SWAD–Europe Deliverable 8.3: RDF Encoding of Multilingual Thesauri

Project name:
Semantic Web Advanced Development for Europe (SWAD-Europe)

Project Number:
IST-2001-34732

Workpackage name:
8. Thesaurus Research Prototype

Workpackage description:

Deliverable title:
8.3: Thesaurus Multilingual Report

URI:
http://www.w3c.rl.ac.uk/SWAD/deliverables/8.3.html

Authors:
Alistair Miles, CCLRC, UK
Brian Matthews, CCLRC, UK
Michael Wilson, CCLRC, UK

Abstract:
This document is an introduction and guide to using SKOS-Core and SKOS-Mapping RDF schemas for encoding of multilingual thesauri. It gives recommendations for encoding different types of multilingual thesauri, where there are variations in structure and implied semantics.

Status:
Completed Report

Comments on this document are welcome and should be sent to Alistair Miles or to the public-esw-thes@w3.org list. An archive of this list is available at http://lists.w3.org/Archives/Public/public-esw-thes/

Contents

1. Introduction
2. Multilingual Labelling
3. Interlingual Mapping
4. Summary and Conclusions
References
Appendix I - SKOS-Core RDF Schema
Appendix II - SKOS-Mapping RDF Schema

1. Introduction

This document is an introduction to using the SKOS-Core [1] and SKOS-Mapping [3] schemas for the RDF encoding of multilingual thesauri.

Every multilingual thesaurus can be analysed as a set of component thesauri, one for each language covered.

Commonly, there are two scenarios:

- The first scenario, a multilingual thesaurus is defined where the conceptual structure of each of the language components is identical. I.e. the hierarchical and associative relationships are mirrored exactly. In this scenario, it is most convenient to define a single conceptual structure, and provide alternative labelling information for each of the languages covered. This is the multilingual labelling approach.

- The second scenario, there exist two independent monolingual thesauri, defined in different languages, but covering the same or similar subject domain. A mapping between the two thesauri is desired, so that for example the two may be used interchangeably for language independent browsing of collection metadata [see e.g. 11]. In this scenario, it is most convenient to model the conceptual structure of each thesaurus independently, and provide a complete, bidirectional mapping between the two. This is the interlingual mapping approach.

Notably, there is also a third, hybrid scenario, in which a single authority is defining a multilingual thesaurus, but the separate language components do not have identical conceptual structures. This scenario is essentially identical to the second, if each language component is treated as an independent monolingual thesaurus. Therefore, this scenario is also best covered by the interlingual mapping approach.
2. Multilingual Labelling

This is the simplest treatment of a multilingual thesaurus. A single conceptual structure is defined. Each concept is given one descriptor label for each of the languages covered. Each concept may also be given any number of alternative labels from each of the languages covered.

The SKOS-Core schema is used to generate an RDF encoding of this information (see also SKOS-Core Guide [2]). To express the language of a label, add a language tag to that label (see also RDF/XML Syntax Specification [13]).

For example ...

```xml
<rdf:RDF xmlns="http://www.w3c.rl.ac.uk/2003/11/21-skos-core#" xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
  <Concept>
    <descriptor xml:lang="en">Sausages</descriptor>
    <altLabel xml:lang="en">Bangers</altLabel>
  </Concept>
</rdf:RDF>
```

An example of a thesaurus that is well covered by the multilingual labelling approach is the HPMulti thesaurus [8]. An excerpt from this thesaurus, encoded in RDF using SKOS-Core, is provided below ...

```xml
<?xml version="1.0" encoding="UTF-8"?>
  <rdfs:Class rdf:ID="Concept">
    <rdfs:label>HPMulti Concept</rdfs:label>
    <rdfs:subClassOf rdf:about="#skos:Concept"/>
  </rdfs:Class>
  <hpm:Concept>
    <descriptor xml:lang="en">Abortion</descriptor>
    <descriptor xml:lang="fr">Avortement</descriptor>
    <descriptor xml:lang="da">Abort</descriptor>
    <descriptor xml:lang="ne">Abortus</descriptor>
    <descriptor xml:lang="fi">Abortti</descriptor>
    <descriptor xml:lang="de">Schwangerschaftsabbruch</descriptor>
  </hpm:Concept>
  <hpm:Concept>
    <descriptor xml:lang="en">ADHD</descriptor>
    <altLabel xml:lang="en">Attention deficit and hyperactive disorder</altLabel>
    <altLabel xml:lang="fr">Trouble de l'attention</altLabel>
    <altLabel xml:lang="da">DAMP/MBD</altLabel>
    <altLabel xml:lang="ne">Aktivisvudan ja tarkkaavan häiriö</altLabel>
    <altLabel xml:lang="de">Minimale zerebrale Dysfunktion und Hyperaktivität</altLabel>
  </hpm:Concept>
  <hpm:Concept>
    <descriptor xml:lang="en">Advertising</descriptor>
    <altLabel xml:lang="en">Publicity</altLabel>
    <altLabel xml:lang="fr">Publicité</altLabel>
    <altLabel xml:lang="da">Annoncering</altLabel>
    <altLabel xml:lang="ne">Reclame</altLabel>
    <altLabel xml:lang="fi">Mainonta</altLabel>
    <altLabel xml:lang="de">Werbung</altLabel>
  </hpm:Concept>
  <hpm:Concept>
    <descriptor xml:lang="en">Assisted conception</descriptor>
    <altLabel xml:lang="en">Artificial insemination</altLabel>
    <altLabel xml:lang="en">In vitro fertilization</altLabel>
    <altLabel xml:lang="da">Kunstig befrugtning</altLabel>
    <altLabel xml:lang="ne">Kunstmatige inseminatie</altLabel>
    <altLabel xml:lang="de">Künstliche Befruchtung</altLabel>
  </hpm:Concept>
</rdf:RDF>
```
3. Interlingual Mapping

In this treatment, a separate conceptual structure is defined for each language component. A complete, bidirectional mapping is then defined between each of the language components. The SKOS-Mapping RDF schema is used to express the mappings.

Interlingual mapping is considered to be a special case of inter-thesaurus mapping [5]. A more complete description and guide to the use of SKOS-Mapping may be found in the SWAD-Europe report on inter-thesaurus mapping [4].

A situation that is well covered by the interlingual mapping approach is the combined use of the French Merimee thesaurus [9] and the English Art and Architecture Thesaurus [10]. These two thesauri have been defined by independent authorities, but are highly overlapping in their subject domain. A complete mapping has already been defined between these two thesauri [5]. An example of how these mappings could be expressed using the SKOS-Mapping schema is provided below ...

```xml
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xmlns="http://www.w3c.rl.ac.uk/2003/11/skos-core#"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:map="http://www.w3c.rl.ac.uk/2003/11/skos-mapping#"
  xmlns:aat="http://www.getty.edu/research/conducting_research/vocabularies/aat/"
  xmlns:mer="http://www.culture.fr/documentation/merimee#">
  <rdfs:Class rdf:about="http://www.culture.fr/documentation/merimee#Concept">
    <rdfs:label>Merimee Thesaurus Concept</rdfs:label>
  </rdfs:Class>
  <rdfs:Class rdf:about="http://www.getty.edu/research/conducting_research/vocabularies/aat/Concept">
    <rdfs:label>AAT Concept</rdfs:label>
  </rdfs:Class>
  <mer:Concept>
    <descriptor xml:lang="fr">Academie</descriptor>
    <map:exactMatch>
      <map:AND>
        <map:memberList rdf:parseType="Collection">
          <aat:Concept>
            <descriptor xml:lang="en">Academy</descriptor>
          </aat:Concept>
          <aat:Concept>
            <descriptor xml:lang="en">Buildings</descriptor>
          </aat:Concept>
        </map:memberList>
      </map:AND>
    </map:exactMatch>
  </mer:Concept>
  <mer:Concept>
    <descriptor xml:lang="fr">Aire de concassage</descriptor>
    <map:exactMatch>
      <map:AND>
        <map:memberList rdf:parseType="Collection">
          <aat:Concept>
            <descriptor xml:lang="en">Crushing</descriptor>
          </aat:Concept>
          <aat:Concept>
            <descriptor xml:lang="en">Floors</descriptor>
          </aat:Concept>
        </map:memberList>
      </map:AND>
    </map:exactMatch>
  </mer:Concept>
</rdf:RDF>
```
Note that language tags should still be attached to the labels, even though every concept has only labels from one language.

4. Summary and Conclusions

The SKOS-Core [1] and SKOS-Mapping [2] vocabularies may be used to achieve the RDF encoding of multilingual thesauri. The style and nature of this encoding is consistent with current standards and best practice in the construction and use of multilingual thesauri [3] [4] [5] [6].

When approaching the task of generating an RDF encoding for a multilingual thesaurus, it should first be considered whether the multilingual labelling approach, or the interlingual mapping approach is most suitable. If the thesaurus essentially describes a single, coherent conceptual structure, then the multilingual labelling approach is
If each language component of the thesaurus defines a unique conceptual structure, then the interlingual mapping approach is recommended. It is worth bearing in mind that, if a complete interlingual mapping has not already been defined for your target thesaurus, to generate such a mapping from scratch is a labour-intensive task.

This document should be accompanied by SWAD-Europe Deliverable 8.4, which gives a fuller guide to the generation, use and encoding of interthesaurus mappings.

References

[1] SKOS-Core RDF Schema
http://www.w3c.rl.ac.uk/2003/11/21-skos-core
http://www.w3c.rl.ac.uk/SWAD/deliverables/8.1.html
http://www.w3c.rl.ac.uk/2003/11/21-skos-mapping
http://www.w3c.rl.ac.uk/SWAD/deliverables/8.4.html
http://jodi.ecs.soton.ac.uk/Articles/v01/i08/Doerr/
http://www.dublincore.org/documents/dco-rdf-xml/
http://hilt.cdr.strath.ac.uk/hilt/Reports/HILT литRev%201a.doc
[8] HPMulti: The European multilingual thesaurus on health promotion in 12 languages.
http://www.hpmulti.net/
http://www.culture.gouv.fr/documentation/thesearch/pres.htm
http://www.getty.edu/research/conducting_research/vocabularies/aat/
http://www.limber.rl.ac.uk/
http://www.w3.org/TR/rdf-syntax-grammar/

Appendix I – SKOS–Core RDF Schema

```xml
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE rdfs [ 
  <!ENTITY rdf "http://www.w3.org/1999/02/22-rdf-syntax-ns#" >
  <!ENTITY rdfs "http://www.w3.org/2000/01/rdf-schema#" >
  <!ENTITY dc "http://purl.org/dc/elements/1.1/" >
  <!ENTITY dct "http://purl.org/dc/terms/" >
 ]>
<rdfs:RDF xmlns:rdf="&rdf;" xmlns:rdfs="&rdfs;" xmlns:dc="&dc;"
xmlns:dct="&dct;"
xml:base="http://www.w3c.rl.ac.uk/2003/11/21-skos-core" >
  <!-- Description of this schema -->
  <rdf:Description rdf:about=""/>
  <!-- SKOS RDF Vocabulary -->
  <dct:title>SKOS RDF Vocabulary</dct:title>
  <dct:description>An RDF vocabulary for defining simple conceptual schemes such as thesauri, taxonomies and classification schemes. This is a pure RDF schema version. Features such as transitivity of properties and equivalence with properties from other schema are described in the comments. These will be formally specified in a later OWL version of this schema.</dct:description>
  <dcterms:modified>2003-11-21</dcterms:modified>
  <dc:creator>Alistair Miles</dc:creator>
  <dc:generator> trademark: OWL Reasoner</dc:generator>

  <!-- Fundamental classes -->
  <rdfs:Class rdf:ID="Concept">
    <rdfs:label>Concept</rdfs:label>
    <rdfs:comment>Use this class to indicate that a node may be considered to be an abstract concept that is part of some conceptual scheme such as a thesaurus.</rdfs:comment>
  </rdfs:Class>
  <rdfs:Class rdf:ID="Facet">
    <rdfs:label>Facet</rdfs:label>
    <rdfs:comment>Facets provide a means of organising concepts along orthogonal dimensions. A facet is treated as a concept. A facet may have member concepts. A concept may be a member of only one
```
```
the preferred label for a resource. If a resource has this property, all other rdfs:label properties are considered to be the 'alternative' (i.e. non-preferred) labels. 

<!-- Generic semantic relation properties -->

<!-- Semantic relation property extensions -->

</rdfs:Class>

</rdf:Property rdf:ID="broader"> 
  <rdfs:label>has-broader-concept</rdfs:label> 
  <rdfs:subPropertyOf rdf:resource="#semanticRelation"/> 
  <rdfs:comment>This property carries very weak semantics. It may be used to state that the object is in some way more general in meaning than the subject. Essentially it provides a means of organizing concepts into hierarchical structures, without being restrictive about the exact semantic implications of the hierarchical structure itself. Extend this property to create properties that carry stronger semantics, but may be reduced to a hierarchical structure for simple visual displays. This property may be considered to be transitive. </rdfs:comment>
</rdf:Property>

</rdf:Property rdf:ID="narrower"> 
  <rdfs:label>has-narrower-concept</rdfs:label> 
  <rdfs:subPropertyOf rdf:resource="#semanticRelation"/> 
  <rdfs:comment>This property should be considered to be the inverse of the 'broader' property. </rdfs:comment>
</rdf:Property>

</rdf:Property rdf:ID="related"> 
  <rdfs:label>has-related-concept</rdfs:label> 
  <rdfs:subPropertyOf rdf:resource="#semanticRelation"/> 
  <rdfs:comment>This property carries very weak semantics. It may be used to state that the object is in some way related to the subject, and the nature of that relationship is NOT to be treated as hierarchical. Extend this property to create properties with stronger semantics, but may still be reduced to an associative structure for simple visual display. This property should be considered to be symmetric. </rdfs:comment>
</rdf:Property>

</rdf:Property rdf:ID="inFacet"> 
  <rdfs:label>member-of-facet</rdfs:label> 
  <rdfs:subPropertyOf rdf:resource="#broader"/> 
  <rdfs:range rdf:resource="#Facet"/> 
  <rdfs:comment>This property indicates that a concept is a member of a facet. A concept may have only one inFacet property. This property is a sub-property of the 'broader' property. Thus faceted conceptual structures may be reduced to simple hierarchical displays by applications that do not comprehend facets. </rdfs:comment>
</rdf:Property>

</rdf:Property rdf:ID="broaderInstantive"> 
  <rdfs:label>instance-of</rdfs:label> 
  <rdfs:subPropertyOf rdf:resource="#broader"/> 
  <rdfs:comment>An extension of the 'broader' property to specify the instantiation relationship between two concepts. This property is semantically equivalent to the 'rdf:type' property. </rdfs:comment>
</rdf:Property>

</rdf:Property rdf:ID="narrowerInstantive"> 
  <rdfs:label>has-instance</rdfs:label> 
  <rdfs:subPropertyOf rdf:resource="#narrower"/> 
  <rdfs:comment>An extension of the 'narrower' property to specify the instantiation relationship between two concepts. This property should be considered the inverse of 'broaderInstantive'. </rdfs:comment>
</rdf:Property>

</rdf:Property>

</rdf:Property rdf:ID="broaderGeneric"> 
  <rdfs:label>sub-class-of</rdfs:label> 
  <rdfs:subPropertyOf rdf:resource="#broader"/> 
  <rdfs:comment>
An extension of the 'broader' property to specify the class subsumption relationship between two concepts. This property is semantically equivalent to the rdfs:subClassOf property.

An extension of the 'narrower' property to specify the class subsumption relationship between two concepts. This property should be considered the inverse of 'broaderGeneric'.

An extension of the 'broader' property to specify a partitive relationship between two concepts. This property should be considered to be the inverse of 'broaderPartitive'.

An extension of the 'narrower' property to specify a partitive relationship between two concepts. This property should be considered to be the inverse of 'narrowerPartitive'.

A scope note is a piece of text that in some way helps to further elucidate the intended meaning of a concept.

A note on the hierarchical location of a concept.

A note providing information about the history of change of a concept.

Appendix II – SKOS–Mapping RDF Schema
of resources properly indexed against the first concept is identical to the set of resources properly indexed against the second. Therefore the two concepts may be interchanged in queries and subject-based indexes.

If 'concept A has-broad-match concept B' then the set of resources properly indexed against concept A is a subset of the set of resources properly indexed against concept B.

If 'concept A has-narrow-match concept B' then the set of resources properly indexed against concept A is a superset of the set of resources properly indexed against concept B.

If 'concept A has-major-match concept B' then the set of resources properly indexed against concept A shares more than 50% of its members with the set of resources properly indexed against concept B.

If 'concept A has-minor-match concept B' then the set of resources properly indexed against concept A shares less than 50% but greater than 0 of its members with the set of resources properly indexed against concept B.

This class is a shorthand for an intersection-like construct. So the statement 'concept A has-exact-match AND (concept B, concept C)' implies that the set of resources properly indexed against concept A is identical to the intersection of the sets properly indexed against concepts B and C.

This class is a shorthand for a union-like construct. So the statement 'concept A has-exact-match OR (concept B, concept C)' implies that the set of resources properly indexed against concept A is identical to the union of the sets properly indexed against concepts B and C.

This class is a shorthand for a negation-like construct. So the statement 'concept A has-exact-match NOT (concept B, concept C)' implies that the set of resources properly indexed against concept A is identical to the intersection of the set properly indexed against concept B and the complement of the set properly indexed against concept C.