Integrated Vocabulary Service for Health Data Interoperability

Sarah N. Lim Choi Keung, Lei Zhao, Edward Tyler, Theodoros N. Arvanitis
University of Birmingham
Edgbaston, United Kingdom
e-mail: {s.n.limchoikeung, l.zhao, e.tyler, t.arvanitis}@bham.ac.uk

F. D. Richard Hobbs
University of Oxford
Oxford, United Kingdom
e-mail: richard.hobbs@phc.ox.ac.uk

***Abstract*— The paper addresses the problem of interoperation when searching across patient data represented in several medical vocabularies. This is an important issue of relevance to eHealth integration that will allow clinical information to be used in clinical research. We propose a novel way to semantically integrate a number of vocabularies for reference using a vocabulary service.**

Keywords–interoperability, controlled vocabularies, electronic health record.

I. INTRODUCTION

The increasing amount of health-related data available calls for new ways of analysing them together as a unified set, despite their heterogeneity. The interoperability of healthcare data is an active research topic, especially with the investigation of the reuse of routine clinical data for clinical research. In many instances, users of healthcare data are familiar with only one medical vocabulary. For instance, in the English primary care, Read Codes Version 2 (RCV2) is still widely used. However, activities, such as patient cohort identification and recruitment, often need to query heterogeneous patient data repositories that use different coding systems. In this paper, we describe our approach to integrating medical vocabularies semantically by providing a vocabulary service for vocabularies commonly used in Europe. The remainder of the paper is organised as follows. Section II gives an overview of the medical vocabularies, existing work on vocabulary collections, their features and limitations. We then introduce our approach to an integrated vocabulary service in Section III, followed by details of the application architecture and web service implementation in Sections IV and V. Finally, we give our conclusions and future work.

II. MEDICAL VOCABULARIES

Medical vocabularies have been in use since patient information needed to be coded for statistical reporting, long before computers started being used in health care [1]. Different vocabularies have been created for various purposes and there is currently no single agreed vocabulary in use. The two main uses for medical vocabularies are for classification of diseases for statistics and reporting, and for the coding of clinical data for patient care. Reviews of the common medical terminology include the works of [2, 3]. The first category of vocabularies includes classifications, such as ICD-10 [4] and the OPCS Classification of Interventions and Procedures (OPCS-4) [5]. They are mainly used to simplify data and create abstractions for statistical reporting and for reimbursements [1]. The second category includes those that can represent detailed information as part of a patient's electronic healthcare record (EHR). Examples include SNOMED CT [6], Read Codes [7], ICPC-2 [8]. As noted by Cimino [1], As a number of vocabularies are used to represent healthcare data, a large amount of data can potentially be used to support clinical research, provided there is an effective means to semantically integrate them.

The Unified Medical Language System (UMLS) [9] is a collection of many controlled vocabularies in the biomedical sciences, developed in an attempt to unify disparate vocabularies and to facilitate the sharing of medical knowledge [1, 3]. UMLS consists of three Knowledge Sources (Metathesaurus, Semantic Network and SPECIALIST Lexicon) and also provides a set of software tools to provide a mapping structure among the different vocabularies. The UMLS Metathesaurus is a large, multi-purpose and multi-lingual vocabulary database, containing information about biomedical and health related concepts, synonyms and relationships among them [10]. The Semantic Network defines how the concepts in the Metathesaurus are assigned semantic types, which are broad subject categories, such as Disease, Finding and Clinical Drug, which are linked to one another through semantic relationships [11]. Despite the UMLS including a number of commonly used vocabularies, it does not support RCV2 and other widely used mappings and languages used in Europe.

III. INTEGRATED VOCABULARY SERVICE

Our approach to resolve some of the limitations of the UMLS is to develop an integrated vocabulary service, which will enable the support of more European terminologies and languages. It follows the lessons learnt and therefore extends on the work achieved in the ePCRN project [12], a US National Institute of Health project investigating an electronic infrastructure to support the design and implementation of randomised clinical trials, while facilitating translational research in primary care in the United States.

A. Background: ePCRN

In the ePCRN project, the National Cancer Institute (NCI) Enterprise Vocabulary Services (EVS) provided the

 ISBN: 978-1-61208-179-3

controlled terminology and ontology. The EVS is a project of the US National Cancer Institute Center for Biomedical Informatics and Information Technology (NCI CBIIT) and forms the semantic base for the Cancer Common Ontologic Representation Environment (caCORE), the NCI cancer Biomedical Informatics Grid (caBIG®), and the new NCI CBIIT semantic infrastructure [13], for supporting interoperability in translational research. As the NCI Metathesaurus is more US-oriented, it has fewer European coding systems and only supports vocabularies in the English language. The NCI EVS develops and supports the NCI Metathesaurus vocabulary source (NCIm), which is based on the UMLS Metathesaurus and thus provides a mapping of concepts to terms within multiple vocabularies. With the aim to include more European terminologies and support other European languages, we have developed an integrated vocabulary service (as part of the TRANSFoRm project [14]), using LexEVS. LexEVS is the open-source software package, on which the NCI EVS is built. LexEVS provides a comprehensive set of software and services to load, publish and access controlled terminologies. It supports a wide variety of ontology formats including UMLS RRF, OWL, OBO, HL7 RIM, and LexGrid XML. Other related works include i2b2 SHRINE [15] and the Ontology Lookup Service [16].

B. *Extending UMLS with Read Codes V2*

The first version of the integrated vocabulary service (VS) allows RCV2 to be semantically interoperable with the UMLS. A customised database of RCV2 has been created to cater for the English primary care domain. The UK Terminology Centre (UKTC) [17] provides the mapping from RCV2 to SNOMED CT and the integrated VS is based on this mapping, such that RCV2 concepts can be linked to search on the UMLS. A subset of the UMLS Metathesaurus 2010AA release is hosted by the integrated VS. It includes the NCI Thesaurus and vocabularies such as SNOMED, ICD10, and ICPC2, in different languages.

IV. APPLICATION ARCHITECTURE

Following a standard tiered architecture, the VS server is divided into three logical tiers: presentation, service and database tiers as depicted in Figure 1.

A. *Presentation Tier*

The presentation tier provides both a Web interface for web client access and a Web service interface for programmatic access. The web interface is designed as an Ajax application using Google Web Toolkit (GWT) technology. The Ajax web client renders the web page on the user's web browser and calls a GWT Remote Procedure Call (RPC) servlet, on the server side, when the user initiates a search. The web services interface is implemented using Metro; the web service stack was originally developed by Sun Microsystems. Both the GWT RPC servlet and the web service component, in turn, invoke the query service, the entry point of the service tier, to execute a search.

B. *Service Tier*

At the service tier, the query service component coordinates access to different vocabulary databases and combines the search results together. The UMLS vocabulary database is accessed through LexEVS, but additional vocabulary databases are accessed through direct JDBC calls. The LexEVS runtime is a Java class library, which provides various APIs to access vocabulary contents in the LexEVS format.

C. *Database Tier*

LexEVS builds an index on the vocabulary databases using Lucene technology, when the vocabulary data are imported into the LexEVS database. During query execution, the LexEVS runtime consults the Lucene index [18], in order to speed up the data search.

V. WEB SERVICE IMPLEMENTATION

The TRANSFoRm Integrated VS provides a Web service API for remote programmatic access via Web Services Description Language (WSDL) and XML schema files. WSDL is an XML-based language for describing Web services and how to access them. It specifies the location of a service and the operations the service exposes. A client can therefore communicate with the service through the WSDL-provided interfaces, regardless of programming language.

A. *Data Model*

Figure 2 shows the data model used in the integrated VS. LexEVS allows meta concepts, such as UMLS concepts

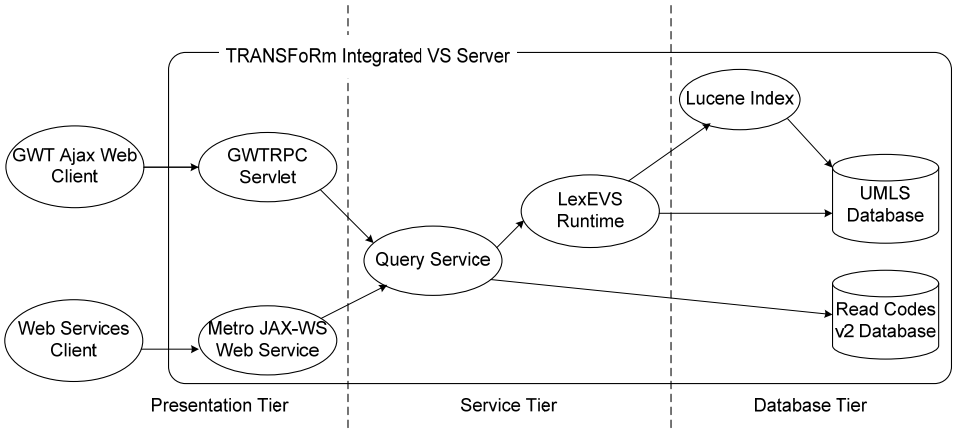

Figure 1. TRANSFoRm VS Server Architecture

 ISBN: 978-1-61208-179-3

to be represented using different formats. UMLS concepts have equivalent concepts in different source vocabularies, such as SNOMED CT, ICD-10. The integrated VS uses the RCV2-to-SNOMED CT map to link the Read Codes V2 to the UMLS concepts. The user is presented with only the mapping from meta concept to source vocabulary concept,

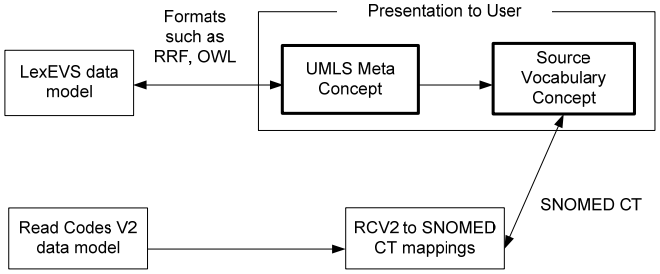

Figure 2. Integrated Vocabulary Service Data Model

depending on the functions requested.

B. Services

Three basic operations are provided as a basic service:

- Search for a term – searches a medical term and returns a list of matching concepts, sorted by relevance.
- Search for a code – searches for a specific code and the function returns the meta concept matching the source vocabulary code searched.
- Return the semantic type for a specific code – the coding system may be specified.

C. Web Interface

The aim for using the vocabulary service is to enable searching by concept, which is a key feature of the eligibility criteria and query formulation that can support researchers to identify eligible patients for their studies. Through an example search, we demonstrate how the vocabulary service Web interface works in helping to map concepts to a wide range of medical terminologies, hence enabling searches across heterogeneous healthcare data. Figure 3 shows a screenshot of the results for an example search for a clinical term ("Type II Diabetes") on the web-based interface of the TRANSFoRm integrated vocabulary service. In this scenario, the user wants to know the RCV2 for Type II diabetes. The term is entered in the search box and the ReadCodes2 option is ticked to include the terminology as it is not provided within UMLS. The elements of the result are described as in Figure 3:

1. UMLS concepts matching the search term are shown in the top left box. The search term "Type II Diabetes" is matched to all the available terminologies (from UMLS and RCV2), and 25 concepts were returned.
2. Each UMLS concept has a unique UMLS code, semantic type, and definition. For instance, the selected concept is "Diabetes Mellitus, Non-Insulin-Dependent", for which further information is displayed on the top right of Figure 3. The UMLS Code "C0011860" provides the unique UMLS identifier for that concept, while the Semantic Type "Disease or Syndrome" indicates the subject category within the categories of UMLS concepts.
3. For the selected concept, its super and sub concepts are rendered on the display, describing the important relationships with other concepts. For instance, a super concept for "Diabetes Mellitus, Non-Insulin-Dependent" is "Diabetes Mellitus", while the sub concept "Diabetes mellitus type 2 in obese" is a more specific concept.
4. Codes from hosted vocabularies (including language and textual description) are also provided for the selected concept. In Figure 3, only the vocabularies in English have been selected. For instance, in RCV2, three terms have been matched to "Type II Diabetes".

VI. Conclusion and Future Work

We have developed an integrated vocabulary service in response to the need for interoperability among medical terminologies used more commonly in Europe. Based on our previous work in the ePCRN project, we have extended the UMLS to support Read Codes Version 2, which are still commonly used in primary care EHR systems in England. The current vocabulary service implementation is based on LexEVS 5.1. A newer version of LexEVS, namely LexEVS 6 has been released, which adds comprehensive support of emerging HL7 terminology service standards [19]. We plan to investigate and migrate to LexEVS 6 to align with the emerging standard. With a vision to support clinical research across Europe in the longer term, we plan to investigate and add more European focused. Additionally, vocabularies related to medication present a big challenge, which definitely needs significant future work.

Acknowledgment

This work was supported in part by the European Commission – DG INFSO (FP7 247787) for the TRANSFoRm project, and the National Institute for Health Research Birmingham and Black Country Comprehensive Local Research Network (NIHR BBC CLRN).

References

[1] J. J. Cimino, "Review paper: coding systems in health care," *Methods of Information in Medicine*, vol. 35, no. 4-5, pp. 273-284, 1996.

 ISBN: 978-1-61208-179-3

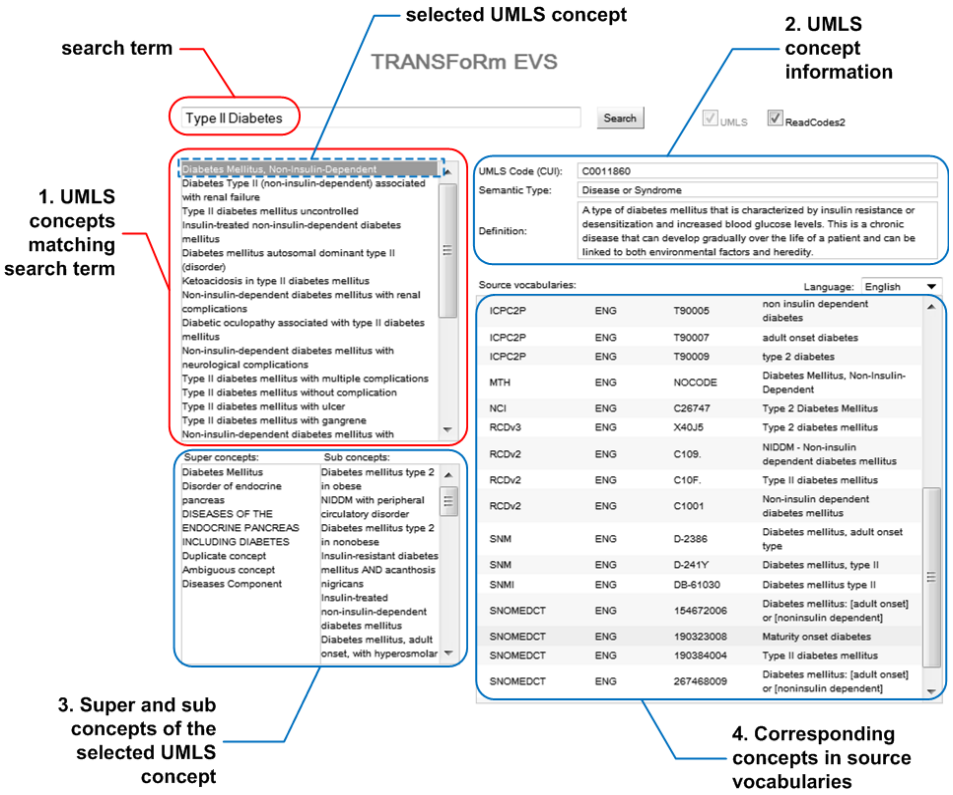

Figure 3. Annotated Screenshot of the Web-based Interface of the TRANSFoRm Integrated Vocabulary Service.

[2] J. S. Rose et al., “Common medical terminology comes of age, Part One: Standard language improves healthcare quality,” *Journal of Healthcare Information Management: JHIM*, vol. 15, no. 3, pp. 307-318, 2001.

[3] J. S. Rose et al., “Common medical terminology comes of age, Part Two: Current code and terminology sets--strengths and weaknesses,” *Journal of Healthcare Information Management: JHIM*, vol. 15, no. 3, pp. 319-330, 2001.

[4] World Health Organization, “International Classification of Diseases (ICD).” [Online]. Available: http://www.who.int/classifications/icd/en/. [Accessed: 19-Sep-2011].

[5] NHS Connecting for Health, “OPCS-4 Classification.” [Online]. Available: http://www.connectingforhealth.nhs.uk/systemsandservices/data/clinicalcoding/codingstandards/opcs4. [Accessed: 19-Sep-2011].

[6] NHS Connecting for Health, “SNOMED CT.” [Online]. Available: http://www.connectingforhealth.nhs.uk/systemsandservices/data/uktc/snomed. [Accessed: 15-Aug-2011].

[7] NHS Connecting for Health, “Read Codes,” *NHS Connecting for Health*, 10-Aug-2011. [Online]. Available: http://www.connectingforhealth.nhs.uk/systemsandservices/data/uktc/readcodes. [Accessed: 10-Aug-2011].

[8] Wonca International Classification Committee, “ICPC-2.” [Online]. Available: http://www.globalfamilydoctor.com/wicc/sensi.html. [Accessed: 19-Sep-2011].

[9] US National Library of Medicine, “Unified Medical Language System (UMLS),” 29-Jul-2009. [Online]. Available: http://www.nlm.nih.gov/research/umls/. [Accessed: 15-Aug-2011].

[10] US National Library of Medicine, “Chapter 2: Metathesaurus,” in *UMLS Reference Manual*, 2009.

[11] US National Library of Medicine, “Chapter 5: Semantic Network,” in *UMLS Reference Manual*, 2009.

[12] ePCRN, “The electronic Patient Care Research Network.” [Online]. Available: http://www.epcrn.org/.

[13] National Cancer Institute, “Enterprise Vocabulary Services (EVS).” [Online]. Available: https://cabig.nci.nih.gov/concepts/EVS/. [Accessed: 15-Aug-2011].

[14] “TRANSFoRm.” [Online]. Available: http://www.transformproject.eu/. [Accessed: 16-Aug-2011].

[15] G. M. Weber et al., “The Shared Health Research Information Network (SHRINE): A Prototype Federated Query Tool for Clinical Data Repositories,” *Journal of the American Medical Informatics Association : JAMIA*, vol. 16, no. 5, pp. 624-630, 2009.

[16] R. G. Côté, P. Jones, R. Apweiler, and H. Hermjakob, “The Ontology Lookup Service, a lightweight cross-platform tool for controlled vocabulary queries,” *BMC Informatics*, vol. 7, p. 97.

[17] NHS Connecting for Health, “UK Terminology Centre.” [Online]. Available: http://www.connectingforhealth.nhs.uk/systemsandservices/data/uktc. [Accessed: 15-Aug-2011].

[18] Apache, “Apache Lucene Project.” [Online]. Available: http://lucene.apache.org/. [Accessed: 19-Sep-2011].

[19] Healthcare Services Specification Project (HSSP), “Common Terminology Services 2.” [Online]. Available: http://hssp.wikispaces.com/cts2. [Accessed: 19-Sep-2011].

 ISBN: 978-1-61208-179-3