Mash-up of LexWiki and Web-Protégé for Distributed Authoring of Large-Scale Biomedical Terminologies

Guoqian Jiang*, Harold R. Solbrig, Christopher G. Chute
Division of Biomedical Statistics and Informatics,
Mayo Clinic College of Medicine, Rochester, MN, 55906

Abstract. In this presentation, we propose a framework for distributed authoring of large-scale biomedical ter-minologies, which comprises three modules: a structured proposal creation module using semantic wiki machinery, a proposal harvesting module using a formal ontology editing platform and a backend module with a formal terminology model. We devel-oped a prototype of the framework based on a real world use case through a mash-up of LexWiki and Web-Protégé.

1. Introduction

Complete, well-defined, high quality ontologies have been regarded as essential for enabling global interoperability and realizing the vision of the semantic web [1, 2]. However, the weakness of the traditional ontology engineering has been recognized as that 1) the ontology evolution is not under the full control of the ontology user community; and 2) the communication between ontology engineers and domain experts is weak [3]. Furthermore, it is important to note that ontologies are not just formal representations of a domain, but much more community contracts about such formal representations, i.e. must be able to reflect the community consensus at any point in time [4]. With the advent of Web 2.0 technologies and applications, it is natural that both researchers and practitioners are now beginning to explore how the power of creating web content in a social environment can be used to acquire, formalize and structure knowledge [5].

Wiki as a collaborative system provides tools for user participation into common tasks within a community, e.g., discussion pages. Combined with Semantic Web technology, semantic wiki provides the ability to capture (by humans), store and later identify (by machines) further meta-information or metadata about those articles and hyperlinks, as well as their relations [6]. This has been demonstrated as an appropriate platform for knowledge engineering methods to work on the different levels of the continuum [7]. Open content development based on semantic wiki technology is considered crucial for two reasons: 1) it ensures by design community acceptance and content relevant to community's needs; 2) it supports rapid publication cycles. In traditional ontology authoring, ontologists must overcome dual challenges as they build, refine and

maintain an abstract terminology model. According to [8], the ontologists not only need to understand the logical formalism, but they must also find effective ways to explain the model to subject matter experts from many end user communities, including its formal and operational properties. We consider that the adoption of semantic wiki machinery may provide a solution for dividing these challenges, leaving formal refinement as a separate step in an iterative authoring process.

In this presentation, we propose a framework for distributed authoring of large-scale biomedical terminologies, and develop a prototype of the framework based on the requirements elicited from a real world use case through a mash-up of LexWiki and Web-Protégé.

2. Background

2.1. Web-Protégé

A number of collaboration features are being developed for the traditional ontology editors. Notably, Collaborative Protégé is developed as an extension of the existing multi-user Protégé system and supports the association of annotations to any component of the ontology or to any change that occurs in the ontology [9]. The core component of the collaborative Protégé is the integration of a *Change and Annotation Ontology* (CHAO) [10] into the system and the CHAO provides the basis for annotation of ontology elements, such as classes, properties, individuals and annotation of ontology changes, such as class creation, deletion, renaming, etc. To better support the collaborative development process in a web environment, Web-Protégé has been developed as a web front-end for the Collaborative Protégé server.

2.2. LexWiki

LexWiki is a collaborative effort led by Mayo Clinic for development of a collaborative authoring platform for large-scale biomedical terminologies [11]. The LexWiki environment based on Semantic MediaWiki [12] enables the wider community to make both structured and unstructured proposals on the definitions of classes and property values, suggest new values, and corrections to the current ones. LexWiki currently is at the core of community-based development of Biomedical Grid Terminology (BiomedGT) [13] and has also been successfully implemented to support the Common Terminology Criteria for Adverse Events (CTCAE) revision project [14] and the CDISC Shared Health and Research Electronic Library (CSHARE) project [15].

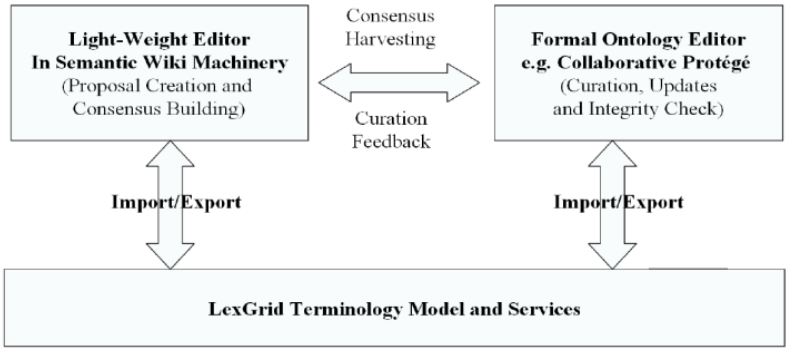

Figure 1. The LexWiki framework for distributed terminology authoring

2.3. LexGrid

The LexGrid Model [16] is a community proposal for standard storage of controlled vocabularies and ontologies, adopted by HL7, NCI, and the National Center for Biomedical Ontology (NCBO). The open-specification LexGrid Model defines how vocabularies should be formatted and represented programmatically, and is intended to be flexible enough to accurately represent a wide variety of vocabularies and other lexically-based resources. In upper class level, LexGrid Model defines *codingScheme*, *codedEntry (i.e. concept)* and *association (i.e. relation)*. Each *codingScheme* has a set of *concepts* associated with it. A concept is described by a set of lexical properties (e.g. *conceptCode*, *definition*, *presentation*, etc.), a set of non-lexical properties (*conceptProperty*) and a set of associations.

3. System Architecture

Fig. 1 depicts our framework for a distributed authoring system comprising three modules: 1) a light-weight editor in semantic wiki machinery for proposal creation and consensus building, 2) a formal ontology editor for change commitment, formalism rendering and consistency checking, and 3) a terminology service module based on the LexGrid formal terminology model.

Our framework focuses on a proposal based workflow process, which is hypothesized as follows.

1) The target terminology is rendered into the LexGrid model;
2) The contents of target terminology are then loaded into semantic wiki machinery;
3) Subject matter experts (SMEs) from different subdomains review the contents and make proposals using the semantic wiki machinery;
4) The change set from the proposals are detected and loaded into Collaborative Protégé;

5) Curators commit the changes to the target terminology and check the formalism and consistency in Collaborative Protégé.
6) The updated contents are rendered into the LexGrid model and re-enter into the circle. The functional components of this framework are described as follows.

3.1. Content representation component

At the concept level, each concept is represented as a page in a category namespace in wiki. To keep both human readability and the unique nature of a wiki page, we combined the preferred name and the concept code of a concept together as a wiki page name. This avoids the conflicts when two different concepts have the same preferred names as the wiki requires the page name to be unique. We proposed a set of templates to represent the core semantics of LexWiki contents of each concept. These templates provide a common mechanism for importing terminological data. Different wiki implementations can assign different renderings to these components, as long as the underlying semantic mapping remains constant. The data elements are derived from the LexGrid terminology model, defining the semantic core of LexWiki, so that different templates can be used as long as they are mapped to the LexGrid model.

3.2. Terminology import component

We developed a Protégé Tab Plug-in that supports terminology import into the Semantic MediaWiki (SMW). LexWiki Tab employs the existing concept-oriented terminology service LexGrid API and transforms the concepts and their attributes from a LexGrid node into LexWiki templates. The Java Wiki Bot Framework API [17] was used to develop an export/import interface between Protégé and the SMW. The LexWiki Tab supports exporting a single concept into the wiki and also supports exporting a group of concepts that are a subset of a coding scheme or all concepts of a coding scheme into the wiki.

In addition, we developed a mapping schema between the constructs of a coding scheme in the LexGrid model and the underlying knowledge model of SMW. Basically, each concept in a coding scheme is mapped with a SMW page in category namespace. The properties and associations of each concept are mapped with the SMW properties. The hierarchical concept relationships are mapped onto the SMW category-subcategorization schema. Using this schema, a coding scheme can be easily represented as a hierarchical tree within a wiki, a browsing view familiar to the medical terminology community.

3.3. Proposal consensus building component

In the course of terminology authoring and editing, we assert that there are two types of proposals that are potentially created by domain experts. The first type is free text comment, which we label “unstructured proposal”. The second type is the structured proposal, in which we developed a mechanism to allow the semantic forms in the LexWiki editor to clone the contents of an original category page where users can change virtually all the clone contents.

Through this mechanism, multiple users may create their own proposals without physical data conflicts, since each operates within their own cloned proposal space.

Each proposal thereby begins as a copy of original concept, retaining semantic annotations (i.e. structured data elements). Similarly, semantic change annotations may leverage semantic queries for change tracking or rendered into OWL/RDF format for change detection and export to the Protégé formal editing platform.

3.4. Change set detection component

After a proposal consensus is achieved in LexWiki, the proposal is harvested into Collaborative Protégé, a distributed ontology editing platform, for change commitment, formal rendering in OWL and consistency checking. Based on the semantic annotations of articles, SMW generates machine-readable documents in OWL/RDF format. This RDF specification may render the semantic annotations of all proposal articles of this category. For change set detection, the RDF specifications for each concept are retrieved and parsed using the existing Protégé OWL API. A change set detection algorithm was developed to compare the semantic annotations between a proposal and its original concept. After the change set are detected from a proposal, we represented the change set in the CHAO ontology through generating the CHAO instances.

3.5. Workflow management component

To adapt the proposal mechanism to different use cases, a workflow management component was designed as follows. For each proposal, a set of workflow curation statuses is defined and used to indicate the current curation status. For instances, the curation status "New" may indicate the proposal is a newly created one; "InProgress" may indicate the proposal is under processing by the curators; "Accepted" may indicate the proposal is accepted by the curators; and "Rejected" may indicate the proposal is rejected by the curators. The set of workflow curation status can be adjusted according to different workflow requirements defined in different use cases. A new wiki namespace "WorkFlow" is defined and a workflow package can be generated by wiki users within this "WorkFlow" namespace to group a set of proposals and to define the working domain.

4. Prototype implementation

4.1. Use case description

The ICD is one of the main longstanding examples of how health information can assist people and countries in managing their health statistics. The 11th revision of the International Classification of Disease (ICD) was officially launched by the World Health Organization (WHO) in April 2007 [18]. The development of a web-based ICD revision platform is part of this revision process. The WHO initially adopted Web-Protégé for the alpha phase of ICD-11 development and the tool is called "iCAT".

iCAT is a variant of Web-Protégé. It is a web based application using Google Web Toolkit (GWT) technology [19]. The main features include

1) ICD content model (customized) browsing and editing;

2) community collaboration, including issue discussion and peer review (users may add notes and discussions to a category or a term attached to a category);
3) linkage to other terminologies (e.g., SNOMED, GO, etc.) through the BioPortal [20] ontology service;
4) hierarchy management, including moving in hierarchy, adding/removing parents, creating class, and retiring class;
5) change history tracking based on Change and Annotation Ontology (ChAO) mechanism; and
6) limited workflow support, e.g., modify access policies.

For the alpha process, the user community is relatively small, as the main task is to augment the definitions of rubrics and the review of elements in the Foundation Component of the ICD. However, in the beta phase, the ICD is supposed to be publicly reviewed by a large user community. The scalability issue of the iCAT tool will be emerging when multiple users work on the same copy of an evolving ICD category.

4.2. Mash-up of LexWiki and Web-Protégé

To meet the requirements of the beta phase of ICD development, we propose an extension that integrates the LexWiki environment with the existing iCAT system. The benefits of this proposal include:

1) Synergizing the strengths of both environments. The key point is that the structured proposal creation mechanism of LexWiki can be leveraged. This will enormously extend the flexibility and scalability of the iCAT functionalities.
2) Incremental updates for both environments. Proposals created in wiki platform can be harvested into the iCAT system for content curation and the updated contents can be published back to the wiki platform.
3) Separating the sophisticated user roles defined for ICD11 revision. Content experts will focus on proposal creation in the wiki platform that will not interfere with the contents authoring by classification experts in Protégé.
4) Keeping the iCAT user interface unchanged.

Fig. 2 shows a screenshot of the user interface of iCAT with a plug-in extension, demonstrating the connection of the ICD categories with the proposal mechanism of an instance platform of LexWiki. In this prototype implementation, users can

1) browse the contents of each category;
2) create structured proposals for an individual category;
3) harvest the proposals created within the wiki platform into the iCAT environment;
4) publish the updated contents of an ICD category back to the wiki.

For the next step, we will investigate

1) synchronization of user management process;
2) synchronization of category labels;
3) enhancement of proposal harvesting mechanism;
4) generalization of the system to support other ontology authoring.

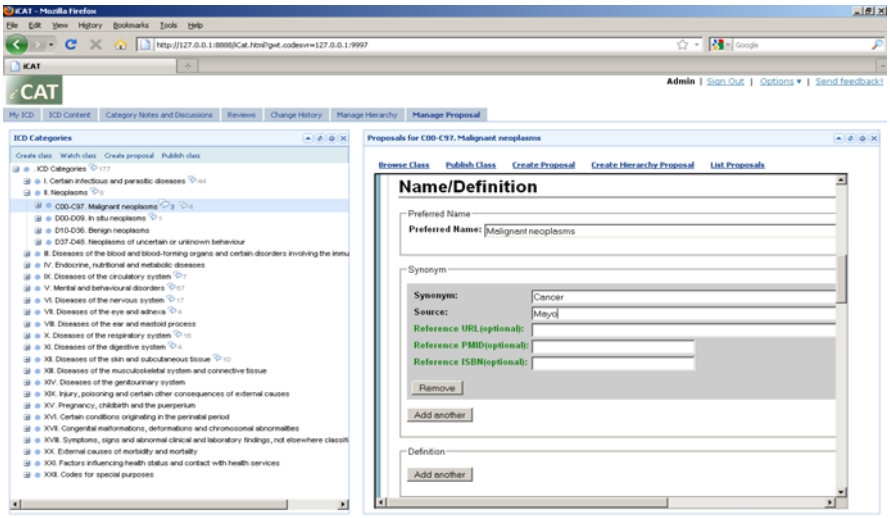

Figure 2. User interface of iCAT authoring tool with a plug-in extension, which connects the ICD categories in Web-Protégé with proposal mechanism in a LexWiki instance.

5. Conclusion

We demonstrated a mash-up of LexWiki and Web-Protégé within the framework for distributed authoring of large-scale biomedical terminologies. We consider that the proposed framework is feasible and can be useful to tackle the scalability issue the terminology authoring community is facing.

References

[1] Shadbolt N, Hall W: The semantic web revisited. IEEE Intelligent Systems 2006, 21 (3):96 –101.

[2] Seidenberg J, Rector AL: A methodology for asynchronous multi-user editing of semantic web ontologies. Proceedings of the 4th international conference on Knowledge capture Whistler, BC, Canada 2007:127 –134.

[3] Hepp MB, D; Siorpaes K: Community-driven ontology engineering and ontology usage based on Wikis. Proceedings of the 2006 international symposium on Wikis 2006:143 –144

[4] Hepp MB, Daniel; Siorpaes, Katharina; Harvesting wiki consensus - using wikipedia entries as ontology elements. IEEE INTERNET COMPUTING 2007, 11(5).

[5] Noy NF, Chugh A, Alani H: The CKC Challenge: Exploring Tools for Collaborative Knowledge Construction. Stanford Medical Informatics Technical Report 2007.

[6] Kamel Boulos MN. Semantic Wikis: A comprehensible introduction with examples from the health sciences. Journal of Emerging Technologies in Web Intelligence. 2009; (1): 94 –96.

[7] Baumeister J, Reutelshoefer J, and Puppe F. Engineering on the Knowledge Formalization Continuum. Proceedings of the Forth Semantic Wiki Workshop (SemWiki 2009), co-located with 6th European Semantic Web Conference (ESWC 2009). Hersonissos, Heraklion, Crete, Greece, June 1st, 2009

[8] Hartel FW, de Coronado S, Dionne R, Fragoso G, Golbeck J. Modeling a description logic vocabulary for cancer research. Journal of Biomedical Informatics. 28 (2005) 114 –129.

[9] Tudorache T, Noy N. Collaborative Protégé. WWW2007, May 8-12, 2007, Banff, Canada.

[10] Noy NF, Chugh A, Liu W, Musen MA: A framework for ontology evolution in collaborative environments. 5th International Semantic Web Conference, Athens, GA 2006.

[11] Jiang G, Solbrig H. LexWiki framework and use cases. The first meeting of Semantic MediaWiki users. Nov. 22-23, 2008. Boston, MA, USA. https://cabig-kc.nci.nih.gov/Vocab/KC/index.php/LexWiki#Presentations; last visited at December 4,2010.

[12] Krötzsch M, Vrandecic D, Völkel M, Haller H, Studer R. Semantic Wikipedia. Journal of Web Semantics 5: 251–261. September 2007.

[13] BiomedGT: http://biomedgt.nci.nih.gov/index.php/Main_Page; last visited at December 4, 2010.

[14] CTCAE: https://cabig-kc.nci.nih.gov/Vocab/KC/index.php/CTCAE; last visited at December 4, 2010.

[15] Jiang G, Solbrig H, Ibeson-Hurst D, Kush RD, Chute CG. A Collaborative Framework for Representation and Harmonization of Clinical Study Data Elements Using Semantic MediaWiki. AMIA Clinical Research Informatics Summit 2010. March 12-13, 2010. San Francisco.

[16] LexGrid Model: https://cabig-kc.nci.nih.gov/Vocab/KC/index.php /LexGrid; last visited at December 4, 2010.

[17] Java Wiki Bot Framework: http://sourceforge.net/projects/jwbf/ ; last visited at December 4, 2010.

[18] WHO: Production of ICD-11: The overall revision process. April 2007.

[19] iCAT: http://icat.stanford.edu/; last visited at December 4, 2010.

[20] BioPortal: http://bioportal.bioontology.org/; last visited at December 4, 2010.