# A BioPortal-based Terminology Service for Health Data Interoperability

Lei ZHAO[a,1], Sarah N. LIM CHOI KEUNG[a], and Theodoros N. ARVANITIS[a]
[a] *Institute of Digital Healthcare, WMG, University of Warwick, Coventry, UK*

**Abstract.** A terminology service makes diverse terminologies/ontologies accessible under a uniform interface. The EU TRANSFoRm project built an online terminology service for European primary care research. The service experienced performance limitations during its operation. Based on community feedback, we evaluated alternative solutions and developed a new version of the service. Based on BioPortal's scalable infrastructure, the new service delivers more features with improved performance and reduced maintenance cost. We plan to extend the service to meet Fast Healthcare Interoperability Resources specifications.

**Keywords.** Terminology Service, BioPortal, LexEVS, FHIR

## 1. Introduction

The need for a centralised terminology service becomes critical to querying and analyzing distributed healthcare and biomedical databases in a heterogeneous environment, where a diverse set of code systems and ontologies have been used to encode health data. Healthcare code systems and biomedical ontologies are managed and developed by different organisations and have their own distribution formats. A centralised service provides a uniform interface to browse, search and manage terminologies and concept maps, facilitating terminology interoperability.

### *1.1. Related Work*

Several US organisations and projects have developed software solutions and online services to serve terminology resources on a large scale, most notably Mayo Clinic's LexEVS 0, National Cancer Institute (NCI) Thesaurus and Metathesaurus [1], National Library of Medicine (NLM) Unified Medical Language System (UMLS) [2], and National Center for Biomedical Ontology (NCBO) BioPortal [3]. LexEVS is an open source software package which provides a common terminology model and application programming interface (API) to access a wide range of terminology formats, value sets, and cross-terminology mappings. LexEVS is used by NCI as the software infrastructure to implement its Thesaurus and Metathesaurus services. The NCI Metathesaurus supplements NLM UMLS Metathesaurus with cancer-centric terminologies. UMLS Metathesaurus is a comprehensive multi-purpose and multi-lingual thesaurus that contains millions of biomedical and health related concepts, their synonymous names and their relationships. It combines over 150 classifications, thesauri and lists of controlled terms in the biomedical domain. NCBO BioPortal is the largest repository of biomedical ontologies with over 300 ontologies. It hosts ontologies developed in OWL, OBO and other formats, as well as a large number of medical terminologies from UMLS Metathesaurus. Users can publish their ontologies to BioPortal, submit new versions, browse the ontologies, and access the ontologies and their components through a set of RESTful services, SPARQL and dereferenceable URIs.

The European FP7 project TRANSFoRm (2010-2015) (*www.transformproject.eu*) has developed an integrated vocabulary service (VS) [4], based on LexEVS, with a focus on European primary care systems. The service extracted and loaded a subset of UMLS Metathesaurus, which contained the most important code systems for European primary care such as SNOMED CT, ICPC, ICD10, Clinical Terms v3, LOINC, DICOM, HL7, including their European language variants, as well as genetic ontologies such as Gene Ontology and OMIM. A number of national code systems e.g. UK Read Codes v2 and British National

[1] Corresponding Author. E-mail: Lei.Zhao@warwick.ac.uk

Formulary (BNF) were also integrated. The service provided a web-based terminology browser and web services API to integrate with other TRANSFoRm software e.g. patient cohort identification and eCRFs. The service has experienced certain limitations in performance and scalability during its 4 years operation in TRANSFoRm community. This paper describes our attempt to address some of these limitations in a new version of the service which is based on BioPortal.

## 2. Methods

In order to support federated medical database query and analysis, a terminology service, as a minimal set of functional operations, needs to allow users to browse a classification's subsumption hierarchies, expand a concept to a set of all the subsumed codes, lookup a concept by search text, and translate a concept from one code system to another if precise cross mappings are available. TRANSFoRm VS specifically used UMLS Metathesaurus as the main source of concept maps, and built its infrastructure on LexEVS 6.1. As the size of the content database grew however, expanding a parent concept to its children became an expensive operation in LexEVS due to its relational implementation. The raw database files (MySQL 5.1) in TRANSFoRm VS were about 30GB. Parent-child navigation involves joining tables with millions of rows, resulting in the high latency for a single navigation operation. We selected 10 clinical concepts (e.g. Type 2 diabetes, UMLS C0011860) from TRANSFoRm use cases and tested the latency of LexEVS API call for each concept to return its child concepts. The typical response time varied from 5 seconds to 30 seconds from our experience with the use case-related queries. Another common feature request by TRANSFoRm users is to make autocomplete search suggestions while users type the medical terms into the search field. LexEVS internally builds Lucene index to speed up full-text search for all the text fields when loading a terminology. However the index is not exposed and cannot be leveraged to implement the autocomplete user interface (UI). We evaluated existing services as alternative solutions due to these limitations.

Both NCI Metathesaurus and NLM UMLS Metathesaurus have implemented web services for public access. Nevertheless, these services only allow to retrieve provided contents and cannot be replicated locally. BioPortal, on the other hand, not only has a public repository free to access and publish but also provides virtual appliance for local installation. The latest version of BioPortal organises ontology data in RDF format and uses a triple store as the primary storage. Parent-child navigation is much faster compared to LexEVS (less than one second on average for one level expansion). BioPortal provides a rich set of RESTful APIs which include pagination and search suggestions, convenient for UI development. BioPortal's own web UI is slow and cannot select codes to include in medical database queries. We therefore develop a new terminology browser based on its RESTful service, using modern responsive web design techniques e.g. HTML5, Bootstrap, JQuery, Fancytree and Typeahead.

## 3. Results

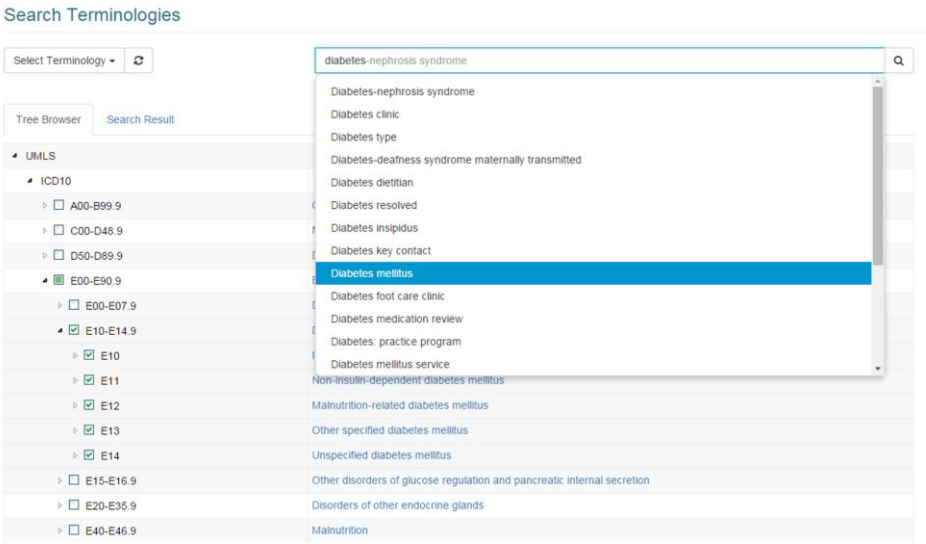


Figure 1: Screenshot of our web-based terminology browser with autocomplete suggestions

Built on BioPortal's REST infrastructure, our new terminology service allows our users to access terminologies in the public repository and license protected contents through a local instance. Many UMLS terminologies we have to manually load into LexEVS are available from BioPortal's public repository, having significantly reduced our maintenance overhead. As shown in [5], users can explore a terminology's subsumption hierarchy in a tree browser, expand a concept, and select all subsumed codes. The selected code set can be saved and included in a database query. Each concept has a hyperlink which will present more information about the concept, including its definition, semantic type, identifier (i.e. Concept Unique Identifiers for UMLS terminologies), and cross-mappings to other code systems. The search field allows users to search concepts by either terms or codes. As user types in the field, a list of suggested terms are presented in a dropdown window to save user input. The search results are listed in a paged table where users can navigate and filter further.

## 4. Discussion

The main focus of our current terminology service is easy access to healthcare code systems in diverse formats and facilitate integration with patient data analytics. However, the service needs to extend its functional scope in order to be used in more healthcare contexts, e.g. to support the functions related to value set as specified by HL7 Fast Healthcare Interoperability Resources (FHIR) [6]. FHIR is a next generation standards framework created by HL7 for sharing healthcare data. Over fifty organisations worldwide use FHIR to exchange information. FHIR makes important differentiation between the concepts of code system and value set. A code system defines a set of codes with meanings. A value set selects a set of codes from one or more code systems to specify which codes can be used in a particular context. A coded element is bound to a value set instead of a code system. A value set can either contain an in-line code system or describe a set of rules to select codes defined in other code systems. These rules can be simply a direct list of codes from a specified version of a code system or can be complex query expressions. FHIR terminology server will expand the value set to a collection of enumerated codes ready to use for data entry or validation. For example, a value set of Myocardial infarction using SNOMED CT is defined as "*is-a SCT::22298006*". The value set will be expanded to the set of all descendants of *22298006* including *22298006* itself. BioPortal provides the service infrastructure to distribute code systems but lacks support for value sets. We plan to investigate the FHIR terminology specification and extend our service to support definition and resolution of value sets, particularly with dynamic expressions.

## 5. Conclusions

The robust and scalable software infrastructure makes BioPortal a good candidate for building online terminology service. We have developed a new version of our terminology service based on BioPortal with improved performance, more UI features and reduced maintenance cost. We plan to extend our service to a full-featured HL7 FHIR terminology server in next version, and also explore BioPortal's SPARQL capability for semantic reasoning in future.

**Acknowledgements:** This work has been supported by the West Midlands Academic Health Science Network, UK.

## References

[1] National Cancer Institute, 'LexEVS - NCI Wiki'. [Online]. Available: https://wiki.nci.nih.gov/display/LexEVS/LexEVS. [Accessed: 10-May-2016].

[2] National Cancer Institute, 'NCI Metathesaurus'. [Online]. Available: https://ncimeta.nci.nih.gov/ncimbrowser/. [Accessed: 10-May-2016].

[3] US National Library of Medicine, 'Unified Medical Language System (UMLS)', 2016. [Online]. Available: http://www.nlm.nih.gov/research/umls/. [Accessed: 10-May-2016].

[4] M. Salvadores, P. R. Alexander, M. A. Musen, and N. F. Noy, 'BioPortal as a Dataset of Linked Biomedical Ontologies and Terminologies in RDF', *Semantic Web*, vol. 4, no. 3, pp. 277–284, 2013.

[5] S. N. Lim Choi Keung, L. Zhao, E. Tyler, T. N. Arvanitis, and F.D.R. Hobbs, 'Integrated Vocabulary Service for Health Data Interoperability', in 4th *International Conference on eHealth, Telemedicine and Social Medicine (eTELEMED 2012)*, Valencia, Spain, 2012, pp. 124–127.

[6] HL7, 'Terminology-service - FHIR v1.0.2'. [Online]. Available: https://www.hl7.org/fhir/terminology-service.html. [Accessed: 10-May-2016].