Migrating Thesauri to the Semantic Web
Guidelines and case studies for generating RDF encodings of existing thesauri

Abstract:

This document presents guidelines, methods and examples for migrating current thesaurus systems to RDF based thesaurus systems. Thesauri with both 'standard' and 'non-standard' structure are considered. Three case studies are presented.

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:
8.8: old_thesaurus_migration_report

This version:
http://www.w3.org/2001/sw/Europe/reports/thes/8.8/version02.html

Latest version:
http://www.w3.org/2001/sw/Europe/reports/thes/8.8/

Previous version:
http://www.w3.org/2001/sw/Europe/reports/thes/8.8/draft01.html

Status:
Completed

Authors:
- Alistair J. Miles, CCLRC
- Nikki Rogers, ILRT
- Dave Beckett, ILRT

Status of this document

This section describes the status of this document at the time of its publication. This is a draft document and may be updated, replaced, or obsoleted by other documents at any time. The latest status of this document series is maintained at the W3C.

This document is a public DRAFT for discussion. These guidelines and the SKOS-Core schema are an output of the research work of the Semantic Web Advanced Development for Europe Project, which is associated with the W3C Semantic Web Activity. This document is made available by W3C for discussion only. Publication of this document by W3C does not
1. Introduction

This document describes principles, guidelines and examples for the process of publishing a thesaurus on the semantic web.

The stages of this process are: (1) Generate an RDF encoding of the thesaurus, (2) error checking and validation of the encoding, (3) publishing the encoding on the web.

This document first describes an approach for migrating 'thesauri with standard structure' (Section 2). Here, a 'thesaurus with standard structure' includes any thesaurus whose structural features coincide with the set of structural features described by ISO 2788:1986 [ISO2788] even if the principles of construction differed from those described by the standard.

Thus a 'thesaurus with standard structure' consists of two types of term: 'preferred' and 'non-preferred'; and five types of term-to-term relationship: 'broader' (BT), 'narrower' (NT), 'related' (RT), 'use' (USE) and 'use for' (UF); where BT NT and RT are allowed only between preferred terms, USE relates a non-preferred term to a preferred term, and UF is the inverse of USE; terms may be annotated with 'scope notes'.

This document then describes an approach for migrating 'thesauri with non-standard structure' (Section 3). Here a 'thesaurus with non-standard structure' includes any thesaurus with structural features that are not described by the standard ISO 2788:1986 [ISO2788].

The goal when migrating thesauri with non-standard structure is to preserve all the unique features of a thesaurus, while also providing a framework for interoperability with other thesauri.

Section 4 comprises three worked examples of migrating thesauri to the semantic web:

- The first example (Section 4.1) is the APAIS Thesaurus (APAIS), a standard monolingual thesaurus, currently available in a native XML format. An RDF/XML encoding this thesaurus is generated using a single XSLT transformation.
- The second example (Section 4.2) is the English Heritage Aircraft Type Thesaurus (ATT), a standard monolingual thesaurus, available as a relational database. An RDF encoding of this thesaurus is generated from comma-delimited database output using a short java program.
- The third example (Section 4.3) is GEMET (GEMET), a non-standard multilingual...
thesaurus, currently available in a native XML format. An RDF/XML encoding of GEMET is generated using multiple XSLT transformations.

2. Thesauri with Standard Structure

A 'thesaurus with standard structure' is here defined as any thesaurus conforming to the following structural restrictions:

- The thesaurus consists of a set of terms, a set of term-to-term relationships, and a set of term annotations.
- A 'term' is any word or phrase.
- There are two types of term, 'non-preferred' and 'preferred'.
- The following term-to-term relationships are allowed between preferred terms only: 'broader' (BT), 'narrower' (NT), 'related' (RT).
- The term-to-term relationship 'use' (USE) directs a user from a non-preferred term to its preferred alternative; 'use for' (UF) is the inverse of 'use'.
- Terms may be annotated with 'scope notes'.

An example of an extract from a typical rendering of a standard thesaurus is below:

<table>
<thead>
<tr>
<th>Therapy</th>
</tr>
</thead>
<tbody>
<tr>
<td>NT Back care</td>
</tr>
<tr>
<td>Back care</td>
</tr>
<tr>
<td>BT Therapy</td>
</tr>
<tr>
<td>RT Back pain</td>
</tr>
<tr>
<td>Back pain</td>
</tr>
<tr>
<td>UF Backache</td>
</tr>
<tr>
<td>RT Back care</td>
</tr>
<tr>
<td>Musculoskeletal disorders</td>
</tr>
<tr>
<td>Musculoskeletal disorders</td>
</tr>
<tr>
<td>UF Bone disorders</td>
</tr>
<tr>
<td>RT Back pain</td>
</tr>
</tbody>
</table>

This is the traditional, *term-oriented*, view of a thesaurus, oriented towards use by people via print media.

It is recommended to use the SKOS-Core RDF vocabulary [SKOS SCHEMA] [SKOS GUIDE] as the basis for an RDF encoding of a thesaurus with standard structure. Re-using existing RDF vocabularies whenever possible is highly desirable, as it provides a basis for programmatic interoperability.

The SKOS-Core RDF vocabulary is oriented towards use by computer programs via the semantic web. It is based on a concept-oriented view of a thesaurus, primarily because this makes programmatic manipulation and maintenance of the data much simpler. In the concept-oriented view:

- Each preferred term becomes the 'preferred label' for a 'concept'.
- Non-preferred terms become 'alternative labels' for concepts.
- A 'label' can be any string of characters, or any symbol or image.
- Relationships such as 'broader' 'narrower' and 'related' are relationships of meaning, and are hence relationships between concepts ('semantic relations').
- Concepts may have annotations such as scope notes and definitions.
- The meaning (intension) of a concept should be inferred from the combination of its preferred label, its alternatives labels, any annotations, and its neighbouring concepts.

An illustration of the RDF model of the above thesaurus extract, based on the SKOS-Core schema, is below:
The above model has the following RDF/XML serialisation:

```xml
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:skos="http://www.w3.org/2004/02/skos/core#"
  xml:base="http://www.examplenamespace.com/">
  <skos:Concept rdf:about="001">
    <skos:prefLabel>Therapy</skos:prefLabel>
    <skos:narrower rdf:resource="002"/>
  </skos:Concept>
  <skos:Concept rdf:about="002">
    <skos:prefLabel>Back care</skos:prefLabel>
    <skos:broader rdf:resource="001"/>
    <skos:related rdf:resource="003"/>
  </skos:Concept>
  <skos:Concept rdf:about="003">
    <skos:prefLabel>Back pain</skos:prefLabel>
    <skos:altLabel>Backache</skos:altLabel>
    <skos:related rdf:resource="002"/>
    <skos:related rdf:resource="004"/>
  </skos:Concept>
  <skos:Concept rdf:about="004">
    <skos:prefLabel>Musculoskeletal disorders</skos:prefLabel>
    <skos:altLabel>Bone disorders</skos:altLabel>
    <skos:related rdf:resource="003"/>
  </skos:Concept>
</rdf:RDF>
```

Case studies 4.1 [→ Section 4.1] and 4.2 [→ Section 4.2] below provide in depth examples of generating an RDF encoding of a standard thesaurus.

### 3. Thesauri with Non-Standard Structure [← back to contents]

Here a 'thesaurus with non-standard structure' includes any thesaurus with structural features that are not described by ISO 2788:1986 [→ ISO2788].
When migrating such a thesaurus to the semantic web, it is desirable to preserve all of the unique features of the thesaurus, i.e. to generate an RDF encoding that preserves all of the information encoded in the thesaurus. However, it is also desirable to provide at least some basis for interoperability. In this way, specialised applications can operate on the unique and specific features of the thesaurus, and generalised thesaurus applications can still recognise the standard features of the thesaurus and operate on them.

To achieve this aim, it is recommended that a schema extension be created for each thesaurus with non-standard structure. This is an RDF Schema that captures all of the features of the thesaurus, but where all of the new classes and properties are derived from the classes and properties of the SKOS-Core RDF vocabulary \[<SKOS_SCHEMA>\][<SKOS_GUIDE>] via sub-class and sub-property statements.

As an example of this, consider the GEMET thesaurus \[<GEMET>\]. GEMET uses most of the standard thesaurus features:

- It consists of a set of 'descriptors' (preferred terms), which may be related to each other via the properties 'broader' 'narrower' and 'related'.

It also has some non-standard features:

- Each 'descriptor' is related to a 'theme'.
- Each 'descriptor' is related to a 'group'.
- Each 'group' is related to a 'super-group'.

Both the groups, super-groups and themes have been given labels that indicate that they represent concepts in their own right. Therefore, three new classes may be created as part of the extended schema to capture these types \(\text{gemet:Theme, gemet:Group, gemet:SuperGroup}\) and these classes may be defined as sub-classes of the \text{skos:Concept} class. The \text{gemet:SuperGroup} class is defined as a sub-class of \text{skos:TopConcept} class, because this type represents the nodes that sit at the very top of the hierarchy.

```
@prefix gemet: <http://www.eionet.eu.int/gemet/schema#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .

gemet:Theme a rdfs:Class;
    rdfs:label 'Theme';
    rdfs:subClassOf skos:Concept.


gemet:Group a rdfs:Class;
    rdfs:label 'Group';
    rdfs:subClassOf skos:Concept.


gemet:SuperGroup a rdfs:Class;
    rdfs:label 'Super Group';
    rdfs:subClassOf skos:TopConcept.
```

Because every 'theme' is broader in sense than the 'descriptor' to which it is related, the property linking a concept to a theme may be defined as a sub-property of the \text{skos:broader} property. The same is true for 'descriptors' and 'groups', and also for 'groups' and 'super-groups'.

```
@prefix gemet: <http://www.eionet.eu.int/gemet/schema#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .

gemet:broaderTheme a rdf:Property;
    rdfs:label 'broader theme';
    rdfs:subPropertyOf skos:broader;
    rdfs:domain skos:Concept;
    rdfs:range gemet:Theme.


gemet:broaderGroup a rdf:Property;
    rdfs:label 'broader group';
    rdfs:subPropertyOf skos:broader;
    rdfs:domain skos:Concept;
    rdfs:range gemet:Group.


gemet:subGroupOf a rdf:Property;
    rdfs:label 'sub-group of';
```
So to recap, for a thesaurus with non-standard structure, first design a schema extension based on the SKOS-Core RDF vocabulary to capture any non-standard features, then generate an RDF encoding of the thesaurus based on the SKOS-Core RDF vocabulary and the schema extension.

Case study 4.3 [↑ Section 4.3] provides an in depth example of this method for the GEMET thesaurus.

### 4. Case Studies [↑ back to contents]

#### 4.1. APAIS Thesaurus [↑ back to contents]

The Australian Public Affairs Information Service Thesaurus [↑ APAIS] is available as an XML download [↑ APAIS XML], with an accompanying Document Type Definition (DTD) [↑ APAIS DTD] which is based on the Z39.50 profile for thesaurus navigation [↑ ZTHES]. The XML document root `<thes>` element contains a number of `<term>` elements, an example of which is below:

```xml
<thes>
  <term>
    <termId>R1722</termId>
    <termName>Aboriginal archaeology</termName>
    <termType>PT</termType>
    <termNote>Use for the study of Aboriginal and Torres Strait Islander prehistoric culture chiefly through excavation and description of its material remains</termNote>
    <termCreatedBy>1986</termCreatedBy>
    <termModifiedDate>9/04/2002</termModifiedDate>
    <relation>
      <relationType>UF</relationType>
      <termId>N0003</termId>
      <termName>Aboriginal prehistory</termName>
      <termType>ND</termType>
    </relation>
    <relation>
      <relationType>UF</relationType>
      <termId>N0651</termId>
      <termName>Prehistory, Aboriginal</termName>
      <termType>ND</termType>
    </relation>
    <relation>
      <relationType>BT</relationType>
      <termId>R0064</termId>
      <termName>Archaeology</termName>
      <termType>PT</termType>
    </relation>
    <relation>
      <relationType>RT</relationType>
      <termId>R1723</termId>
      <termName>Aboriginal history</termName>
      <termType>PT</termType>
    </relation>
    <relation>
      <relationType>RT</relationType>
      <termId>R0099</termId>
      <termName>Aborigines</termName>
      <termType>PT</termType>
    </relation>
  </term>
  ...
</thes>
```

The APAIS thesaurus is classified here as a thesaurus with standard structure, because it consists only of terms, with the relations UF USE BT NT RT between them. Therefore it can be mapped directly into the SKOS-Core 1.0 schema with no loss of information.

An RDF/XML encoding of this thesaurus can be generated from the source XML using an...
XSLT transformation. The full XSLT stylesheet is linked from this reference [APAIS XSLT].

Appendix I [Appendix I] walks through this stylesheet in detail, explaining its features.

The complete XSLT stylesheet, when applied to the above excerpt from the APAIS thesaurus, generates the following snippet of RDF/XML:

RDF/XML generated from transformation

```xml
<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
xmlns:skos="http://www.w3.org/2004/02/skos/core#"
xmlns:dc="http://purl.org/dc/elements/1.1/"
    <dc:title>Australian Public Affairs Information Service (APAIS) thesaurus</dc:title>
  </skos:ConceptScheme>
  <skos:Concept rdf:about="R1722">
    <skos:prefLabel>Aboriginal archaeology</skos:prefLabel>
    <skos:altLabel>Aboriginal prehistory</skos:altLabel>
    <skos:altLabel>Prehistory, Aboriginal</skos:altLabel>
    <skos:related rdf:resource="R0009"/>
    <skos:broader rdf:resource="R0064"/>
    <skos:related rdf:resource="R1723"/>
    <skos:scopeNote>Use for the study of Aboriginal and Torres Strait Islander prehistoric culture chiefly through excavation and description of its material remains</skos:scopeNote>
  </skos:Concept>
  ...
</rdf:RDF>
```

One point to note is that the numbering scheme from the native XML format has been used as the basis for the concept URIs. It is recommended wherever possible to base concept URIs on existing identifiers.

Once an RDF/XML transformation has been applied, the final step is to run well-formedness, syntax, and validation checks (see [RDF VALIDATE]) on the output.

The full output of this conversion is at this reference [APAIS OUT].

4.2. English Heritage Aircraft Type Thesaurus [back to contents]

The English Heritage Aircraft Type Thesaurus (ATT) [ATT] is a monolingual thesaurus with standard structure. English Heritage uses a relational database to store and maintain this thesaurus. For this case study EH provided us with comma delimited files, one file for each table in the database.

The EH database schema for thesauri consists of the following tables (here each table illustrated with field headings and an example tuple):

'classification groups table'

<table>
<thead>
<tr>
<th>CLA_GR_UID</th>
<th>DESCRIPTION</th>
<th>NAME</th>
<th>CLASS_TYPE</th>
<th>STATUS</th>
<th>OPS_SET</th>
</tr>
</thead>
<tbody>
<tr>
<td>225</td>
<td>&quot;AIRCRAFT TYPE&quot;</td>
<td>&quot;T&quot;</td>
<td>&quot;O&quot;</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

One tuple for each thesaurus (thesaurus metadata).

'terms table'

<table>
<thead>
<tr>
<th>CLA_GR_UID</th>
<th>THE_TE_UID</th>
<th>TERM</th>
<th>INDEX_TERM</th>
<th>SCOPE_NOTE</th>
<th>STATUS</th>
</tr>
</thead>
<tbody>
<tr>
<td>225</td>
<td>111367</td>
<td>&quot;LANCASTER&quot;</td>
<td>&quot;Y&quot;</td>
<td>&quot;Four-engined bomber, developed by Avro from the&quot;</td>
<td>&quot;P&quot;</td>
</tr>
</tbody>
</table>
twin-engined Manchester.
Entered service in 1942 as the RAF's principal night bomber and took part in the 1,000 bomber raids as well as the famous Dam Busters raid."

One tuple for each term. A 'STATUS' value of "P" indicates a preferred term, "N" indicated a non-preferred term.

'term preferences table'

<table>
<thead>
<tr>
<th>THE_TE_UID_1</th>
<th>THE_TE_UID_2</th>
</tr>
</thead>
<tbody>
<tr>
<td>111425</td>
<td>111363</td>
</tr>
</tbody>
</table>

One tuple for each 'USE' relationship.

'term uses table'

<table>
<thead>
<tr>
<th>THE_T_U_UID</th>
<th>TERM</th>
<th>CLA_GR_UID</th>
<th>BROAD_TERM_U_UID</th>
<th>TOP_TERM_U_UID</th>
</tr>
</thead>
<tbody>
<tr>
<td>175943</td>
<td>&quot;HAMPDEN&quot; 225</td>
<td>153785</td>
<td>167503</td>
<td></td>
</tr>
</tbody>
</table>

One tuple for each 'BT' relationship.

'term relation table'

<table>
<thead>
<tr>
<th>THE_T_U_UID_1</th>
<th>THE_T_U_UID_2</th>
</tr>
</thead>
<tbody>
<tr>
<td>135606</td>
<td>133678</td>
</tr>
</tbody>
</table>

One tuple for each 'RT' relationship.

The generation of an RDF encoding of this thesaurus was done programmatically, using the Jena 2 [🔗 JENA] Java API for RDF. A link to the Java main class written to perform the conversion is at this reference [🔗 EH_CONV], supporting classes here [🔗 EH TOK] [🔗 SKOS JAVA].

To summarise the operations of this class, basically, an in memory RDF model is created, each comma delimited file (the output of a table) is parsed in turn and used to add statements to the model. More specifically:

- From the single tuple of the 'classification groups table' a resource representing the thesaurus is created, and typed as a skos:ConceptScheme. To this resource dc:title and dc:description properties are added with values from fields of this tuple.
- From each tuple of the 'terms table', if the 'STATUS' value is "P" (i.e. the term is a preferred term), a resource is created and typed as a skos:Concept. To this resource skos:inScheme, skos:prefLabel and skos:scopeNote properties are added.
- From each tuple of the 'term preferences table' a skos:altLabel property is added to the appropriate concept.
- From each tuple of the 'term uses table' skos:broader/skos:narrower statements are added between the appropriate concepts.
- From each tuple of the 'term relation table' skos:related statements are added between the appropriate concepts.

The model is then serialised, and this RDF file is then the output of the conversion.

Examine the Java class file itself for full details of the conversion program [🔗 EH_CONV].

An extract of the RDF output of this conversion is at these references [🔗 EH
4.3. GEMET [- back to contents] -

GEMET is a multilingual thesaurus, publicly available as an XML download [→ GEMET]. There is one XML file for each of the 16 languages in which GEMET is available. These XML files share the same markup structure and element names - only the element contents change with the language.

The extracts below are from the GEMET EN-GB (British English) XML file.

The 'thesaurus' element contains a list of 'descriptor' elements, an example of which is below:

```
<thesaurus>
  <descriptor>
    <descriptor-term desc-id="7" desc-type="1" top="0">abandoned industrial site</descriptor-term>
    <broader-term desc-ref-id="4666">land setup</broader-term>
    <narrower-term desc-ref-id="2275">disused military site</narrower-term>
    <theme accronym="IND" theme-id="19">industry</theme>
    <theme accronym="NAT" theme-id="23">natural areas, landscape, ecosystems</theme>
    <theme accronym="POLL" theme-id="26">pollution</theme>
    <theme accronym="URB" theme-id="38">urban environment, urban stress</theme>
    <group group-id="1062" super-group-id="5499">ANTHROPOSHERE (built environment)</group>
  </descriptor>
</thesaurus>
```

The 'thesaurus' element also contains some 'super-group' elements, for example:

```
<thesaurus>
  <super-group super-group-id="2894">
    <super-group-name>SOCIAL ASPECTS, ENVIRONMENTAL POLICY MEASURES</super-group-name>
  </super-group>
</thesaurus>
```

Finally the 'thesaurus' element contains some 'descriptor-related' elements, for example:

```
<thesaurus>
  <descriptor-related desc-ref-id="9147" rel-desc-ref-id="9214"/>
  <descriptor-related desc-ref-id="9160" rel-desc-ref-id="4505"/>
  <descriptor-related desc-ref-id="12214" rel-desc-ref-id="1808"/>
</thesaurus>
```

Because the 'group' 'super-group' and 'theme' constructs are non-standard thesaurus constructs, a schema extension must be defined for GEMET.

The GEMET schema extension is below:

```
<!DOCTYPE skos [ <!ENTITY skos "http://www.w3.org/2004/02/skos/core#" ]>
<rdf:RDF
  xmlns:rdfs="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns: rdf="http://www.w3.org/2000/01/rdf-schema#"
  xmlns: skos="http://www.eionet.eu.int/GEMET/skos-ext">
  <!-- This is the extension of SKOS-Core for the GEMET Thesaurus -->
  <rdfs:Class rdf:ID="SuperGroup">
    <rdfs:label>Super Group</rdfs:label>
    <rdfs:subClassOf rdf:resource="#skos:TopConcept"/>
  </rdfs:Class>
</rdfs:RDF>
```
Because GEMET is a multilingual thesaurus, generation of an RDF encoding was done in two parts. First a file defining the conceptual backbone of GEMET in RDF was generated. Then files defining the labels for each of the concepts themes and groups were generated, one file for each language.

An extract from the GEMET conceptual backbone in RDF is below:

Extract of GEMET concept backbone in RDF

```
<rdf:RDF
 xmlns:gemet="http://www.eionet.eu.int/GEMET/skos-ext#"
 xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
 xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
 xmlns:skos="http://www.w3.org/2004/02/skos/core#"
 xmlns:dc="http://purl.org/dc/elements/1.1/"
 xml:base="http://www.eionet.eu.int/GEMET/">
  <rdf:Description rdf:about="c_204">
    <skos:inScheme rdf:resource="http://www.eionet.eu.int/GEMET/skos-ext#"/>
    <skos:broader rdf:resource="c_10217"/>
    <gemet:broaderTheme rdf:resource="t_23"/>
    <skos:narrower rdf:resource="c_7969"/>
    <gemet:broaderGroup rdf:resource="g_7956"/>
  </rdf:Description>
  <rdf:Description rdf:about="c_11786">
    <skos:inScheme rdf:resource="http://www.eionet.eu.int/GEMET/skos-ext#"/>
    <skos:broader rdf:resource="c_11124"/>
    <skos:narrower rdf:resource="c_7969"/>
    <gemet:broaderGroup rdf:resource="g_7956"/>
  </rdf:Description>
  <rdf:Description rdf:about="c_7962">
    <skos:inScheme rdf:resource="http://www.eionet.eu.int/GEMET/skos-ext#"/>
    <skos:related rdf:resource="t_36"/>
    <gemet:broaderTheme rdf:resource="t_4"/>
  </rdf:Description>
  <rdf:Description rdf:about="g_14979">
    <gemet:subGroupOf rdf:resource="sg_5499"/>
  </rdf:Description>
</rdf:RDF>
```
The stylesheet used to generate this file is at this reference [GEMET BB XSLT]. Appendix II [Appendix II] walks through this stylesheet in detail. Any of the GEMET XML source files could be used equally as the source for this transformation.

An extract from the GEMET Portuguese labels in RDF is below:

```
<rdf:RDF
  xmlns:gemet="http://www.eionet.eu.int/GEMET/skos-ext#"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:skos="http://www.w3.org/2004/02/skos/core#"
  xml:base="http://www.eionet.eu.int/GEMET/">
  <rdf:Description rdf:about="c_204">
    <skos:prefLabel xml:lang="pt">paisagens agrícolas</skos:prefLabel>
  </rdf:Description>
  <rdf:Description rdf:about="c_11786">
    <skos:prefLabel xml:lang="pt">índices bióticos</skos:prefLabel>
  </rdf:Description>
  <rdf:Description rdf:about="c_4657">
    <skos:prefLabel xml:lang="pt">ecologia paisagística</skos:prefLabel>
  </rdf:Description>
  <rdf:Description rdf:about="g_8575">
    <skos:prefLabel xml:lang="pt">COMÉRCIO, SERVIÇOS</skos:prefLabel>
  </rdf:Description>
  <rdf:Description rdf:about="t_29">
    <gemet:acronymLabel>REC</gemet:acronymLabel>
    <skos:prefLabel xml:lang="pt">turismo</skos:prefLabel>
  </rdf:Description>
</rdf:RDF>
```

The stylesheet used to generate this file is at this reference [GEMET LB XSLT]. This stylesheet was run on each of the source XML files, to generate one labels file for each language.

Note that the extracts of GEMET in RDF/XML printed above are not exactly what would be expected as the outcomes of the XSLT stylesheets printed in the appendices. This is because the output of the XSLT transformations was run through an RDF parser to validate the RDF, and the RDF parser serialised the data using a slightly different form of RDF/XML. To learn more about RDF and the RDF/XML serialisation go to these references [RDF PRIMER][RDF XML].

The full output of the GEMET transformations is at the following reference [GEMET OUT].

References

[SKOS GUIDE]
SKOS-Core 1.0 Guide. Miles, A.J., Rogers, R., Beckett, D. SWAD-Europe Thesaurus Activity.
- [http://www.w3.org/2001/sw/Europe/reports/thes/1.0/guide/]

[SKOS SCHEMA]
- [http://www.w3.org/2004/02/skos/core.rdf]

[AAPAIS]

[AAPAIS XML]

Aircraft Type Thesaurus (ATT). National Monuments Record, English Heritage.

GEneral Multilingual Environmental Thesaurus (GEMET). European Environment Information and Observation Network (EIONET).

Zthes: a Profile for Thesaurus Navigation in Z39.50 and SRW.

W3C RDF Validation Service.

Jena - A Semantic Web Framework for Java.


## Associated Files

[AAPAIS XSLT]
- http://www.w3.org/2001/sw/Europe/reports/thes/1.0/migrate/apais/apais.xslt

[AAPAIS OUT]
- http://www.w3.org/2001/sw/Europe/reports/thes/1.0/migrate/apais/apais.rdf.xml

[EH CONV]
- http://www.w3.org/2001/sw/Europe/reports/thes/1.0/migrate/eh/Converter.java

[EH TOK]
- http://www.w3.org/2001/sw/Europe/reports/thes/1.0/migrate/eh/Tokenizer.java

[SKOS_JAVA]
- http://www.w3.org/2001/sw/Europe/reports/thes/1.0/migrate/eh/SKOS.java

[EH RDF/XML]
- http://www.w3.org/2001/sw/Europe/reports/thes/1.0/migrate/eh/eh_aircraft_extract.rdf.xml

[EH RDF/N3]
- http://www.w3.org/2001/sw/Europe/reports/thes/1.0/migrate/eh/eh_aircraft_extract.rdf.n3

[GEMET BB XSLT]
- http://www.w3.org/2001/sw/Europe/reports/thes/1.0/migrate/gemet/gemet_backbone.xslt

[GEMET LB XSLT]
- http://www.w3.org/2001/sw/Europe/reports/thes/1.0/migrate/gemet/gemet_labels.xslt

[GEMET OUT]
- http://www.w3.org/2001/sw/Europe/reports/thes/1.0/migrate/gemet/gemet.zip

## Appendix I. APAIS XSLT Stylesheet Walkthrough

This section guides you through the elements of the XSL stylesheet that transforms the APAIS XML into SKOS/RDF.

The root element of the stylesheet is as follows:

```xml
<xsl:stylesheet version="1.0"
  xmlns:xsl="http://www.w3.org/1999/XSL/Transform"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:skos="http://www.w3.org/2004/02/skos/core#">
```

Stylesheet root element
The first template of the stylesheet matches the `thes` element, and sets up the root element for the target document, including all required namespace declarations:

```xml
<xsl:template match="thes">
  <rdf:RDF
    xmlns:dc="http://purl.org/dc/elements/1.1/"
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
    xmlns:skos="http://www.w3.org/2004/02/skos/core#"
      <dc:title>Australian Public Affairs Information Service (APAIS) thesaurus</dc:title>
    </skos:ConceptScheme>
    <xsl:apply-templates select="term"/>
  </rdf:RDF>
</xsl:template>
```

Note I’ve included the `ConceptScheme` declaration in this template.

Note also that the root element contains an `xml:base` declaration. This allows us to use relative URIs throughout the generated document.

The next template is used to map each preferred term in the APAIS thesaurus to an `skos:Concept` declaration:

```xml
<xsl:template match="term">
  <xsl:variable name="id" select="termId"/>
  <xsl:variable name="type" select="termType"/>
  <xsl:if test="$type='PT'">
    <xsl:choose>
      <xsl:when test="count(relation[relationType='BT'])=0">
        <skos:TopConcept rdf:about="{$id}"
          prefLabel><xsl:value-of select="termName"/>
          apply-templates select="termNote"/>
        </skos:TopConcept>
      </xsl:when>
      <xsl:otherwise>
        <skos:Concept rdf:about="{$id}"
          prefLabel><xsl:value-of select="termName"/>
          apply-templates select="termNote"/>
        </skos:Concept>
      </xsl:otherwise>
    </xsl:choose>
  </xsl:if>
</xsl:template>
```

Note that an `xsl:if` test is first used to determine whether the current term element is a preferred term. A second `xsl:choose` test is then used to determine whether the current preferred term is a top-level term in the thesaurus or not. If it is, an `skos:TopConcept` declaration is generated, otherwise an `skos:Concept` declaration is generated.

Also note that the `termId` values from the XML documents are being used as the basis for generating URIs for each concept in the thesaurus.
Further templates are then applied within the skos:Concept declaration, to generate the alternative labels, semantic relations and scope notes.

The 'relation' template generates the alternative labels and the semantic relations:

```xml
<xsl:template match="relation">
    <xsl:variable name="id" select="termId"/>
    <xsl:variable name="type" select="relationType"/>
    <xsl:if test="$type='UF'">
        <skos:altLabel><xsl:value-of select="termName"/></skos:altLabel>
    </xsl:if>
    <xsl:if test="$type='BT'">
        <skos:broader rdf:resource="{$id}"/>
    </xsl:if>
    <xsl:if test="$type='NT'">
        <skos:narrower rdf:resource="{$id}"/>
    </xsl:if>
    <xsl:if test="$type='RT'">
        <skos:related rdf:resource="{$id}"/>
    </xsl:if>
</xsl:template>
```

Finally the 'termNote' template generates any skos:scopeNote declarations:

```xml
<xsl:template match="termNote">
    <skos:scopeNote><xsl:value-of select="."/></skos:scopeNote>
</xsl:template>
```

The complete stylesheet, when applied to the above excerpt from the APAIS thesaurus, generates the following snippet of RDF/XML:

```xml
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
    xmlns:skos="http://www.w3.org/2004/02/skos/core#"
    xmlns:dc="http://purl.org/dc/elements/1.1/"
>
        df:title="Australian Public Affairs Information Service (APAIS)"
    />

    <skos:Concept rdf:about="R1722">
           rdf:label="Aboriginal archaeology"
        />
        <skos:prefLabel>Aboriginal archaeology</skos:prefLabel>
        <skos:altLabel>Aboriginal prehistory</skos:altLabel>
        <skos:altLabel>Prehistory, Aboriginal</skos:altLabel>
        <skos:related rdf:resource="R0009"/>
        <skos:broader rdf:resource="R0064"/>
        <skos:related rdf:resource="R1723"/>
        <skos:scopeNote>Use for the study of Aboriginal and Torres Strait Islander prehistoric culture chiefly through excavation and description of its material remains</skos:scopeNote>
    </skos:Concept>
    ...
</rdf:RDF>
```
The root template sets up the root element for the generated document, and applies further templates:

```xml
<xsl:template match="cds-thes">
  <rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
    xmlns:skos="http://www.w3.org/2004/02/skos/core#"
    xmlns:dc="http://purl.org/dc/elements/1.1/"
    xmlns:gemet="http://www.eionet.eu.int/GEMET/skos-ext#"
    xml:base="http://www.eionet.eu.int/GEMET/"
  >
    <skos:ConceptScheme rdf:about="http://www.eionet.eu.int/GEMET">
      <dc:title>General Multilingual Environment Thesaurus (GEMET)</dc:title>
    </skos:ConceptScheme>
    <xsl:apply-templates select="//super-group"/>
    <xsl:apply-templates select="//descriptor"/>
    <xsl:apply-templates select="//theme" mode="resource"/>
    <xsl:apply-templates select="//group" mode="resource"/>
    <xsl:apply-templates select="//descriptor-related"/>
  </rdf:RDF>
</xsl:template>
```

The 'super-group' template generates individuals of the gemet:SuperGroup class:

```xml
<xsl:template match="super-group">
  <xsl:variable name="id" select="@super-group-id"/>
  <gemet:SuperGroup rdf:about="sg_{id}">
    <skos:inScheme rdf:resource="http://www.eionet.eu.int/GEMET"/>
  </gemet:SuperGroup>
</xsl:template>
```

The 'group' template generates individuals of the gemet:Group class:

```xml
<xsl:template match="group" mode="resource">
  <xsl:variable name="id" select="@group-id"/>
  <xsl:variable name="sid" select="@super-group-id"/>
  <gemet:Group rdf:about="g_{id}">
    <skos:inScheme rdf:resource="http://www.eionet.eu.int/GEMET"/>
    <gemet:subGroupOf rdf:resource="sg_{sid}"/>
  </gemet:Group>
</xsl:template>
```

The 'theme' template generates individuals of the gemet:Theme class:

```xml
<xsl:template match="theme" mode="resource">
  <xsl:variable name="id" select="@theme-id"/>
  <gemet:Theme rdf:about="t_{id}">
    <skos:inScheme rdf:resource="http://www.eionet.eu.int/GEMET"/>
  </gemet:Theme>
</xsl:template>
```

The 'descriptor' template generates individuals of the skos:Concept class:

```xml
<xsl:template match="descriptor">
  <xsl:variable name="id" select="descriptor-term/@desc-id"/>
  <skos:Concept rdf:about="c_{id}">
    <skos:inScheme rdf:resource="http://www.eionet.eu.int/GEMET"/>
  </skos:Concept>
</xsl:template>
```
Note again that throughout identifiers from the source XML document were used as the basis for relative URIs in the generated document.

Further templates are applied within the 'descriptor' template, to generate the statements linking a concept to broader and narrower concepts, and to themes. Also, if the concept has no broader concepts, then a statement linking the concept to the corresponding group is also added via an applied template. These templates are below:

Finally the 'descriptor-related' template generates skos:related statements:

A snippet of what is generated by the backbone transformation is below:
<skos:inScheme rdf:resource="http://www.eionet.eu.int/GEMET"/>
<skos:narrower rdf:resource="c_10217"/>
<skos:broader rdf:resource="t_4648"/>
<gemet:broaderTheme rdf:resource="t_23"/>
<gemet:broaderTheme rdf:resource="t_2"/>
</skos:Concept>
...
<gemet:Group rdf:about="g_7007">
  <gemet:subGroupOf rdf:resource="sg_4044"/>
  <skos:inScheme rdf:resource="http://www.eionet.eu.int/GEMET"/>
</gemet:Group>
...
<gemet:SuperGroup rdf:about="sg_5499">
  <skos:inScheme rdf:resource="http://www.eionet.eu.int/GEMET"/>
</gemet:SuperGroup>
...
<gemet:Theme rdf:about="t_2">
  <skos:inScheme rdf:resource="http://www.eionet.eu.int/GEMET"/>
</gemet:Theme>
...
</rdf:RDF>