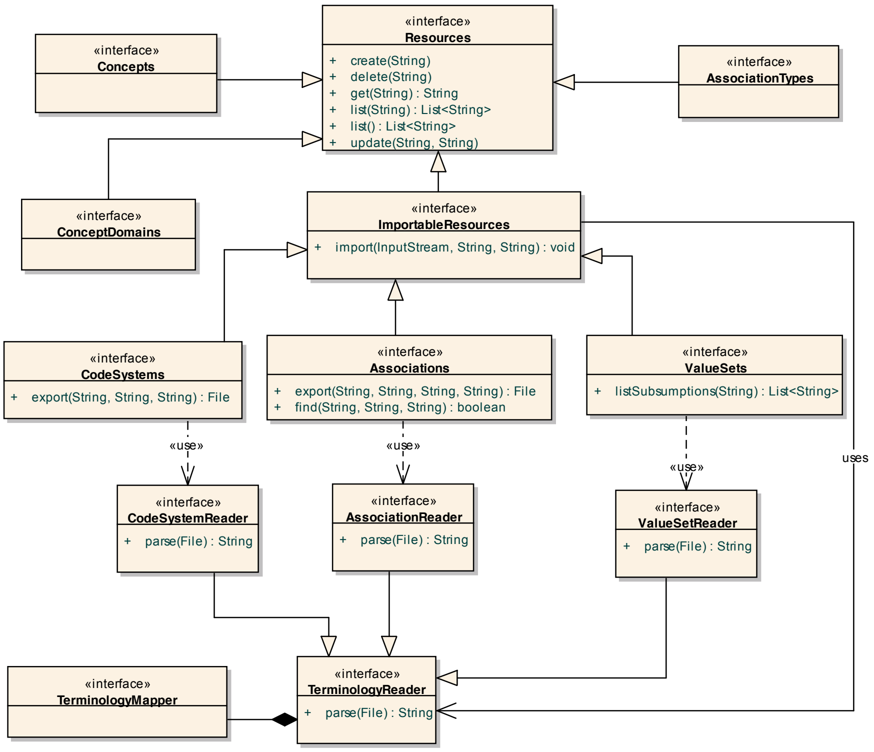

Figure 7.2: UTS API

Chapter 8

Results and Evaluation

8.1 Data Set and Test Environment

Unified Terminology Service (UTS) is tested and evaluated against 7 heterogenous medical terminologies. These terminologies vary in size and representation. The terminologies having larger size and higher representational complexity are considered to be more complex terminologies. Table 8.1 shows the complexity criteria of heterogenous terminologies.

Table 8.1: Terminologies Complexity

Terminology	**Size**	**Representation**	**Complexity**
SNOMED	1,730,787	OWL-DL	High
LOINC	730,369	Tabular List	Medium
ICD	39,343	OWL-DL	High
DRG	749	Tabular List	Low
UCUM	500	Tabular List	Low
Fhir Valuseset	31,108	JSON	Medium
CPT	12,480	Tabular List	Low

The evaluation is performed on a regular desktop with following specifications:Core i7, 2.53GHz, 4 GB memory, 64 bit Windows OS, Hard Disk 900 GB.

8.2 Results

UTS is evaluated in terms of load/import time of each terminology. With the increased size of number concepts and relationships in terminology, the import took longer time. Figure 8.1 shows the results of load time of selected medical terminologies.

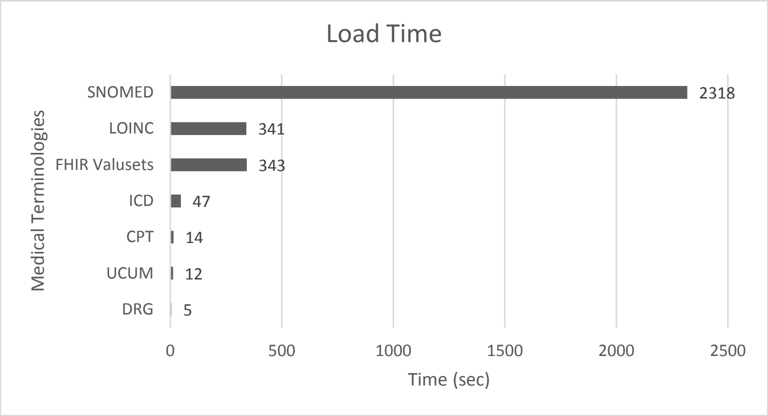

Figure 8.1: Terminologies Load Time

The terminology statistics including number of concepts, associations before the import of terminology and total number of triples in the datastore after the import of terminology is shown in Table 8.2.

Table 8.2: Terminology Statistics

Terminology	Concepts	Associations	Triples
CPT	12,480	0	59,989
FHIR v2	3,265	0	26,027
FHIR v3	1,854	25,998	250,275
ICD	11,373	27,970	341,031
LOINC	73,115	657,369	4,628,653
SNOMED	392,914	1,430,787	15,214,679
MS-DRG	749	749	11,330

In the import process, each terminology is parsed and mapped to the UTS model. The mapped content is then stored as RDF triples in datastore. Example RDF triples of SNOMED code system and SNOMED concept is shown below.

```
<http://snomed.info/sct/v1.0>
    a <http://www.semr.ts/CodeSystem> ;
    <http://www.w3.org/2000/01/rdf-schema#label>
        "SNOMED CT"^^<http://www.w3.org/2001/XMLSchema#string> ;
```

```
    <http://www.semr.ts/codeSystemID>
        "2.16.840.1.113883.6.96"^^
            <http://www.w3.org/2001/XMLSchema#string> ;
    <http://www.semr.ts/codeSystemStatus>
        <http://www.semr.ts/codesystem.active> ;
    <http://www.semr.ts/copyright>
        "IHSTDO"^^<http://www.w3.org/2001/XMLSchema#string> ;
    <http://www.semr.ts/effectiveDate>
        "2015-05-11T11:37:25.595Z"^^
            <http://www.w3.org/2001/XMLSchema#dateTime> ;
    <http://www.semr.ts/flagged> false ;
    <http://www.semr.ts/releaseDate>
        "2014-01-02T19:00:00Z"^^
            <http://www.w3.org/2001/XMLSchema#dateTime> ;
    <http://www.semr.ts/source>
        "IHSTDO"^^<http://www.w3.org/2001/XMLSchema#string> ;
    <http://www.semr.ts/statusDate>
        "2015-05-11T11:37:25.595Z"^^
            <http://www.w3.org/2001/XMLSchema#dateTime> ;
    <http://www.semr.ts/versionID>
        "1.0"^^<http://www.w3.org/2001/XMLSchema#string> ;
    <http://www.w3.org/2004/02/skos/core#note>
        "SNOMED Clinical Terms (SNOMED CT) is the most comprehensive,
        multilingual clinical healthcare terminology in the
        world."^^<http://www.w3.org/2001/XMLSchema#string> .

<http://snomed.info/sct/v1.0/sno.300367007>
    a <http://www.semr.ts/Concept> ;
    <http://www.w3.org/2000/01/rdf-schema#label>
        "Finding of measures of vomit
        (finding)"^^<http://www.w3.org/2001/XMLSchema#string> ;
    <http://www.semr.ts/codeSystem>
        <http://snomed.info/sct/v1.0> ;
    <http://www.semr.ts/conceptCode>
        "300367007"^^<http://www.w3.org/2001/XMLSchema#string> ;
    <http://www.semr.ts/entityStatus>
        <http://www.semr.ts/enity.active> .

<http://snomed.info/sct/v1.0/sno.300367007>
    <http://www.w3.org/2008/05/skos-xl#altLabel>
        <http://snomed.info/sct/v1.0/sno.1226794019> ;
    <http://www.w3.org/2008/05/skos-xl#prefLabel>
        <http://snomed.info/sct/v1.0/sno.441415016> .

<http://snomed.info/sct/v1.0/sno.1226794019>
    a <http://www.w3.org/2008/05/skos-xl#Label> ;
    <http://www.semr.ts/entityStatus>
        <http://www.semr.ts/enity.active> ;
    <http://www.w3.org/2008/05/skos-xl#literalForm>
```

```
        "Observation of measures of vomit"@en .

<http://snomed.info/sct/v1.0/sno.441415016>
    a <http://www.w3.org/2008/05/skos-xl#Label> ;
    <http://www.semr.ts/entityStatus>
        <http://www.semr.ts/enity.active> ;
    <http://www.w3.org/2008/05/skos-xl#literalForm>
        "Finding of measures of vomit"@en .
```

The mapping of each terminology to the UTS model is validated by translating the terminology content (RDF data) back to the format, provided by the terminology providers. Table **??** shows the results of correctness measure of SNOMED medical terminologies.

Table 8.3: SNOMED Reverse Mapping

UTS	**SNOMED**	**Result**
codeSystemID	uid	True
Concept	Concept	True
Association	Relationship	True
Valueset	Subset	True
skos:member	SubsetMember	True
EntityStatus	ConceptStatus, SubsetStatus,DescriptionStatus	True
ConceptCode	ConceptCode	True
entityID	SubsetID	True
semanticRelation	RelationshipType	True
DesignationType	Synonym, PrefferedDescription	True
Designation	Term	True
skos:definition	FullySpecifiedName	True
rdfs:label	CodeSystemName, ConceptName, SubsetName	True

Another considered criteria to test UTS is performance evaluation and its comparison with NCI Terminology Browser. Response time is evaluated using user to average response time ratio. UTS is deployed over a CentOS based 64bit server with 32GB memory. After the import of terminologies including SNOMED, LOINC, CPT, ICD, DRG and FHIR value sets, the size of the dataset grows to 3.26 GB and approximately 19million RDF triples.

NCI browser and UTS are tested against multiple scenarios of listing, discovery and getting metadata. Their performance evaluation is shown in graphs, generated using Apache JMeter. In first scenario the UTS and NCI are tested for code system listing. Figure 8.2 shows the average response time per request, if number of users access the application concurrently.

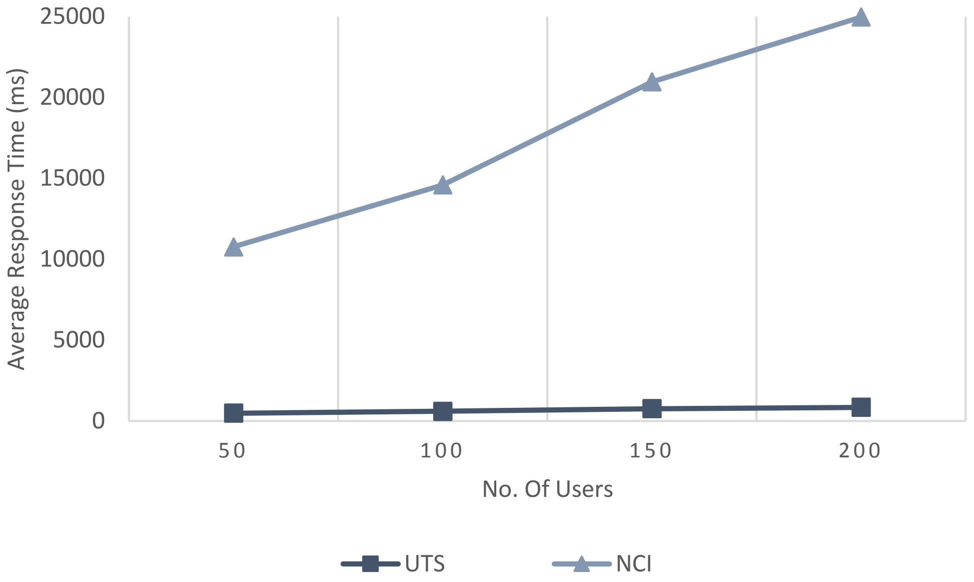

Figure 8.2: Average Code System Listing Response Time

In second scenario the applications are tested for code system discovery. The term snomed is searched in both applications and the generated results of average response time per request are shown in Figure 8.3.

In third scenario the applications are tested for code system metadata. The generated results of average response time per request for viewing the detailed metadata of the code system are shown in Figure 8.4.

In forth scenario the applications are tested for concept discovery in a particular code system. The term cancer is searched in SNOMED-CT in both applications and the generated results of average response time per request are shown in Figure 8.5.

In fifth scenario the applications are tested for concept metadata. The generated results of average response time per request for viewing the detailed metadata of the concept including synonyms, associations and valuesets are shown in Figure 8.6.

The results of considered scenarios have shown that UTS offers relatively lower average response time than NCI browser.

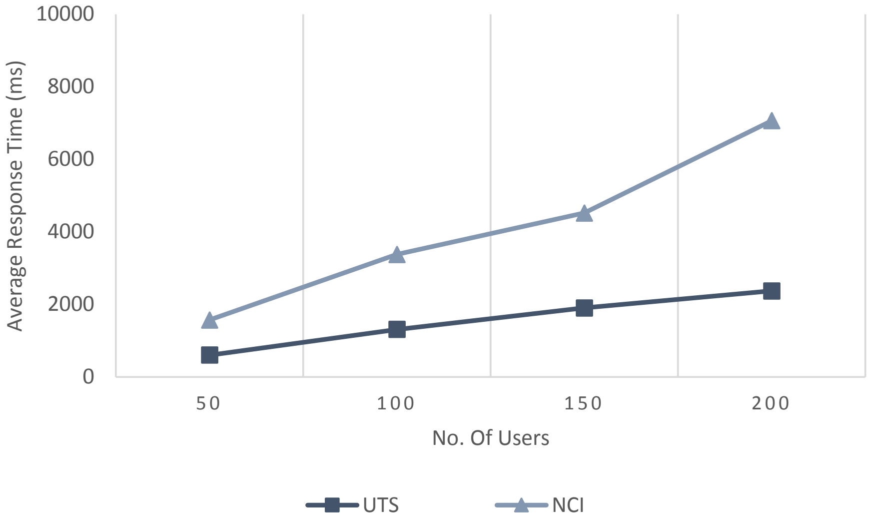

Figure 8.3: Average Code System Discovery Response Time

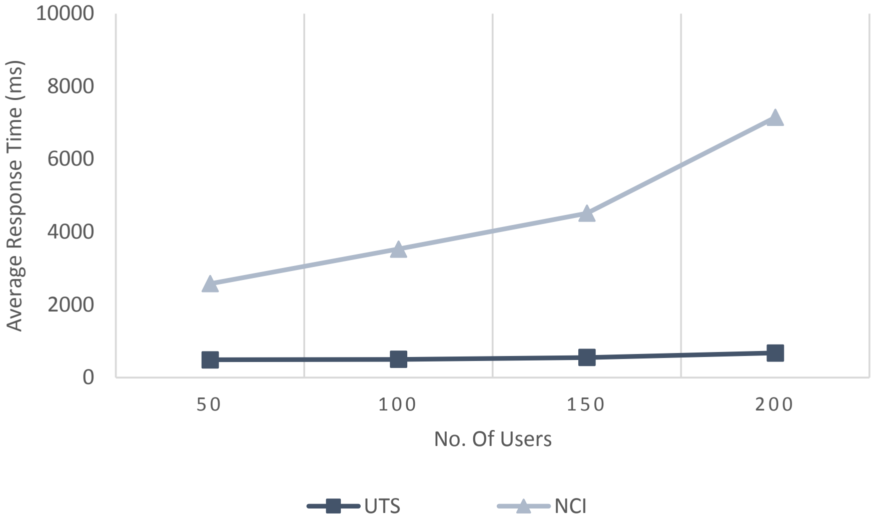

Figure 8.4: Average Code System Metadata View Response Time

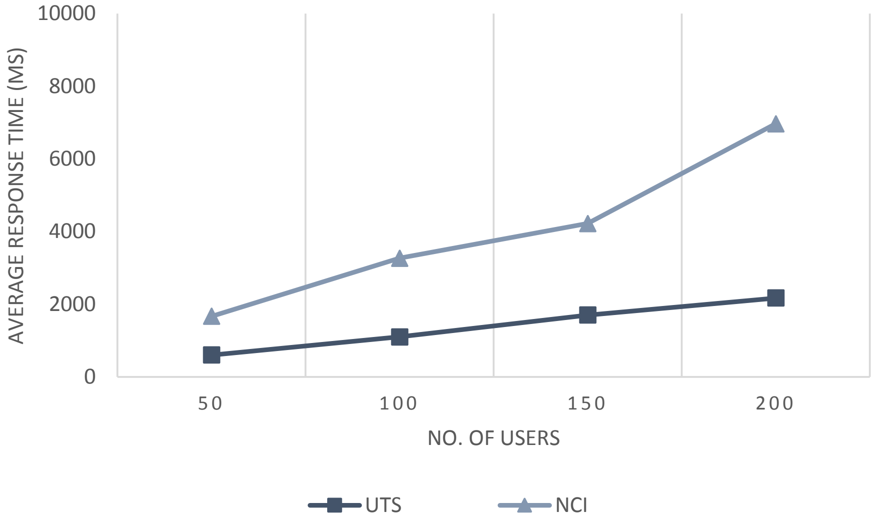

Figure 8.5: Average Concept Discovery Response Time

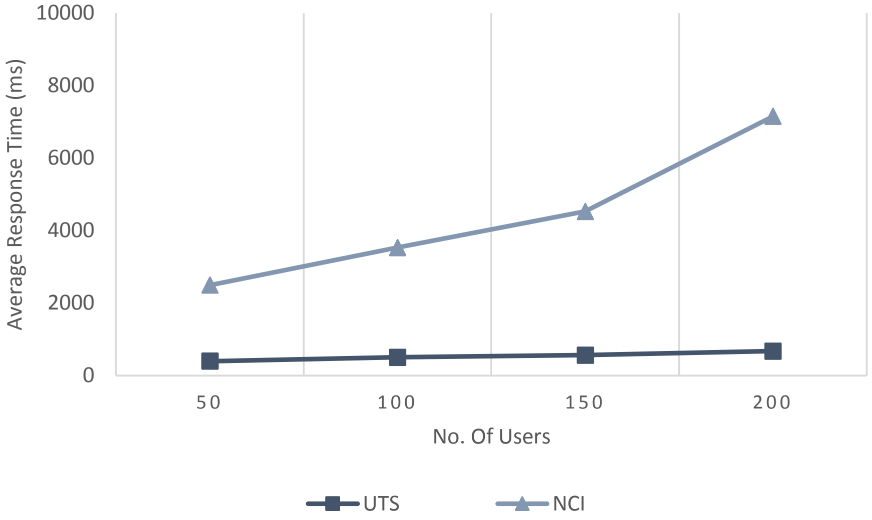

Figure 8.6: Average Concept Metadata View Response Time

Chapter 9

Conclusion and Future Work

9.1 Conclusion

Standard medical terminologies are developed with specific purposes in mind, are structured very differently. The format of the terminology may range from a simple flat list of concepts, to more complex poly-hierarchies. This thesis describes the problem of heterogeneous terminology standards and to cope with the problem a SKOS based knowledge model using CTS2 standard is developed to accommodate diverse structure and formats of terminologies. An effort is made to translate and map the terminological content to UTS model and then store the mapped content as RDF data in the data store. CTS2 compliant Restful services are developed, that intent to support the navigation, searching and authoring of multiple medical terminologies. The proposed model could enhance the semantic interoperability beyond the traditional health care applications. The UTS is evaluated in terms of load time, the load time increased with the increased complexity. The size of terminology before and after load is calculated to provide reader with overall system statistics.

9.2 Future Work

We will continue to include more standard medical terminologies into the UTS. Moreover, biomedical ontologies like cancer ontology can be mapped over UTS model. The efficiency of the service can be improved to give access in minimum response time and maximum throughput.

References

[1] Alistair Miles and Sean Bechhofer. Skos simple knowledge organization system reference. *W3C recommendation*, 18, 2009.

[2] James Cimino. Review paper: coding systems in health care. *Methods of Information in Medicine-Methodik der Information in der Medizin*, 35(4):273–284, 1996.

[3] J. Quinn. An hl7 (health level seven) overview. *Journal of AHIMA / American Health Information Management Association*, 70(7):324; quiz 356, 1999.

[4] BB Wein. Ihe (integrating the healthcare enterprise): a new approach for the improvement of digital communication in healthcare. *RoFo : Fortschritte auf dem Gebiete der Rontgenstrahlen und der Nuklearmedizin*, 175(2):183186, February 2003.

[5] Stella G. Dextre Clarke and Marcia Lei Zeng. Iso 25964 the international standard for thesauri and interoperability with other vocabularies. Technical report, National Information Standards Organization, 2013. http://www.niso.org/schemas/iso25964/.

[6] Russell Hamm, Craig Stancl, Kristi Eckerson, Davera Gabriel, Alan Honey, Dr. Jobst Landgrebe, Lloyd McKenzie, Grahame Grieve, John Carter, Senthil Nachimuthu, Harold Solbrig, Heather Grain, Ted Klein, Beverly Knight, and Jobst Landgrebe. Service functional model specification common terminology services release 2 (cts 2). Technical report, Health Level Seven International, October 2009. http://www.hl7.org/documentcenter/public/ standards/dstu/2009may/V3_CTS_R2_DSTU_2009OCT.pdf.

[7] Apelon dts (distributed terminology system), Visited: September 2014. http://apelon-dts.sourceforge.net/.

[8] F. Manola, E. Miller, and B. McBride. Rdf primer. W3c recommendation, World Wide Web Consortium, February 2004. http://www.w3.org/TR/rdf-primer/.

[9] Pascal Hitzler, Markus Kroetzsch, Bijan Parsia, Peter F. Patel-Schneider, and Sebastian Rudolph. Owl 2 web ontology language primer (second edition). W3c recommendation, World Wide Web Consortium, December 2012. http://www.w3.org/TR/owl2-primer/.

[10] Guoqian Jiang, Harold R. Solbrig, and Christopher G. Chute. Building standardized semantic web restful services to support icd-11 revision. In *Semantic Web Applications and Tools for Life Sciences (WAT4LS)*, 2012.

[11] OMG CORBA and IIOP. Object management group, November 1997. Joint revised submission OMG document orbos/99-02.

[12] L Bos. Snomed-ct: The advanced terminology and coding system for ehealth. *Methods of Information in Medicine-Methodik der Information in der Medizin*, 54, 2006.

[13] Clement McDonald, Stanley M Huff, Jeffrey Suico, Gilbert Hill, Dennis Leavelle, Raymond Aller, Arden Forrey, Kathy Mercer, Georges DeMoor, John Hook, et al. Loinc, a universal standard for identifying laboratory observations: a 5-year update. *Clinical chemistry*, 49(4):624–633, 2003.

[14] Natasha Rekhtman, Justin A Bishop, and Diana Weedman Molavi. Cpt (current procedural terminology) coding quick reference. In *Quick Reference Handbook for Surgical Pathologists*, pages 135–137. Springer, 2011.

[15] World Health Organization. International statistical classification of diseases and related health problems. 2010.

[16] Diagnosis-related groups. *Journal of Nursing Administration*, 103(157):387, December 2005.

[17] AL Rector, JE Rogers, PE Zanstra, and Egbert Van Der Haring. Opengalen: open source medical terminology and tools. In *AMIA Annual Symposium Proceedings*, volume 2003, page 982. American Medical Informatics Association, 2003.

[18] Jyotishman Pathak, Harold R Solbrig, James D Buntrock, Thomas M Johnson, and Christopher G Chute. Lexgrid: a framework for representing, storing, and querying biomedical terminologies from simple to sublime. *Journal of the American Medical Informatics Association*, 16(3):305–315, 2009.

[19] H. Solbrig and T. Brinson. Lexicon query service: Rfp response. *3M Health Information Systems*, pages 1–175, March 1998.

[20] Alan L Rector, Rahil Qamar, and Tom Marley. Binding ontologies and coding systems to electronic health records and messages. *Applied Ontology*, 4(1):51–69, 2009.

[21] Ewout Kramer and Lloyd McKenzie. Fhir named code-systems list. Technical report, Health Level Seven International, 2014. http://hl7.org/implement/standards/fhir/terminologies-systems.html.