Proposed SKOS Extensions for BioPortal Terminology Services

Cui Tao[†], Natalya F. Noy[‡], Harold R. Solbrig[†],
Nigam H. Shah[‡], Mark A. Musen[‡], and Christopher G. Chute[†]

[†] Division of Biomedical Statistics and Informatics,
Mayo Clinic College of Medicine, Rochester, MN
[‡]Stanford Center for Biomedical Informatics Research,
Stanford University, Stanford, CA

Abstract. The National Center for Biomedical Ontology (NCBO) BioPortal provides common access for browsing and querying a large set of ontologies that are commonly used in biomedical communities. One of our missions is to align lexical features (i.e., textual definitions) that are commonly used in these ontologies across different representation formats with standard tags and to represent them in a standard way to the users. The Simple Knowledge Organization System (SKOS) is a recommendation of the World-Wide-Web Consortium (W3C) for a common data model for sharing and linking knowledge organization systems on the Semantic Web. The BioPortal is in the process of adopting SKOS in the backend representation for its content. During this process, we discovered that there exists a set of commonly-used lexical features shared by the biomedical ontologies that SKOS does not yet represent. In this paper, we discuss our proposed SKOS extensions to cover this set of commonly used lexical features, the rationales, and the detailed description of each proposed construct.

1 Introduction

The use of ontologies in the biomedical domain has accelerated dramatically over the last decade [5–7]. . Biomedical ontologies provide the essential domain knowledge needed for different purposes such as data integration, data sharing, semantic annotation, information extraction, and natural language processing. The National Center for Biomedical Ontology (NCBO) BioPortal [4] is an open-source repository of ontologies, terminologies, and thesauri of importance in biomedicine. As of Oct. 2011, BioPortal is hosting 296 biomedical ontologies, including more than 5 million terms. These ontologies cover domain knowledge from basic science, clinical terms, to translational studies.

One important goal of BioPortal is to provide a common access for browsing and querying the shared ontologies that are commonly used in the biomedical communities [10]. BioPortal is designed for harmonizing these resources in a uniform representation and delivering this content in a consistent, standardized fashion for use in biomedical, clinical, and research applications.

This goal coincides with one of the missions of the Semantic Web community: creating technologies to share heterogeneous content and to distribute it over the Web. The World Wide Web Consortium (W3C) has produced recommendations for a stack of Semantic Web languages for structured data, terminologies, and ontologies, including the Resource Description Framework (RDF) for representing data on the Web [13], the SPARQL language for querying RDF data [16], the Web Ontology Language (OWL2) for representing ontologies on the Web [12], the Simple Knowledge Organization System (SKOS) [14], and SKOS eXtension for Labels (SKOS-XL) [15] for representing terminologies and thesauri.

One necessary step toward this goal is to represent commonly existing properties in the biomedical domain in a standard, Semantic-Web compliant way. BioPortal is a host for different kinds of resources: formal ontologies, terminologies, thesauri, and so on. Formal ontologies are mapped to a formal semantics, thus allowing well defined and rather powerful inferences, and use a precise and rigorous set of constructs to represent their content. Thesauri and other knowledge-organization systems, however, have less emphasis on formal structure and focus more on lexical components such as synonyms, multilingual definitions, examples, comments, cross references. The latest SKOS specification describes a standard set of tags for most of these constructs and has been adopted by many organizations, including the Library of Congress [8], NASA [9], and the United National Food and Agriculture Organization [3].

BioPortal assists authors in the identification of the related constructs during terminology submission. For example, when submitting their terminology or ontology, the authors indicate which property in their ontology stores synonyms, authors of each term, definitions of a term and so on.We then relate the corresponding lexical components to the SKOS standard, which, in combination with a structured terminology model, can serve as a baseline for the representation of the next generation of terminological resources. During this process, we have identified several properties and constructs that are missing from SKOS, but are critical to biomedical terminologies, including:

- **multiple definitions** - resources may be accompanied by multiple (textual definitions), derived from multiple sources. The resource authors need to indicate which of these definitions is preferred in a given language or context.
- **flavors of annotation** - resource annotations include information about what was changed and when, the current status of the entry and targeted directions, when and where the resource is applicable, etc. Applications need to be able to differentiate these various flavors of notes as some are applicable to end user situations while others are only of value in editorial or historical contexts. In addition, applications need to be able to clearly differentiate definitions from comments from examples, as when and where each of these is used is different.
- **lexical semantics** - as ontologies begin to be consumed in the NLP space, it becomes important to be able to identify the various forms of labels - the noun form, adjectival form, singular, plural as well as the label derivation - acronym, abbreviation, eponym, etc. In addition, the introduction of multi-

lingual lexical systems creates a need to be able to identify the derivation of specific labels, notes, etc. The fact that a German definition is a literal translation of a corresponding English definition is important when attempting to understand the intended meaning of a term. Similarly, an acronym or abbreviation needs to be associated with the language and term that it is an acronym and abbreviation for.

Building on our experience with LexGrid [1], which provides common terminology services for biomedical terminologies, as well as standard ontology representation guidelines [17], we propose a set of constructs to cover these common lexical properties in the biomedical domain.

2 SKOS Extensions

New Construct	**Description**	**Comment**
skosxl_plus:Comment	Notes for a resource excluding examples and definitions	see Figure 1
skosxl_plus:Note	Similar to skosxl:Label but works with skos:noteRelation	
skosxl_plus:Source	Reified provenance details.	
skos_plus:prefDefinition	The preferred definition of a resource given a specific context or language	subproperty of *skos:definition*
skos_plus:altDefinition	Alternative definitions of a resource	subproperty of *skos:definition*
skos_plus:designationType	Discriminant that determines the type of a particular description	synonyms, acronyms, short names, etc
skosxl_plus:noteRelation	Relationships between two lexical properties	a super property of *skosxl:labelRelation*, domain and range are skosxl:Note
skosxl:literalForm		Expand its domain and range to cover skosxl:Note

Table 1. SKOS Extension Summary

Comments and Notes "Notes are used to provide information relating to SKOS concepts." [14]. The current SKOS specification identifies 6 subclasses, *skos:changeNote, skos:definition, skos:editorialNote, skos:example, skos:historyNote, skos:scopeNote.* Applications that consume terminology services need to be able

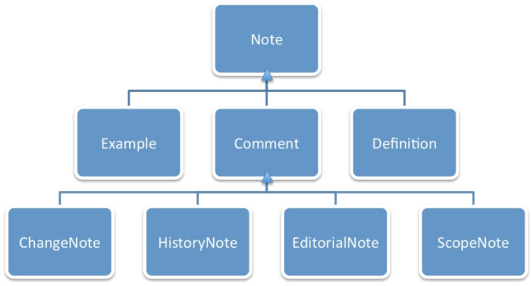

Fig. 1. Proposed Hierarchy of skos:note

to distinguish *definitions*, *examples* from other forms of notes as definitions are used to clarify the meaning of a concept or term, examples help to refine it, while other types of comments provide a variety of secondary purposes. We propose the addition of one additional class, *skos_plus:Comment*, which provides this primary distinction.[1]

Preferred Definition Textual definitions are becoming increasingly common in biomedical ontologies. They provide an important link between the computational formalism of the ontology and the (intended) meaning of the term, class or property. A given class may be accompanied by several definitions that are drawn from multiple reference sources. Definitions may be provided in different languages and may be applicable in different contexts. SKOS currently defines *prefLabel* and *altLabel*, but no analogous distinctions are provided for definitions. To remedy this, we propose the creation of *prefDefinition* and *altDefinition*, as subproperties of *skos:defintion*, with the assertion that each concept may have at most one preferred definition given a language.

Type of Descriptions The use and a meaning of a term label can require additional lexical semantics. It is often important to know whether a label is a noun form, adjectival form, whether it is singular or plural, whether it is a common acronym, an abbreviation or a full form. The existing SKOS specification provides some ability to make these distinctions, as labels can be marked as "preferred", "alternate" or "hidden", which indicates whether they are the primary identifiers, synonyms or deprecated forms such as common misspellings, deprecated acronyms, etc. It doesn't, however, provide the sort of granularity many sophisticated applications need. To address this issue, we propose a new construct, *designationType*, to define the type of a description of a given resource. Typical types would include noun, adjective, acronym, eponym, and ab-

[1] The reason that the existing tag, *rdfs:comment* was not used is that it is often heavily overloaded in existing ontologies— being used to carry definitions, examples, various flavors of comments, etc.

breviation, which we propose can be drawn from the ISO TC37 Data Category Registry[2].

Figure 2 shows an example using *designationType*. Here we use the SKOS-XL data model to represent that concept <C1> has an alternative label "FAO" which serves the role of *acronym*.[2]

```
<C1> skosxl:altLabel <L>.
<L>  rdf:type skosxl:Label;
     skosxl:literalForm "FAO";
     skos_plus:designationType ISOcat:acronym.

ISOCat:acronym rdfs:subClass ISOCat:abbreviatedForm.
```

Fig. 2. An Example of Designation Type

Relations between Lexical Properties SKOS-XL defines *skosxl:labelRelation* to represent binary relations between instances of the class *skosxl:Label*, which allows assertions such as the label, "FAO" is an *ISOcat:acronymFor* "Food and Agriculture Organization". This pattern, however, does not extend to *definition-definition* or *comment-definition* relationships, etc. To address this issue, we propose a new construct, *noteRelation*, which can be used for representing relations between not only two labels, but also any two lexical properties, such as definitions, comments, and examples.

We also propose a new class—*skosxl_plus:Note*—as the domain and range of *noteRelation*. Similar to *designationType*, the types of these relations could be adopted from the ISO TC37 Data Category Registry. Figure 3 provides an example of a *noteRelation*.[3]

SKOS-XL has defined *skosxl:prefLabel*, *skosxl:altLabel*, and *skosxl:hiddenLabel* as instances of the class *skosxl:Label*. These new constructs are necessary because *skos:prefLabel, skos:altLabel* and *skos:hiddenLabel* are sub-properties of *rdfs:label*, and therefore inherit its range of *rdfs:Literal* vs. the class *skosxl:Lablel*. *skos:note* and its sub properties are already defined as instances of *owl:AnnotationProperty*, which means that the target does not have the same restrictions and we can use them directly when connecting a concept (e.g., *<C1>* to note (e.g., *<E1>* or *<E2>* in Figure 3).

[2] ISOCat does not follow the convention of capitalizing the first letter of class names - all identifiers begin with a lower case letter. We have included a type definition in the examples to help clarify this.

[3] Note that ISO TC37 does not, at the moment, appear to provide a property "translation of". We will need to address the difference between the verb / noun (property / class) with this organization.

```
<C1> skos:editoralNote <E1>.
<E1> rdf:type skosxl_plus:Note;
     skosxl:literalForm "needs to be updated later"@en.
<C1> skos:editoralNote <E2>.
<E2> rdf:type skosxl_plus:Note;
     skosxl:literalForm  "moeten later worden bijgewerkt"@de.
<E2> ISOcat:translation <E1>.

 ISOcat:translation rdf:subProperty skosxl_plus:noteRelation.
```

Fig. 3. An Example of Property Relation

Provenance Annotations We also need the ability to provide provenance information for the lexical resources in an ontology.[4] As an example, the OBO [11] allow the specification of the source roles and documents of definitions. Figure 4 shows the definition of "reproduction" in the Gene Ontology.

```
def: "The production by an organism of new individuals that contain some
portion of their genetic material inherited from that organism."
[GOC:go_curators, ISBN:0198506732 "Oxford Dictionary of
Biochemistry and Molecular Biology"]

 GO:0000003 rdf:type skos:Concept;
            skos_plus:prefDefinition <D1>.
<D1> rdf:type skosxl_plus:prefDefinition;
     skosxl:literalForm "The production by an organism of new ...";
     dc:source <S1>.
<S1> rdf:type skosxl_plus:Source;
     crdf:role GOC:go_curators;
     crdf:sourceDocument URN:ISBN:0198506732;
     crdf:sourceDescription "Oxford Dictionary of Biochemistry
                    and Molecular Biology".
```

Fig. 4. Definition of "reproduction" in OBO and RDF

In this case, we need a general class for source to allow further annotations of the instances of the source. Here we propose a new class, *skosxl_plus:Source*, to represent the overall source information for reification purposes. Note that we use the *crdf* namespace for tags that describe further source information. We do not propose that these tags are included in the SKOS extension because we believe that they are out of scope for SKOS. Note that the "<S1>" tag is used in this example because this represents a description of URN:ISBN:019506732 *by* the GO curators rather than the resource itself.

[4] Fine grained provenance is also required in the *formal* or "semantic" assertions, but that is out of the scope of this document.

3 Discussion

The definition of a logic formalism consists of two sections. The first section defines *syntax* of the logic—describing the set of symbols that are valid in the logic and the set of possible transformations that may be applied to them. The second section describes the *semantics*—a set of rules that describe how the various logic symbols and operations correspond with equivalent structures the (or some) "real world". Most of the Ontology efforts to date have been focused on the first section—the representation and manipulation of the formal symbols. The *semantics* of an ontology is represented by the *lexical* aspects, and is the part that allows human beings to map the symbols from and to their intended referents. An ontology that fails to maintain this second, lexical component becomes nothing more than an interesting mathematical artefact—a set of logically consistent symbols that may or may not apply to the real world.

The SKOS [14] specification has formed an excellent starting point for formalizing the lexical aspects of an ontology. Today, many of the ontologies that pay attention to the lexical components at all use ontology specific tags for these components, or, in some situations, embed the meanings of these tags lexically inside *rdfs:comment* constructs. The conversion of these idiosyncratic approaches into one that is semantically consistent requires a sufficiently robust target that the majority of the information that needs conversion can migrated without significant loss.

The proposals that we discuss in this document suggest one such approach. It should be noted, however, that this approach depends on either (a) semantics that do not formally include the distinction made by OWL between *AnnotationProperty* and *ObjectProperty* and *DataProperty* or (b) the use of OWL 2. The reason for this is because the OWL 1 series of specifications restrict AnnotationProperties to the point that they cannot be used as first class resources without lapsing into the computational marshland of OWL Full. This restriction has driven many existing ontologies to resort to a variety of tricks and embedding schemes to allow them to carry the information they need while, at the same time, taking advantage of the formal classifications that are available.

The proposals presented here seem to work in principal. To be truly useful, however, more standardization work still remains. While the ISOCat Data Category Registry [2] appears to carry many of the markup properties that are needed, there are still issues about format, rights and completeness that need to be resolved. The specific provenance tags such *sourceDocument*, *sourceDescription* need standardization as well before the work done here will begin to bear significant fruit.

4 Concluding Remarks

In this paper, we discussed our proposed SKOS extension for representing common lexical information in BioPortal ontologies. One important goal of BioPortal is to provide a common access for browsing and querying the shared ontologies

that are commonly used in the biomedical communities. We believe these additional properties will improve interoperability among biomedical ontologies, and therefore enhance the common service provided by BioPortal.

These proposed extensions were based on our more than a decade experience with common terminology service in the biomedical domain. We believe that SKOS has already provided a reasonably good foundation for common lexical information representation. The proposed tags will be a useful addition to SKOS for harmonizing biomedical ontologies.

Acknowledgement This research is partially supported by the National Center for Biomedical Ontology (NCBO) under the NIH Grant #N01-HG04028 and the NSF under Grant #0937060 to the Computing Research Association for the CIFellows Project.

References

1. LexGrid: The Lexical Grid. `https://cabig-kc.nci.nih.gov/Vocab/KC/index.php/LexGrid`, 2009.
2. ISOcat - data category registry, October 2011. http://www.isocat.org.
3. AGROVOC Thesaurus, Food and Agriculture Organization of the United Nations (FAO). `http://www.fao.org/agrovoc`.
4. NCBO Bioportal. `http://bioportal.bioontology.org/`.
5. O. Bodenreider. Biomedical Ontologies in Action: Role in Knowledge Management, Data Integration and Decision Support. In A. Geissbuhler and C. Kulikowski, editors, *IMIA Yearbook of Medical Informatics*, volume 47, pages 67–79. International Medical Informatics Association, 2008.
6. C.G. Chute. The Copernican Era of Healthcare Terminology: A Re-Centering of Health Information Systems. In *AMIA Annual Symposium*, pages 68–73, 1998.
7. C.G. Chute. Clinical Classification and Terminology: Some History and Current Observations. *JAMIA*, 7(3):298–303, 2000.
8. Library of congress subject headings, the library of congress cataloging distribution service. `http://www.loc.gov/cds/lcsh.html`.
9. NASA Taxonomy. `http://nasataxonomy.jpl.nasa.gov/fordevelopers/`.
10. N.F. Noy, N.H. Shah, P.L. Whetzel, B. Dai, M. Dorf, N. Griffith, C. Jonquet, D.L. Rubin, M.D. Storey, C.G. Chute, and M.A. Musen. Bioportal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Research*, 37:170–173, 2009.
11. The open biomedical ontologies. `http://www.obofoundry.org/`.
12. OWL 2 web ontology language structural specification and functional-style syntax. `http://www.w3.org/TR/owl2-syntax/`.
13. The RDF vocabulary. `http://www.w3.org/1999/02/22-rdf-syntax-ns`.
14. SKOS vocabulary. `http://www.w3.org/2006/07/SWD/SKOS/reference/20090315/skos.rdf`.
15. SKOS XL vocabulary. `http://www.w3.org/2006/07/SWD/SKOS/reference/20090315/skos-xl.rdf`.
16. SPARQL Query Language for RDF. `www.w3.org/TR/rdf-sparql-query/`.
17. C. Tao, J. Pathak, H.R. Solbrig, W. Wei, and C.G. Chute. Common terminology guidelines for representing biomedical ontologies in semantic web notations. *Journal of Biomedical Informaitcs*. submitted.