A Semantic MediaWiki-Empowered Terminology Registry

Qing Zou
School of Information Studies
McGill University, Canada
qing.zou2@mail.mcgill.ca

Wei Fan
National Science Library
Chinese Academy of Sciences, China
fanwei@mail.las.ac.cn

Abstract

Semantic Wikis leverage the power of the Semantic Web and Wiki technologies. This paper introduces an approach of utilizing Semantic MediaWiki in particular as a platform for designing, developing, and constructing a terminology registry. A full-fledged terminology registry should be built on an open integrated infrastructure for terminology services. It not only registers various metadata schemas and controlled vocabularies, but also provides a collaborative development environment for constructing terminology resources. At the current stage, the project has focused on tools and techniques which deploy the semantic wiki method of modeling and managing terminology resources. Preliminary findings of this project will be discussed.

Keywords: metadata registry; terminology registry; semantic wiki; knowledge organization systems.

1. Introduction

A terminology registry lists, describes, identifies and points to sets of vocabularies available for use in information systems and services (Terminology Registry Scoping Study, 2008). Terminology registries provide the fundamental infrastructure for terminology services, such as web navigation, query expansion, cross-language retrieval, and metadata creation. There are already several terminology registries, including the Dublin Core Metadata Initiative (DCMI) registry[1], National Science Digital Library (NSDL) metadata registry[2], and CORES metadata registry[3]. However, as linked data continues to grow, we intend to present this project as a semantic wiki way to design a terminology registry from the linked data perspective.

Terminology is a list or vocabulary of terms (words or phrases) or notation used to describe, navigate, and search content (Proffitt et al., 2007). Terminology resources may contain terms, concepts and their relationships in vocabularies, and metadata schemas. Terminology resources could be considered to be a part of linked data that can be shared, mashed up, and reused by either humans or applications on the Semantic Web. A terminology resource may be listed, described and identified by one or more terminology registries. Meanwhile, a terminology registry registers and points to terminology resources.

A wiki is a type of web page designed so that its content can be edited using a simplified markup language by anyone who accesses it (Wiki, 2007). A wiki is often considered to be a collaborative and easy-to-use tool with a much lower technology barrier for creating and sharing information. A semantic wiki combines the strengths of wiki technology (easy to use and contribute, strongly interconnected, collaborative) and the Semantic Web technologies (machine processable, data integration, complex queries). Semantic wikis have the ability to capture or identify information about the data within pages, and the relationships between pages, in ways that can be queried or exported like database data (Semantic Wiki, 2009).

[1] DCMI Registry - http://dublincore.org/dcregistry/
[2] NSDL registry - http://metadataregistry.org/
[3] CORES - http://www.cores-eu.net/

https://doi.org/10.23106/dcmi.952109612

Many semantic wiki engines already exist, such as AceWiki[4], IkeWiki[5], and Semantic MediaWiki (SMW)[6]. In comparison with other semantic wiki engines, SMW has many advantages. It is built on the popular open source, MediaWiki, which has a large user community, a mature system framework, and various extensions. More importantly, MediaWiki is flexible and can be easily extended for specific use. The integration between MediaWiki and SMW is based on MediaWiki's extension mechanism: SMW registers for certain events or requests, and MediaWiki calls on SMW functions when needed (Krotzsch et al., 2007). SMW enables users to annotate the wiki's contents with explicit, machine-readable information (Krotzsch et al., 2007). Because of its robustness, flexibility, and support, SMW was selected for this project.

In this project, we explored the possibility and feasibility of establishing an open and collaborative terminology registry platform by building a semantic wiki-based terminology registry. More specifically, this project aims to examine the capabilities of a semantic wiki and to answer two main questions: 1) whether a semantic wiki is suitable for facilitating the development and management of terminology resources; and 2) whether open vocabularies integration processes such as the import and export functions of schemas and vocabularies can be seamlessly incorporated into a semantic wiki. In the long run, terminology services can be developed upon a SMW-empowered terminology registry including term expansion and disambiguation.

2. Technical Overview

After examining the capability of SMW, a prototype was constructed around SMW with feasible solutions to the two questions. Figure 1 shows the overview of this design.

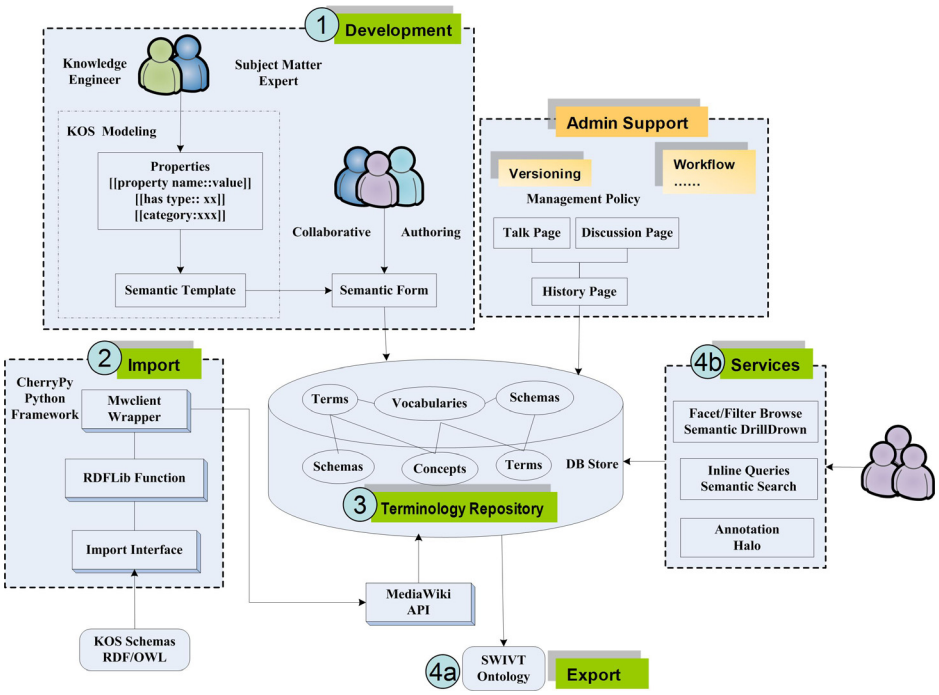

FIG. 1. Architecture of the SMW-empowered Terminology Registry

Components built around SMW include import/export, development, services, and other function modules such as versioning management and workflow. The development part (marked

[4] AceWiki - http://attempto.ifi.uzh.ch/acewiki/
[5] IkeWiki - http://ikewiki.salzburgresearch.at/
[6] Semantic Mediawiki - http://semantic-mediawiki.org

https://doi.org/10.23106/dcmi.952109612

as #1 in the Figure) is for knowledge engineers to model different types of KOS. Subject experts would be able to collaborate with knowledge engineers and build subject content by using semantic tools. The prototype provides an import interface (#2) for using and reusing existing vocabularies and schemas. The terminology repository (#3) currently stores all terminology resources in a MySQL database which can be replaced by a RDF triples store later. In addition to an export function (#4a), the services module (#4b) is built for users to utilize the terminology registry through add-on services.

3. SMW-Empowered Terminology Registry

In SMW, all content is structured by wiki pages. Each page could be considered a resource, a property, or an individual. Pages could be further classified into namespaces to distinguish different types of pages. More importantly, each wiki page has a unique name which can be treated as its Uniform Resource Identifier (URI) in combination with namespaces. Table 1 shows the representation of Web Ontology Language (OWL) and Resource Description Framework (RDF) constructs being mapped to wiki syntax in SMW.

TABLE 1: Representation of OWL terms in Semantic MediaWiki

OWL or RDF	SMW	Wiki Syntax
OWL individual	Normal article page	Page name
Owl:Class	Category	[[Category:*class]]*
Owl:ObjectProperty	Property	[[*property_name::*object]]
Owl:DatatypeProperty	Property	[[*property_name::*value]]
Rdf:type *class_name*	Category	[[Category:*class_name*]]
Rdfs:subClassOf *class_name*	Category	[[Category:*class_name*]]

The experimental SMW can be reached at http://libnt2.lakeheadu.ca/mw/. Currently, various schemas including OWL, Simple Knowledge Organization System (SKOS), Dublin Core (DC), RDF, and RDF Schema have been imported into our experimental wiki. In addition, a bilingual thesaurus, *Government of Canada Core Subject Thesaurus* (Library and Archives of Canada, 2008) (GCCST) has been converted into SKOS format and loaded into the project wiki via the import interface developed for this project.

The content of a terminology registry may be constructed differently in structure as well as syntax in a SMW. However, they all could be modeled in RDF triples. In the following section, we use a thesaurus entry for "Aerospace industry" (taken from GCCST) as an example to illustrate the semantic wiki approach.

3.1 Terms in SMW

Semantic relationships contained in a GCCST entry conform to the guidelines for the establishment and development of monolingual thesauri (ISO 2788-1986) and the guidelines for the establishment and development of multilingual thesauri (ISO 5964-1985)[7]. Shown below is the term "Aerospace industry" from GCCST:

TABLE 2: A thesaurus entry "Aerospace industry" (from GCCST)

Aerospace industry	A preferred term (in English) representing a concept
FRE: Industrie de l'aérospatiale	An equivalent in French
SN: Economic activity dedicated to the design and manufacturing of aircraft, spacecraft and missiles.	A scope note, specifying the meaning of a preferred term or defining its usage limitations for the purpose of indexing.
UF: Avionics, Space industry	One or several synonyms, quasi-synonyms or spelling variants

[7] http://www.thesaurus.gc.ca/default.asp?lang=En&n=0073D232-1

https://doi.org/10.23106/dcmi.952109612

BT: Manufacturing industry	One or several broader terms, preferred terms representing concepts broader in scope.
NT: Aviation industry	One or several narrower terms, preferred terms representing concepts narrower in scope.
RT: Space exploration	One or several RELATED TERMS, preferred terms that are associated in meaning with the preferred term of the record, but that are not part of the same hierarchy.
SC: EC Economics and Industry	One or several SUBJECT CATEGORIES, that helps to determine the meaning of the descriptor in the context of the thesaurus.

For demonstration purposes, the URI for GCCST is pointing to http://libnt2.lakeheadu.ca/cst. In SMW, if cst is the namespace of GCCST, "Category:cst:unique term number" can be defined as the internal identifier of the term. Preferred terms could be identifiers as well. For instance, "Aerospace industry@en" could be another identifier of the term. Then the full URI of the term in the experimental wiki is http://libnt2.lakeheadu.ca/mw/Category:cst:Aerospace industry@en. The following table shows the term expressed in RDF and SMW.

TABLE 3: Expressing the thesaurus entry for "Aerospace industry" in RDF and SMW

RDF	SMW
<rdf:Description rdf:about="http://libnt2.lakeheadu.ca/cst/Aerospace industry ">	[[URI::http://libnt2.lakeheadu.ca/cst/Aerospace industry]]
<skos:prefLabel xml:lang="en">Aerospace industry</skos:prefLabel>	[[skos:prefLabel ::Aerospace industry@en]]
<skos:prefLabel xml:lang="fr">Industrie de l'aérospatiale</skos:prefLabel>	[[skos:prefLabel ::Industrie de l'aérospatiale@fr]]
<skos:altLabel xml:lang="en">Avionics</skos:altLabel>	[[skos:altLabel::Avionics@en]]
<skos:altLabel xml:lang="en">Space industry</skos:altLabel>	[[skos:altLabel::Space industry@en]]
<skos:scopeNote>Economic activity dedicated to the design and manufacturing of aircraft, spacecraft and missiles.</skos:scopeNote>	[[skos:scopeNote::Economic activity dedicated to the design and manufacturing of aircraft, spacecraft and missiles.]]
<skos:related rdf:resource="http://libnt2.lakeheadu.ca/cst/Space exploration"/>	[[skos:related ::Category:cst:Space exploration]]
<skos:broader rdf:resource="http://libnt2.lakeheadu.ca/cst/Manufacturing industry"/>	[[skos:broader ::Category:cst:Manufacturing]]
<skos:narrower rdf:resource="http://libnt2.lakeheadu.ca/cst/Aviation industry"/>	[[skos:narrower ::Category:cst:Aviation industry]]
<skos:inScheme rdf:resource="http://libnt2.lakeheadu.ca/cst/"/>	[[skos:inScheme::http://libnt2.lakeheadu.ca/cst/]]
<skos:hasTopConcept>EC Economics and Industry </skos:hasTopConcept>	[[skos:hasTopConcept ::EC Economics and Industry]]

The real "Aerospace industry" wiki page would be more complicated than shown above. A semantic form and a semantic template have been created to make it easier to model KOS terms or vocabularies. In Table 1 shown in the previous section, we explained the underlying mapping between RDF and SMW. Figure 2 shows properties of the thesaurus term in our experimental wiki.

https://doi.org/10.23106/dcmi.952109612

FIG 2. Presenting the thesaurus entry for "Aerospace industry" in a semantic wiki page

The heading of the entry is the preferred label in English. This can be altered if we would like to use the preferred French label as the heading.

3.2 Import

Existing schemas and vocabularies are valuable terminology resources. It is crucial to encourage the utilization of those terminology resources in order to ensure interoperability. The import function could be considered a fundamental building block of a terminology registry. It allows users to load schema elements or vocabularies in bulk. In the project, the import interface is built on CherryPy[8] – a Python web framework. This project benefits from RDFLib[9] to handle schemas or vocabularies in RDF format. It utilizes mwclient[10] through which wiki pages could be created, deleted, and updated to communicate with a SMW. A web interface has been developed on CherryPy. The following is a code snippet of loading and handling a RDF file:

```
g = ConjunctiveGraph()
g.parse(StringIO(content))
prefixNS = g.namespaces()      '''update namespaces'''
wikins.updatens(prefixNS, myscheme.ns)
subjectList=wikins.uniq(list(g.subjects()))
for subj in subjectList:
   page=wikins.stripns(subj)
   for p,o in g.predicate_objects(subject=subj):
  ...
```

Through the import web interface, a user can specify a file to import. The function of import from a URI is under development. If the file is in RDF format, it would be parsed. Then RDF statements can be retrieved and manipulated. More specifically, RDF statements will be converted into corresponding wiki syntax based on the mapping relationships shown in Table 1 and other criteria. Because other semantic wiki techniques have been used to model terminology resources, some of them need to be applied to the wiki pages at this stage. All these, including loading, importing, and conversion undertaken behind the scene, are transparent to users. Users have two options: (1) to import elements, properties, or terms one by one; or (2) to load them all at once. Finally, the real changes are made to the wiki through mwclient.

4. Discussions and Conclusions

In this project, we have demonstrated the power of SMW to manage metadata schemas and vocabularies. As shown in Figure 1, a fully functional integrated terminology registry platform supports four major functions including terminology development, version management, import/export, and semantic services.

The SMW technology provides some great features for basic platform construction. At the current stage, we utilized various SMW modules and extensions including semantic templates and semantic forms for modeling terminologies. The SMW provides support for managing users with different levels of access. In combination with semantic templates, some basic functions of a registry including managing users with different levels of access can be developed. To some extent, a certain level of trust could be established. Version management is a crucial issue of a fully functional registry. Like MediaWiki, any changes made to a SMW are recorded. We are investigating whether this feature of a SMW can be utilized for versioning management of a registry.

However, the SMW technology has several limitations in terms of using it as a platform for a terminology registry. The SMW exports RDF/OWL descriptions through special formatted URIs

[8] CherryPy - http://cherrypy.org/
[9] RDFLib - http://rdflib.net/
[10] MWClient - http://sourceforge.net/projects/mwclient

https://doi.org/10.23106/dcmi.952109612

(e.g. http://libnt2.lakeheadu.ca/mw/Special:ExportRDF/:term) which may make terms hard to be linked by other applications. This can be resolved by applying content negotiation. Currently, only the RDF format is supported by the export interface. The interface needs to be extended to provide support to other formats such as Notation3[11], Terse RDF Triple Language[12] or JavaScript Object Notation[13] to facilitate linking data.

We are working on setting up a stable experimental SMW, importing and loading more metadata schemas and vocabularies, and incorporating required functionality of a registry into the project. Version management and terminology services will be another important step in our future experimental plan.

Acknowledgements

Special thanks to Dr. Marcia Lei Zeng of Kent State University for her support, valuable comments, and help to this project and this paper. We are grateful to Moira Davidson and Debra Gold of Lakehead University Library for their support to the project. We also would like to thank Lakehead University Library for providing server space for this project.

References

Krotzsch, Markus, Denny Vrandecic, Max Volkel, Heiko Haller, and Rudi Studer. (2007). Semantic Wikipedia. *Web Semantics: Science, Services and Agents on the World Wide Web*, 5(4), 251-261.

Library and Archives of Canada. (2008). *Government of Canada Core Subject Thesaurus*. Retrieved April 7, 2009, from http://www.thesaurus.gc.ca/EAEAD1E6-7DD2-4997-BE7F-40BFB1CBE8A2/Thesaurus_2008_e.pdf

Proffitt, Merrilee, Gunter Waibel, Diane Vizine-Goetz, Andrew Hunghton. (2007). *Terminologies Services Strawman*. Distributed to Terminologies Services Meeting at the Metropolitan Museum of Art, September 12, 2007, New York City. Retrieved May 2, 2009, from http://www.oclc.org/programs/events/2007-09-12a.pdf

Semantic Wiki. (2009). In: *Wikipedia*. Retrieved April 27, 2009, from http://en.wikipedia.org/wiki/Semantic_wiki

Terminology Registry Scoping Study. (2008). UKOLN at the University of Bath and University of Glamorgan. Retrieved April 7, 2009, from http://www.ukoln.ac.uk/projects/trss

Wiki. (2007). In: *Oxford English Dictionary*. Retrieved May 2, 2009, from Oxford English Dictionary Online: http://dictionary.oed.com/cgi/entry/50293088

[11] http://www.w3.org/DesignIssues/Notation3
[12] http://www.w3.org/TeamSubmission/turtle/
[13] http://www.json.org/

https://doi.org/10.23106/dcmi.952109612