The Use of SKOS Vocabularies in Digital Repositories

The DSpace Case

Georgia D. Solomou
and
Theodore S. Papatheodorou

High Performance Information Systems Laboratory

Overview

- The Simple Knowledge Organization System (SKOS)
  - Why SKOS?
  - SKOS applications and tools
    - The Thesaurus of Greek Terms (TGT)
- Methods for SKOSifying Thesauri
  - The SKOSification of TGT
- The DSpace digital repository system
- Controlled Vocabularies in DSpace
  - A SKOSified controlled vocabulary in DSpace
  - Problems and solutions
- Conclusions and future work

Simple Knowledge Organization System (SKOS)

"…a common data model for sharing and linking knowledge organization systems (KOS)[1] via the Web"

[1]*taxonomies, classification schemes, subject headings, thesauri, …*

- Provides a formal language for representing any type of structured controlled vocabularies
  - A practical application of RDF (and RDFS)
  - ⇨ A ***machine-understandable*** representation framework



Why SKOS?

- A widely-recognized and adopted **standard**
  - *A W3C Recommendation since 18 August 2009*
- A means to achieve ***interoperability***

  "SKOS… enables data and technology sharing across diverse applications"
- Enables easy publication of controlled structured vocabularies for the ***Semantic Web***
  - Offers a low-cost migration path for all KOS!
- Produces descriptions in a ***machine readable format***
- … and a ***multilingual*** and ***extensible*** model

Many organizations/institutions worldwide have adopted SKOS!



Library of Congress in SKOS

- Library of Congress Subject Headings are expressed in SKOS

Ad hoc networks (Computer networks)

URI: <http://id.loc.gov/authorities/sh2007004723#concept>

Type: Topical Term

Alternate Labels: MANETs (Computer networks); Mobile ad hoc networks; Wireless ad hoc networks

Broader Terms:

- Computer networks
- Wireless communication systems

Narrower Terms:

- Cognitive radio networks
- Vehicular ad hoc networks (Computer networks)

Sources:

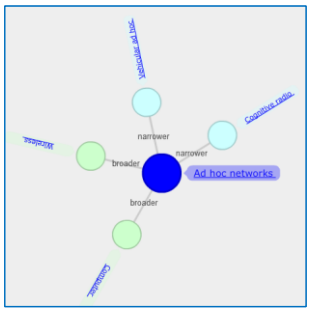
- Work cat.: 2007026336: Policy-driven ad hoc network management, c2008: ECIP galley (Mobile ad hoc networks; MANETs)
- EI, viewed June 28, 2007 (uncontrolled identifiers: Mobile ad hoc networks; MANET)
- Applied Science, viewed June 28, 2007 (Ad hoc networks)
- Inspec, viewed June 28, 2007 (controlled indexing: Ad hoc networks)
- LC data base, June 28, 2007 (in titles: Wireless ad hoc networks; MANETs; Mobile ad hoc networks; Ad hoc networks)

LC Classification: TK5105.77

Created: 2007-07-18

Last Modified: 2007-07-19 07:54:25

Alternate Formats: RDF/XML, N-Triples, JSON



… more SKOS applications

- UKAT - UK Archival Thesaurus
  - Indexes and searches in the UK archive sector
- AAT - Getty Arts and Architecture Thesaurus
  - Characterizes any type of cultural material and items of art and architecture
- AGROVOC – The Food and Agriculture Organization Thesaurus
  - Available the SKOS description for each concept
- GEMET - General Multilingual Environmental Thesaurus
  - Core terminology for the environemnt
- WordNet

… and a lot more running efforts!

The Thesaurus of Greek Terms (TGT)

- Published by the National Documentation Centre of Greece (EKT)
- The **first** official published thesaurus in Greek
  - Multi-subject thesaurus
    - … comprised of **5227 bilingual terms** (Greek, English)
  - Aiming at use/exploitation by Hellenic libraries, museums and Information Centers
- Structured as a controlled vocabulary representing both **vertical** (hierarchical) and **horizontal** associations between concepts

⇨ *But EKT hasn't proceeded yet with the SKOSification of TGT!*



SKOS tools

- ThManager
  - For **creation** and **visualization** of SKOS vocabularies
- SKOSed
  - A Protégé 4 plug-in for SKOS applications
  - Accompanied by the SKOS API
- PoolParty
  - For ***managing***, ***editing*** and ***validating*** SKOS vocabularies
- W3C Validation Service
  - On-line validator for SKOS
- The MONDECA SKOS Reader
  - For ***navigating*** and ***browsing*** SKOS thesauri (provided as files)



Bringing thesauri to SKOS

- Not yet a structured, **standardized** method
  - Only attempts that fit the needs of a particular thesaurus
  - ⇨ *Vocabularies with non-standard features cannot be handled!*

So what?

- Manual effort is required for ***mapping*** thesaurus elements to SKOS notions
- An ***XSL transformation*** usually accomplishes the migration



Structure of TGT

Based on a common subject thesaurus structure (enriched with extensions):

- Expresses three types of relationships between terms:
  - **Hierarchical** (**B**roader/**N**arrower /**M**icrothesauri **T**erm – **BT**, **NT**, **MT**)
  - **Associative** (**R**elated **T**erm - **RT**)
  - **Equivalence** for synonyms (SYN) (**U**sed **F**or - **UF**)
- English translations included (**E**nglish **T**erm - **ET**) (some derived from Eurovoc Thesaurus)
  - ⇨ Bilingual terms
- Correspondence to the **D**ewey **D**ecimal **C**lassification system (**DDC**)



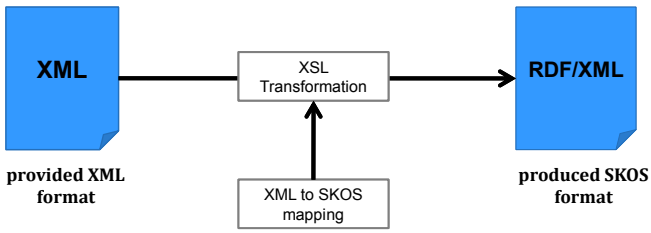
Term structure (XML format)

Term:
"civil_courts"@en
"αστικά_δικαστήρια"@el

```
<TERM>
  <CONTEXT>αστικά δικαστήρια</CONTEXT>
  <USER>EKT</USER>
  <MT>Νομικές Επιστήμες</MT>
  <ET>civil courts</ET>
  <BT>δικαστήρια</BT>
  <NT>Άρειος Πάγος<NT>
  <NT>ειρηνοδικεία<NT>
  <NT>εφετεία<NT>
  <NT>πρωτοδικεία<NT>
  <UF>βλ. πολιτικά δικαστήρια</UF>
  <RT>πολιτική δικονομία</RT>
  <RT>αστικό δίκαιο</RT>
  <SN>some description</SN>
  <dewey>347</dewey>
</TERM>
```


The SKOSification of the EKT thesaurus

1. Manual mapping of the thesaurus (XML) elements to SKOS notions
   - *Based on SKOS specification*
2. Creation of the appropriate XSL Transformation
3. Application of XSLT document to XML file



Mapping Summary

XML element		SKOS notion
<TERM>	The described term	<skos:Concept>
<USER>	Thesaurus' owner	-
<CONTEXT>	Term's label	<skos:prefLabel lang="el">
<MT>	Microthesauri term	<skos:broaderTransitive>
<ET> (first)	English translation	<skos:prefLabel lang="en">
<ET>	Alternative English translation	<skos:altLabel lang="en">
<BT>	Broader term	<skos:broader>
<NT>	Narrower term	<skos:narrower>
<RT>	Related term	<skos:related>
<UF>	A synonym	<skos:altLabel lang="el">
<SN>	A short description	<skos:definition>
<DEWEY>	A number indicating the correspondence to Dewey system	<skos:notation>

Term structure (XML format)

Term:
"civil_courts"@en
"αστικά_δικαστήρια"@el

```
<TERM>
  <CONTEXT>αστικά δικαστήρια</CONTEXT>
  <USER>ΕΚΤ</USER>
  <MT>Νομικές Επιστήμες</MT>
  <ET>civil courts</ET>
  <BT>δικαστήρια</BT>
  <NT>Άρειος Πάγος<NT>
  <NT>ειρηνοδικεία<NT>
  <NT>εφετεία<NT>
  <NT>πρωτοδικεία<NT>
  <UF>βλ. πολιτικά δικαστήρια</UF>
  <RT>πολιτική δικονομία</RT>
  <RT>αστικό δίκαιο</RT>
  <SN>some description</SN>
  <dewey>347</dewey>
</TERM>
```

Example of a concept in SKOS

Term (concept): "civil_courts"@en
"αστικά_δικαστήρια"@el

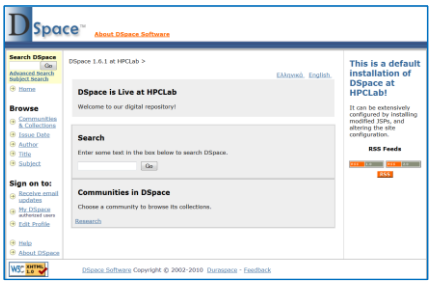
```
<skos:Concept rdf:about="http://www.hpclab.ceid.upatras.gr/skos/ekt#αστικά_δικαστήρια">
  <skos:prefLabel xml:lang="el">αστικά δικαστήρια</skos:prefLabel>
  <skos:prefLabel xml:lang="en">civil courts</skos:prefLabel>
  <skos:altLabel xml:lang="el">πολιτικά δικαστήρια</skos:altLabel>
  <skos:broaderTransitive rdf:resource="http://www.hpclab.ceid.upatras.gr/skos/ekt#Νομικές_επιστήμες"/>
  <skos:broader rdf:resource="http://www.hpclab.ceid.upatras.gr/skos/ekt#δικαστήρια"/>
  <skos:narrower rdf:resource="http://www.hpclab.ceid.upatras.gr/skos/ekt#'Αρειος_Πάγος"/>
  <skos:narrower rdf:resource="http://www.hpclab.ceid.upatras.gr/skos/ekt#ειρηνοδικεία"/>
  <skos:narrower rdf:resource="http://www.hpclab.ceid.upatras.gr/skos/ekt#εφετεία_(αστικά_δικαστήρια)"/>
  <skos:narrower rdf:resource="http://www.hpclab.ceid.upatras.gr/skos/ekt#πρωτοδικεία"/>
  <skos:related rdf:resource="http://www.hpclab.ceid.upatras.gr/skos/ekt#πολιτική_δικονομία"/>
  <skos:notation rdf:datatype="http://dewey.info/schema-terms/Notation">347</skos:notation>
</skos:Concept>
```

The TGT thesaurus in SKOS is available at:
http://swig.hpclab.ceid.upatras.gr/SKOS?action=AttachFile&do=get&target=ekt_to_skos.rdf

About DSpace

*A mechanism for the efficient **description**, **preservation, management**, **exploitation** and **distribution** of any kind of digitized material*

- A **Digital Library System:**
  "… used by museums, state archives, state and national libraries, journal repositories, consortiums, and commercial companies to manage their digital assets"



http://www.dspace.org/

Articles, Books, Journal Papers, Images, Videos, 3D Objects, Data Sets, Learning Objects, …

Semantic Web And Reasoning for Cultural Heritage and Digital Libraries 2010 (*SWARCH-DL*) September 22-24 2010, Pittsburg, PA, USA

Usage of Controlled Vocabularies

- Refinement of the set of keywords used:
  - during item description in the **submission process**
  - when **browsing** by subject
- **Search** in subject fields

Subject Search in DSpace

Controlled Vocabularies in DSpace

- Supported controlled vocabularies are expressed in a simple XML format (*"Node Schema"*)

DSpace Node Schema

```
<node id="acmccs98" label="ACMCCS98">
  <isComposedBy>
    <node id="A." label="General Literature">
      <isComposedBy>
        <node id="A.0" label="GENERAL"/>
        <node id="A.1" label="INTRODUCTORY AND SURVEY"/>
        …
  </isComposedBy>
</node>
```

  - Each term is represented as a `<node>`, characterized by a unique **ID** and a lexical **Label**
  - `<isComposedBy>` is used for narrower relationships



Controlled Vocabulary add-on for DSpace

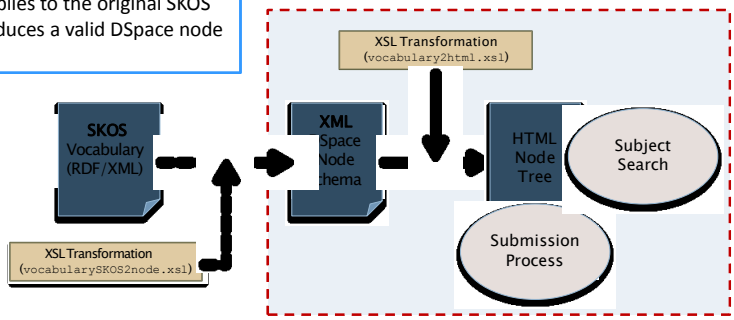
by the Odisseia Research Group at the University of Minho

1. Updated node schema supporting more types of relationships and/or properties
   - Provision for associative relationships (Related Terms)
   - Allows for the use of preferred terms (Use-instead Terms)
2. Recognizes thesaurus/controlled vocabularies **expressed in SKOS**
3. Possibility to assign distinct vocabularies to specific communities



Vocabulary Ingestion Process (add-on)

1st XSLT: Applies to the original SKOS file and produces a valid DSpace node schema.



2nd XSLT: Converts the inherent node schema to an HTML node tree (taxonomy)

The TGT thesaurus in DSpace

Part of the SKOSified TGT Thesaurus in DSpace

Problems

Two main problems in the construction of the presented taxonomy:

1. Incorrect rendering in the tree hierarchy
   a. Some terms may appear in the **wrong level/depth**
   b. … or may be **repeated** (both as top level concepts and sub-terms)
2. Incomplete rendering in the tree hierarchy
   a. Some terms may be **missing**



Why?

- Provided XLST does not handle every case
  - No provision for broader terms
  - ⇨ *Problems 1a, 1b*
- TGT implementation is not exhaustive
  - Not every possible relation is explicitly asserted
    - *but semantically consistent!*
  - ⇨ *Problem 2a*

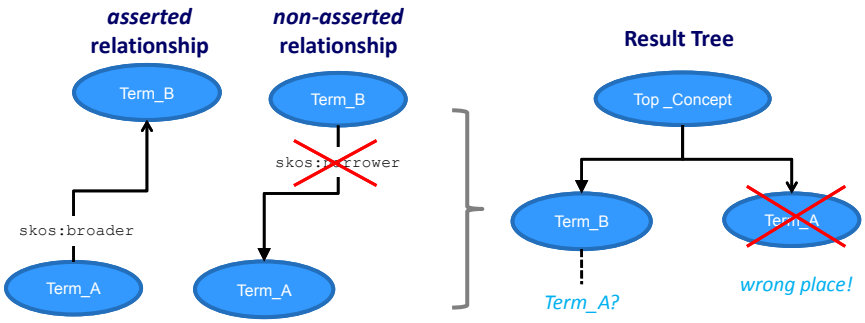


… for example

Problem 1a: *Some terms may appear in the wrong level/depth*

Reason:

- Handling for only narrower (and not broader) terms!



Solving some problems

- Modification of the 1st level XSL Transformation

  (SKOS ⇨ DSpace node schema)
  - Provision for ***all type*** of relationships between concepts
    - … and the ***broader*** ones

Result

⇨ Creation of an accurate taxonomy!

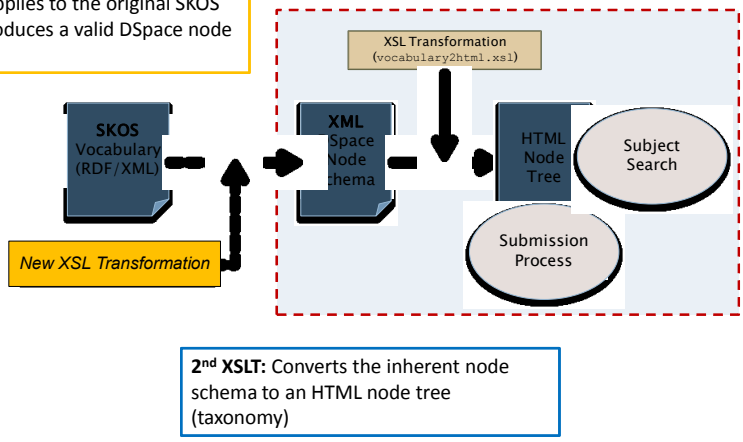
- ⇨ **No** wrong placement of terms (*Problem 1a*)
- ⇨ **No** repetitions (*Problem 1b*)

But…

- *…the "missing terms" problem remains unsolved*

Ingestion Process (with modified XSLT)

1st XSLT: Applies to the original SKOS file and produces a valid DSpace node schema.



2nd XSLT: Converts the inherent node schema to an HTML node tree (taxonomy)

Possible Solution: OWL Ontologies

- SKOS *is* (in) OWL
  - Could exploit semantic relations and axioms
  - Enables reasoning

⇨ The TGT thesaurus as an OWL ontology

- Programming access to the thesaurus elements
- Exploitation of the **OWL API** for parsing thesauri ontologies (expressed in RDF/XML format)
- A simpler way to construct the node tree (instead of complex XSL Transformations)
  - Correct term rendering

Possible Solution: Using Reasoners

- A reasoning based approach
  - Apply an OWL **reasoner** (e.g. FaCT++, Pellet) to the SKOS thesaurus/ontology
  - ⇨ *"Missing" relations could be **inferred***
  - ⇨ *Inferenced-based classification and rendering of the thesaurus*



Summary

- Very important to have controlled vocabularies (thesauri) expressed in SKOS
- Utilization of SKOS vocabularies by digital library systems
- Better handling of SKOS vocabularies when using OWL API and reasoners



Thank you for your attention!

Questions?