A RDF-base Normalized Model for Biomedical Lexical Grid

Cui Tao, Jyotishman Pathak, Harold R. Solbrig, Wei-Qi Wei, and Christopher G. Chute
Division of Biomedical Statistics and Informatics
Mayo Clinic, College of Medicine, Rochester, MN, 55905
{tao.cui, pathak.jyotishman, solbrig.harold, wei.weiqi, chute}@mayo.edu

ABSTRACT

The Lexical Grid (LexGrid) project is an on-going community-driven initiative coordinated by the Mayo Clinic Division of Biomedical Statistics and Informatics. It provides a common terminology model to represent multiple vocabulary and ontology sources as well as a scalable and robust API for accessing such information. While successfully used and adopted in the biomedical and clinical community, LexGrid model now needs to be aligned with emerging Semantic Web standards and specifications. This paper introduces the LexRDF model, which maps the LexGrid model elements to corresponding constructs in W3C specifications such as RDF, OWL, and SKOS. With LexRDF, the terminological information represent in LexGrid can be translated to RDF triples, and therefore allowing LexGrid to leverage standard tools and technologies such as SPARQL and RDF triple stores.

1. INTRODUCTION

The Lexical Grid (LexGrid) [1] project, coordinated by the Mayo Clinic Division of Biomedical Statistics and Informatics, provides a semantic foundation upon which multiple APIs can be developed that support consistent searching, navigation and cross terminology traversal. These open-source tools are used in a variety of projects such as the NCI Cancer Biomedical Informatics Grid, the National Center for Biomedical Ontology, the Biomedical Grid Terminology project, and the World Health Organization International Classification of Diseases (ICD-11) development process. LexGrid hosts a wide variety of terminologies and ontologies including ICD-9-CM, the Gene Ontology, the HL7 Version 3 vocabulary, and SNOMED-CT. LexGrid can also represent complete NLM Unified Medical Language System, which currently includes over 100 source terminologies. Our experience in developing and deploying the LexGrid technology provides an unparalleled basis for using ontologies to represent patient and clinical trial information, thereby enabling semantic information retrieval.

A valuable augmentation to LexGrid is the adoption of Semantic Web technologies. We have designed and implemented a LexOWL system [2], which provides an API bridge between LexEVS and the OWL API. LexOWL represents information in LexOWL using OWL DL specifications. It mainly focuses on mapping the semantic representation and the hierarchical information whereas some of the lexical elements in LexGrid were not covered. Lexical information such as synonyms, acronyms, definitions, and comments, however, serves very important roles in biomedical terminologies and vocabularies. We therefore introduce the LexRDF project, which extends the LexOWL efforts and provides an unified RDF-based model for both semantic and lexical information in the biomedical domain. We discuss the details of our exercise in the remainder of this paper.

2. LEXRDF MAPPING SPECIFICATION

For each LexGrid element, we searched for equivalent constructs or axioms in the W3C recommendation standards such as OWL, RDF, and SKOS, as well as Dublin Core. In the case where appropriate mapping is lack from the standard name spaces, we proposed new constructs in the LexRDF name space. These extensions will be proposed to appropriate W3C committee for future recommendation.

Ontology Information Mapping LexGrid comprises various lexical elements describing meta-data about an ontology. These include provenance (source (*dc:source*), copyright (*dc:right*), version (*owl:verionInfo*)), name (*dc:title, rdf:label*), URI, and language (*dc:language*). LexRDF successfully identified mappings for all the LexGrid ontology-information components except one: *approxNumConcepts*, which indicates the total number of ontological entities present in an ontology. This attribute was intended as a hint to service components. Since this information can be inferred from the ontology itself, we chose to exclude it from this mapping.

Entity Mapping A LexGrid *entity* represents any resource in a terminology or ontology. Figure 1 shows the syntax graph of the LexGrid entity components. A dashed arrow from element A to element B indicates that A is an instance of B. An arrow with a clear arrowhead from A to B indicates that A is a subclass of B. We use *lg* to represent the LexGrid name space. LexGrid has defined *lg:concept*, *lg:association*, and *lg:instance* as subclasses of *lg:entity*. LexRDF maps *lg:concept* to *owl:Class*, meaning that *lg:concept* inherits the definition of *owl:Class*—both an instance and a subclass of *rdfs:Class*. The construct *lg:association* is equivalent to the union of *owl:ObjectProperty* and *owl:DatatypeProperty*, which are both instances of *rdfs:Class* and subclasses of *rdf:Property*. The construct *lg:instance* is a general holder of OWL individuals which are instances of OWL classes. LexRDF uses *owl:Thing* to declare a LexGrid instance in RDF triple representation. LexRDF also maps *lg:entity* to *skos:Concept*, which is defined as an instance of *owl:Class*. This mapping specification preserves the original LexGrid definition without introducing any contradictions of definition in the standard name spaces.

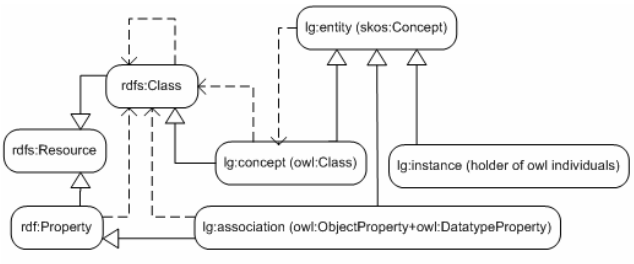

Figure 1: LexRDF Entity Mapping Specification

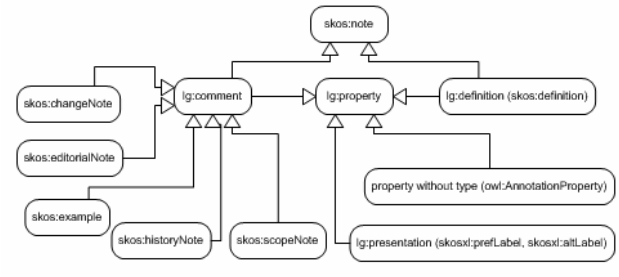

Figure 2: LexRDF Property Mapping Specification

Property Mapping Every instance of a LexGrid entities is associated with a set of properties, which are analogous to annotation properties in OWL. Figure 2 shows the LexRDF property definition overview. Each *lg:property* could have an optional type (comment, presentation, or definition). Each *lg:comment* and *lg:definition* has a *isPreferred* flag which indicates whether it was "preferred" in the given language and context. When no type is specified, a *property* construct is mapped to an *owl:AnnotationProperty*. The *lg:comment* construct is a super property of *skos:changeNote*, *skos:editorialNote*, *skos:example*, *skos:historyNote*, and *skos:scopeNote*. The *lg:presentation* is mapped to *skos:prefLabel* when the *isPreferred* flag is set to *true* and to *skos:altLabel* otherwise. The *lg:definition* is mapped to *skos:definition*. LexRDF uses a *LexRDF:isPreferred* construct for the *lg:isPrefered* flag which is used to reify whether a definition is preferred or not. LexGrid uses *lg:propertyLink* to define relationships between two properties. LexRDF defines each *lg:propertyLink* instance as an *owl:ObjectProperty* and uses RDF reification to define the relations.

Association Mapping LexGrid uses associations to represent relationships between entities. The association definition may also further define the nature of the relationship such as forward and inverse names, transitivity, symmetry, reflexivity, and etc. LexRDF used OWL properties and assertions to represent all of them except *reverseName* and *isAntiTransitive*. LexRDF uses a new construct *LexRDF:reverseName* to represent the name of the association on the reverse direction when a target to source side of the association is meaningful. *LexRDF:isAntiTransitive* is used to represent a property that is not transitive. A *LexRDF:reverseName* represents the name of the association on the reverse direction when a target to source side of the association is meaningful. In addition, an association could be modified by using *lg:associationQualification*. For example, one can define an association *Poland anomaly* $\frac{HAS\_CLINICAL\_SIGN}{Very\ frequent}$ *Dextrocardia*, where `HAS_CLINICAL_SIGN` is the association name and `Very frequent` is the qualifier indicates how frequently the disease has the symptom. LexRDF defines a new OWL AnnotationProperty, *LexRDF:associationQualifier*, and use RDF reification to represent the association qualifiers.

3. DISCUSSION AND CONCLUSION

In this paper, we introduced the LexRDF mapping specification with respective to ontology information, entity, property, and association. LexRDF has successfully mapped 37 out of 45 LexGrid elements, achieving a very high degree of reusability.

We have also discovered some interesting problems where the W3C standard language cannot fully represent our needs in LexGrid. Firstly, LexGrid has a common superclass *lg:property* for comments, presentations, and definitions. In LexRDF, we use *skos:prefLabel* and *skos:altLabel* to represent *lg:presentation*; we use *skos:definition* to represent *lg:definition*; and we use a subset of *skos:note* to represent *lg:comment*. SKOS does not define an common ancestor for all these properties (labels, notes, and definition), which can represent generic properties. We have a similar problem with *lg:comment*. Currently it is mapped to a a set of sub-properties of *skos:note*, but a generic comment class is also preferred. Secondly, SKOS has defined *prefLabel* and *altLabel*, although no such constructs are provided for "definitions". Currently, we are using *LexRDF:isPreferred* as a tag to specify whether a definition is preferred or not. Akin to prefLabel and altLabel, our objective is to propose prefDefinition and altDefinition to the SKOS committee to be introduced in the future specification. Thirdly, LexGrid provides an option for modify an association instance by adding association qualifiers. We have found this to be needed in the clinical domain and believe that it is an important requirement to be considered by the appropriate W3C standards group. In addition, we have a requirement to describe relations among properties. By doing so we can assert that a particular label is an acronym or another, or that a given definition is a literal translation of the same definition in another language. SKOS provides *skosxl:labelRelation* that can represent relations between two labels. The property *skosxl:labelRelation*, however, is defined as a symmetric property with domain and range as *skosxl:Label*. These limitations restrict us from using it for our *lg:propertyLink*. We would like to propose a more general property to represent *lg:propertyLink*.

In summary, this paper introduced our on-going work to map the elements from the LexGrid model to various Semantic Web standards. Although mostly successful, we have identified several limitations of the existing W3C specifications that warrant broader community engagement.

4. REFERENCES

[1] J. Pathak, H. Solbrig, J. Buntrock, T. Johnson, and C. 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.

[2] C. Tao, J. Pathak, H. Solbrig, and C. Chute. LexOWL: A bridge from LexGrid to OWL. In *Proceedings of the First International Conference of Biomedical Ontology (ICBO 09)*, pages 131–134, Buffalo, New York, July 2009.