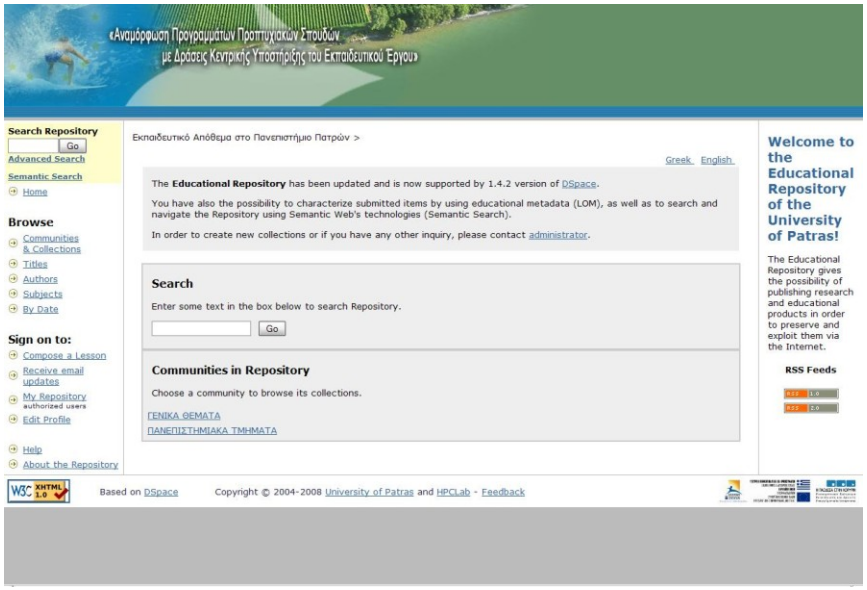
Support of SKOS Vocabularies in Digital Repository Systems –The Case of the University of Patras Institutional Repository

Dimitrios Koutsomitropoulos
Georgia Solomou

High Performance Information Systems Laboratory
University of Patras

University of Patras Institutional Repository

A mechanism for the efficient **description**, **preservation, management**, **exploitation** and **distribution** of the University's educational and scientific material



http://repository.upatras.gr/dspace

- Built upon the open-source **DSpace** digital repository system
- Item description using the **Dublin Core** metadata schema

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

HPCLab
University of Patras

University of Patras Institutional Repository

Additional features

- Multilingual support
  - User Interface (Greek, English, …)
  - Metadata - Characterization of items in more than one languages
- Advanced search service
  - Full text
  - Metadata
  - **Semantic Search**
- Advanced browsing
  - Semantic navigation

Controlled Vocabularies in DSpace

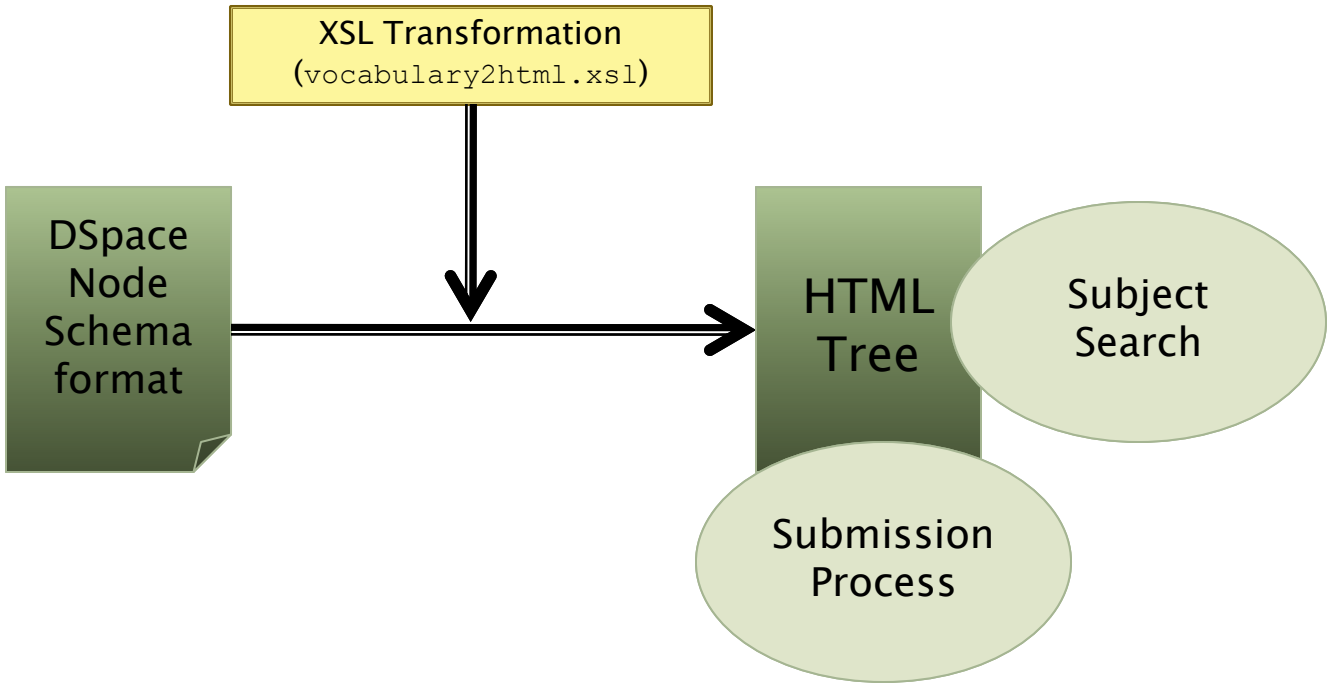
- Support of Controlled Vocabularies expressed in XML format (*"Node Schema"*)

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

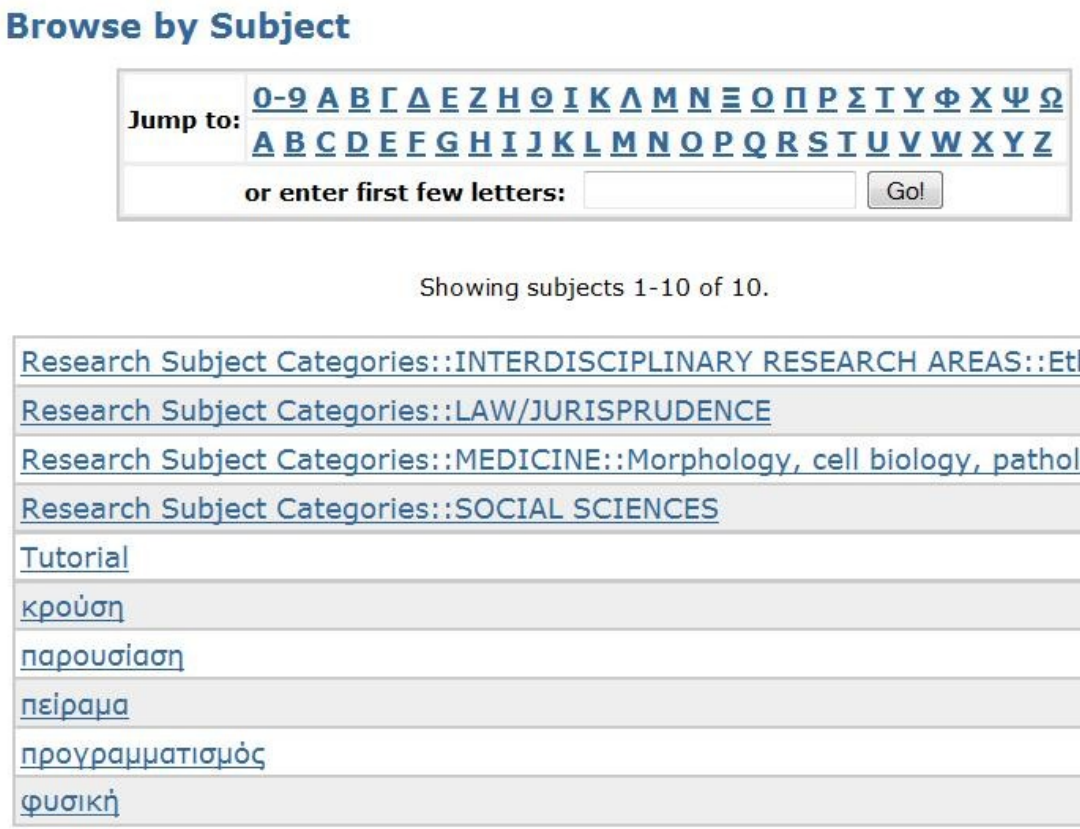
Vocabulary Ingestion Process



Controlled Vocabularies in DSpace

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

Browse by Subject



Browse by Subject

Jump to: 0-9 A B Γ Δ E Z H Θ I K Λ M N Ξ O Π P Σ T Y Φ X Ψ Ω
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

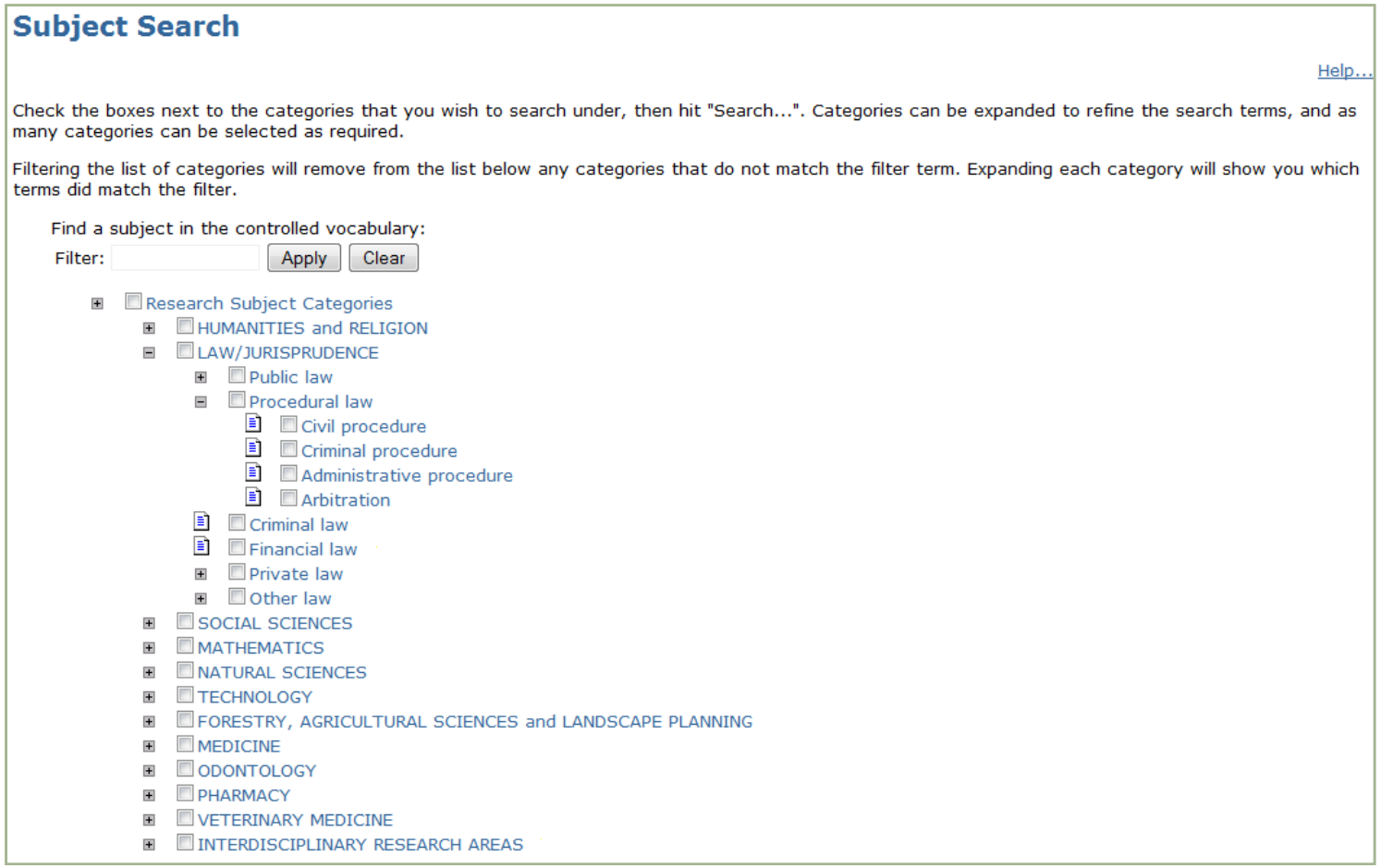
or enter first few letters: Go!

Showing subjects 1-10 of 10.

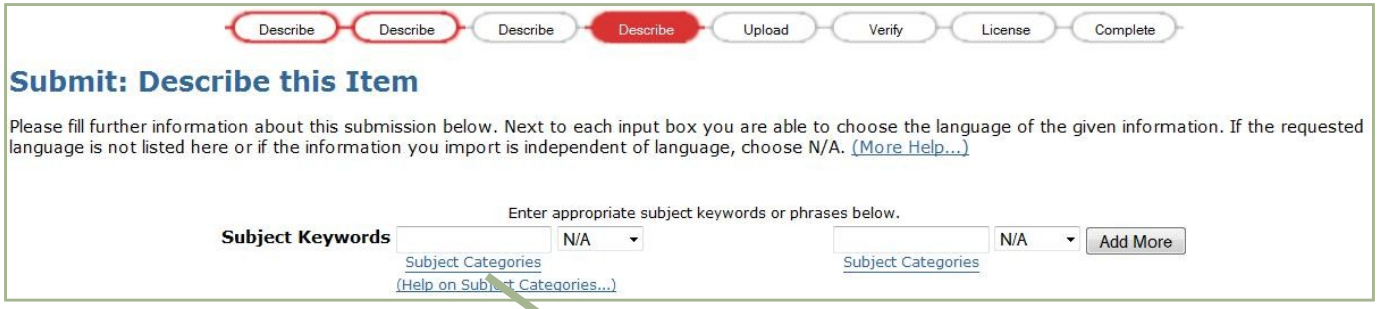
- Research Subject Categories::INTERDISCIPLINARY RESEARCH AREAS::Ethnicity
- Research Subject Categories::LAW/JURISPRUDENCE
- Research Subject Categories::MEDICINE::Morphology, cell biology, pathology
- Research Subject Categories::SOCIAL SCIENCES
- Tutorial
- κρούση
- παρουσίαση
- πείραμα
- προγραμματισμός
- φυσική

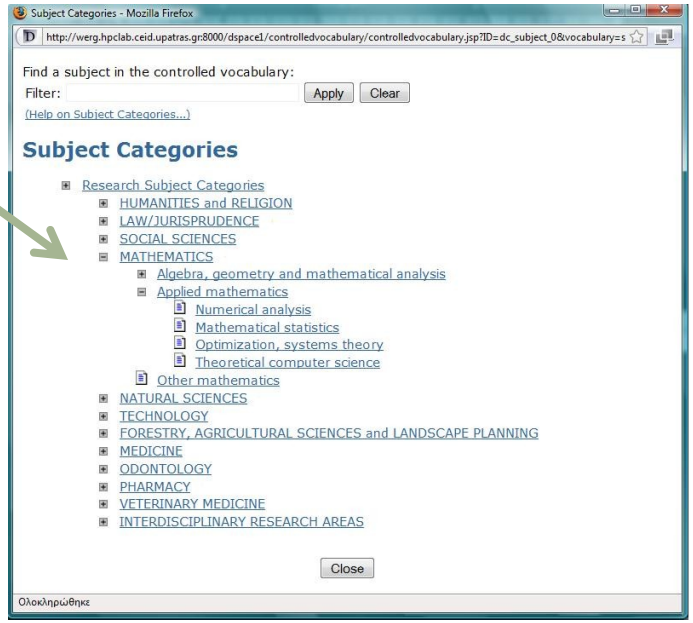
Controlled Vocabulary Terms

Subject Search



Controlled Vocabularies in the Submission Process



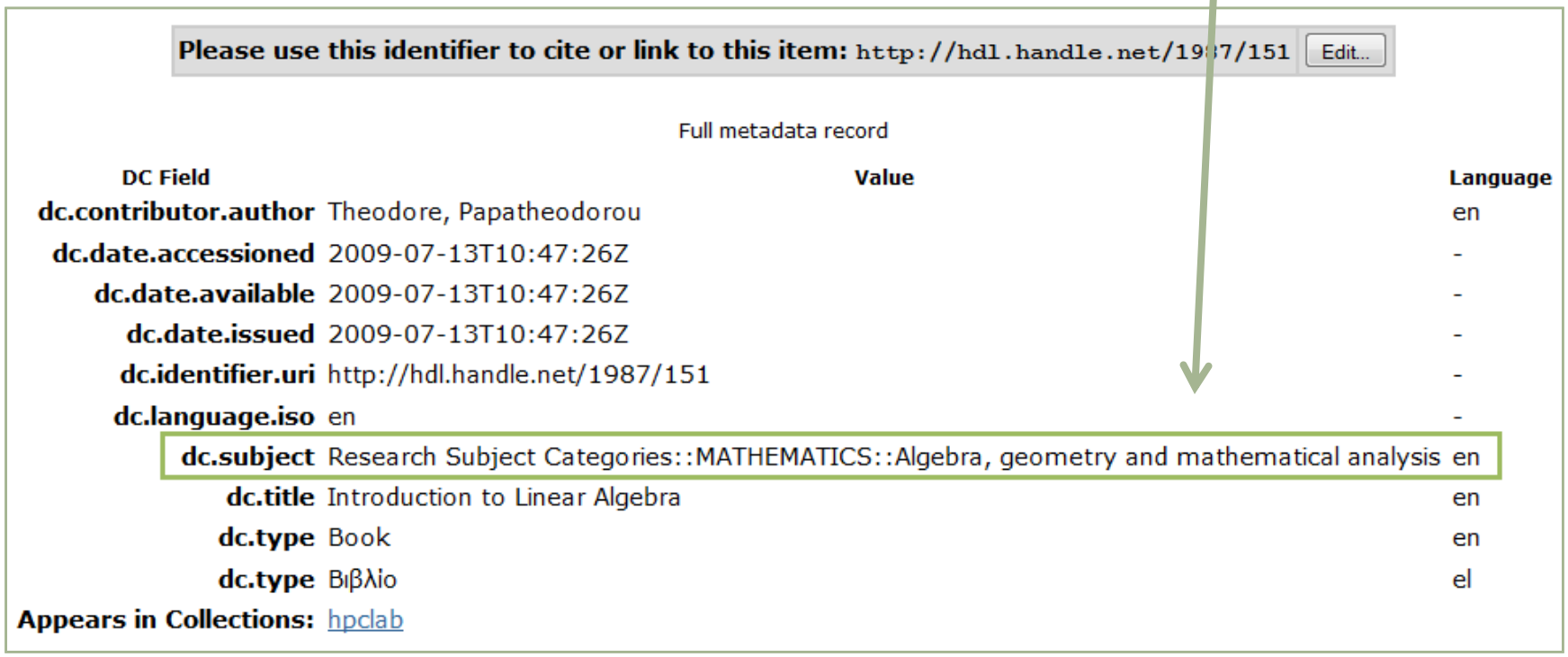


- Auto-fill subject field using proposed terms (pop-up window)
- Use of only one controlled vocabulary

HPCLab
University of Patras

DSpace Item Metadata

Subject coming from the controlled vocabulary terms



Please use this identifier to cite or link to this item: http://hdl.handle.net/1987/151 Edit...

Full metadata record

DC Field	Value	Language
dc.contributor.author	Theodore, Papatheodorou	en
dc.date.accessioned	2009-07-13T10:47:26Z	-
dc.date.available	2009-07-13T10:47:26Z	-
dc.date.issued	2009-07-13T10:47:26Z	-
dc.identifier.uri	http://hdl.handle.net/1987/151	-
dc.language.iso	en	-
dc.subject	Research Subject Categories::MATHEMATICS::Algebra, geometry and mathematical analysis	en
dc.title	Introduction to Linear Algebra	en
dc.type	Book	en
dc.type	Βιβλίο	el
Appears in Collections:	hpclab	

Item's Metadata

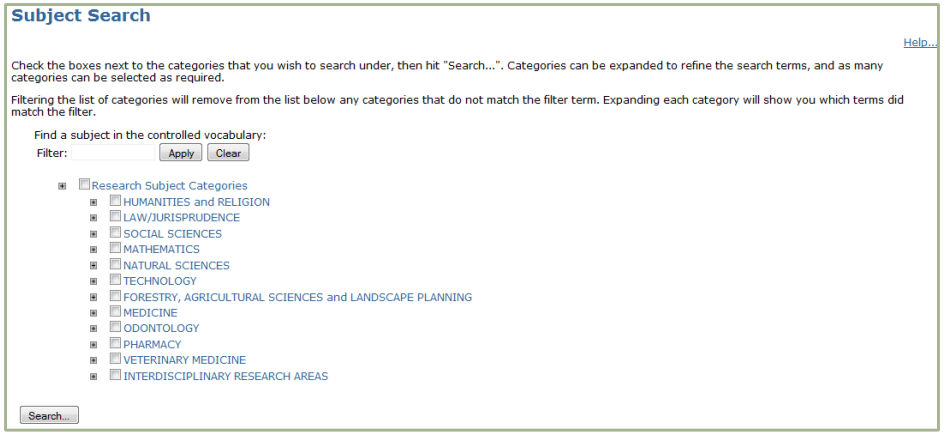
Controlled Vocabularies in DSpace

Additional features

- Support for multilingual vocabularies
  - one file for each translation (language)
  - use of the language code in the name of the file (e.g. `voc_el.xml` for Greek)

implemented by HPCLab, University of Patras

Subject Search in Greek Interface



Subject Search in English Interface

HPCLab
University of Patras

Structure of Controlled Vocabularies in DSpace

- Vocabulary formatted in a simple XML structure ⇨ not a **standard**!
- Only a simple hierarchical (narrower) relationship is expressed
  - `<isComposedBy>`

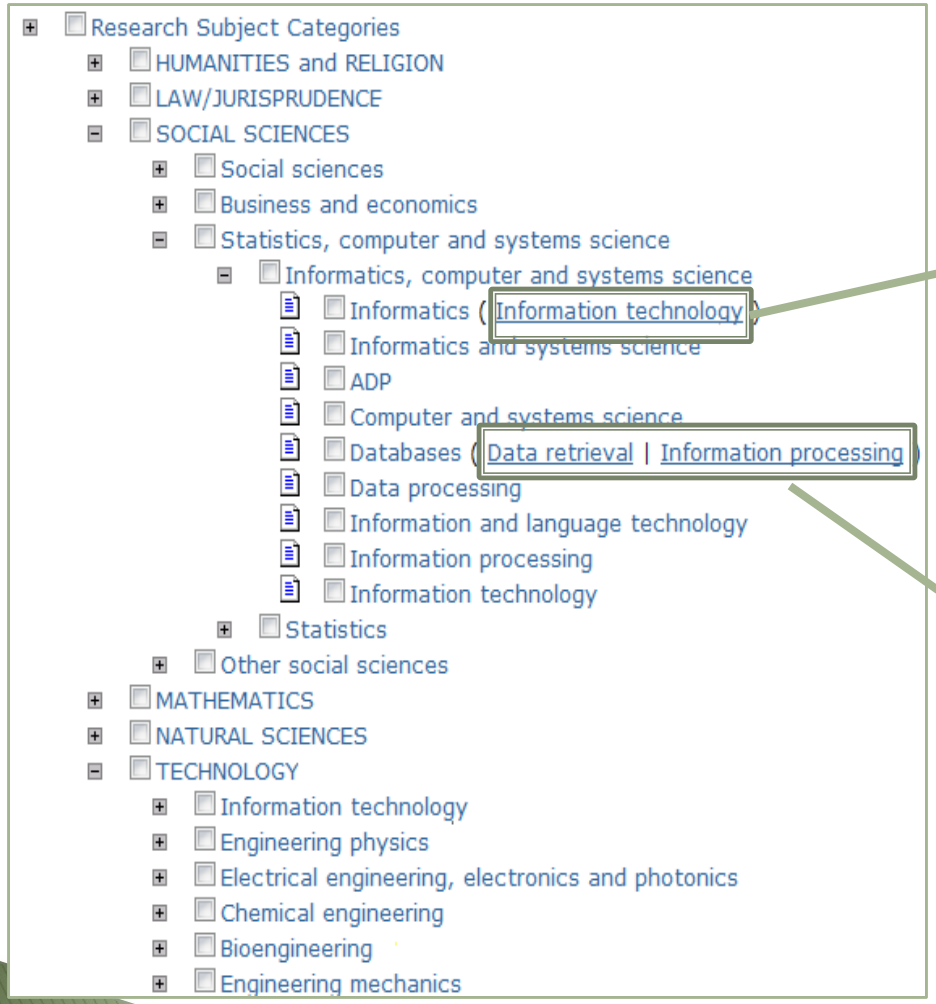
  ⇨ loose type of relationships among terms
- About SKOS
  - very close to becoming a standard
  - provides richer types of relationships
    - Hierarchical (broader/narrower)
    - Associative (related)

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)

Updated Node Schema



Use-instead Term
Use “Information Technology” *instead of* “Informatics”

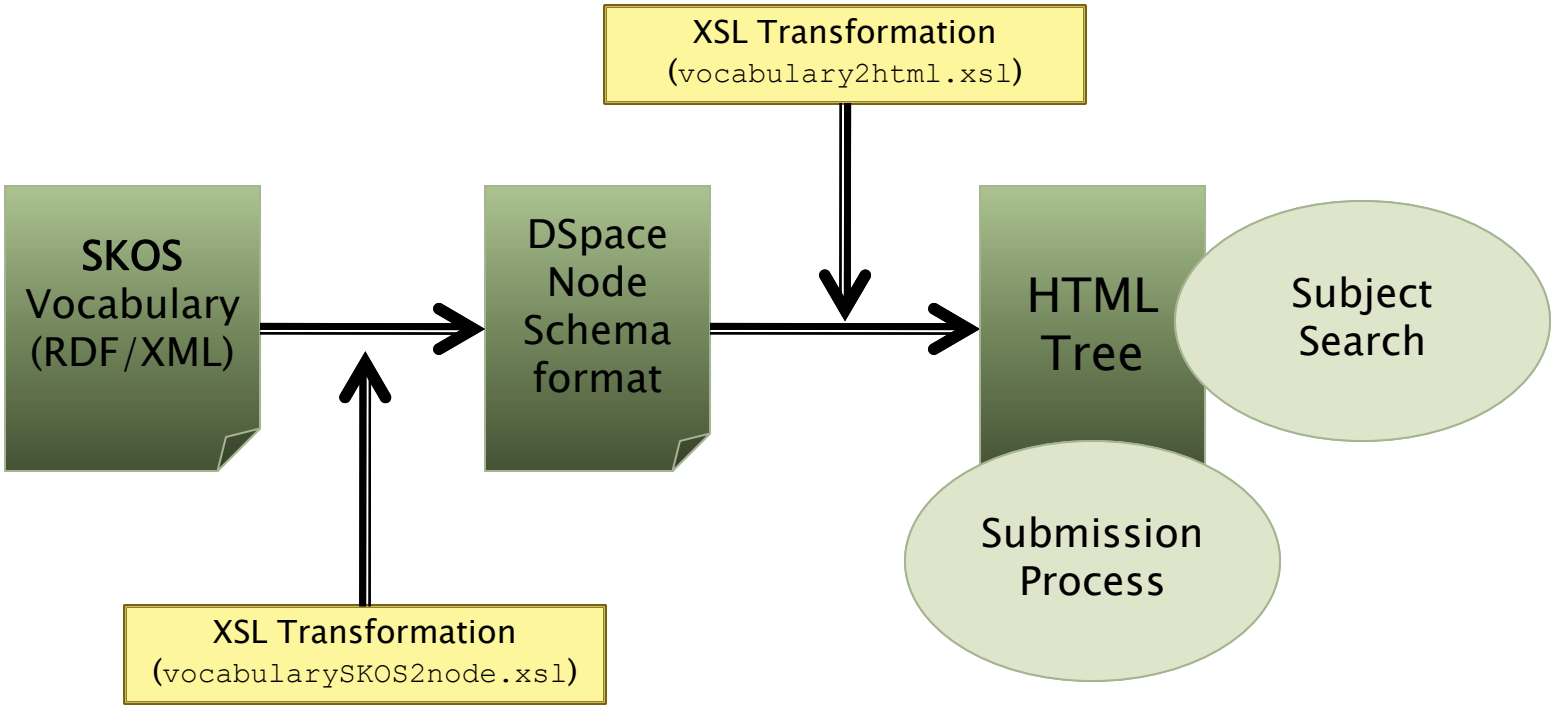
Related Term(s)
“Data retrieval” and “Information processing” *relate to* “Databases”

Controlled Vocabulary add-on for DSpace

What is more:

2. Recognizes thesaurus/controlled vocabularies **expressed in SKOS**
   - RDF/XML format
3. Possibility to assign distinct vocabularies to specific communities
   - Use of the community's particular vocabulary when filling subject fields in the submission process

Vocabulary Ingestion Process *(add-on)*



Application of the Controlled Vocabulary add-on at the University of Patras DSpace Installation

Applied changes

- Updated node schema
  - Parsing and rendering of *Related* and *Use-instead* Terms (`vocabulary2html.xsl`, `ControlledVocabularyTag.java`)
- Support for **SKOS** syntax
  - Adoption of the provided XSL Transformation
  - (`vocabularySKOS2node.xsl`)

Problems

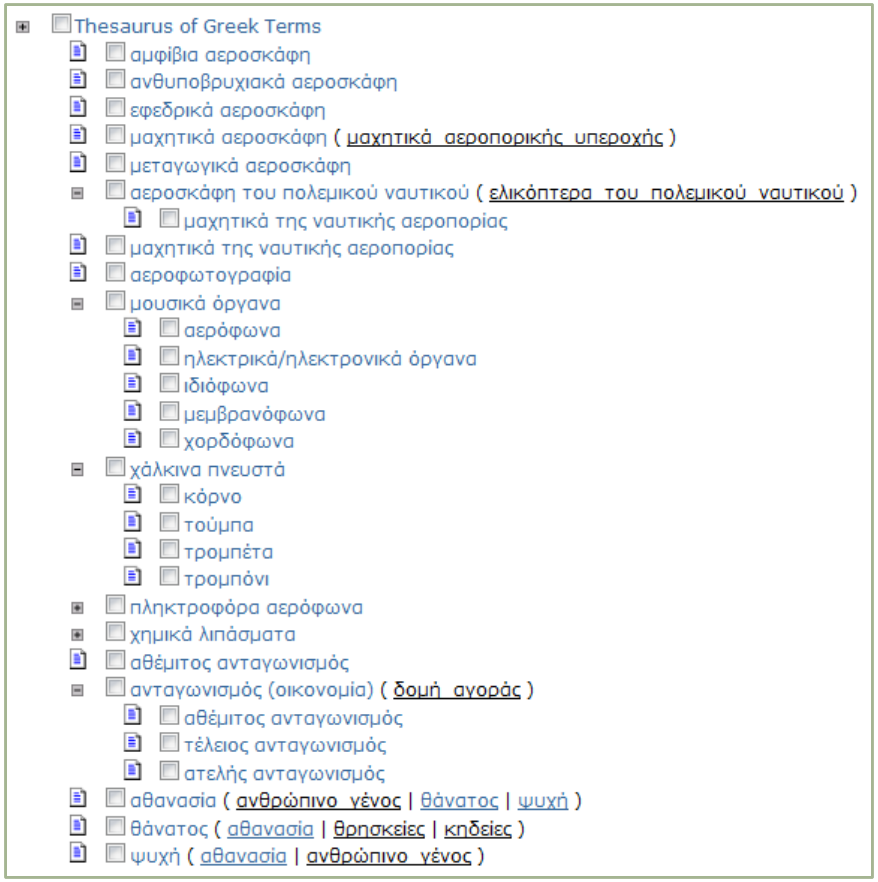
- Only those narrower terms are handled that appear in the thesaurus as separate concepts themselves
  - solved by correcting the skos-to-node XSLT transformation file
- Repetitions/missing terms in the hierarchical form

The EKT Thesaurus in Greece

- Provides a Controlled vocabulary based on
  - the National Library of Greece subject headings
  - subject terms used by the Hellenic Libraries
- Aiming at use/exploitation by all Hellenic Libraries and Information Centres
- Bilingual terms (Greek, English)
- The first such vocabulary aiming at being established as a standard in Greece
  - Incorporation in the Hellenic Public Libraries Union Catalogue

EKT Thesaurus in DSpace

- Part of the SKOSified EKT Thesaurus in DSpace
  - Use of the produced RDF/XML file format
    - requires file extension `.skos`
  - File name augmented by the language code (`_el` for Greek)



EKT Thesaurus in DSpace: Problems

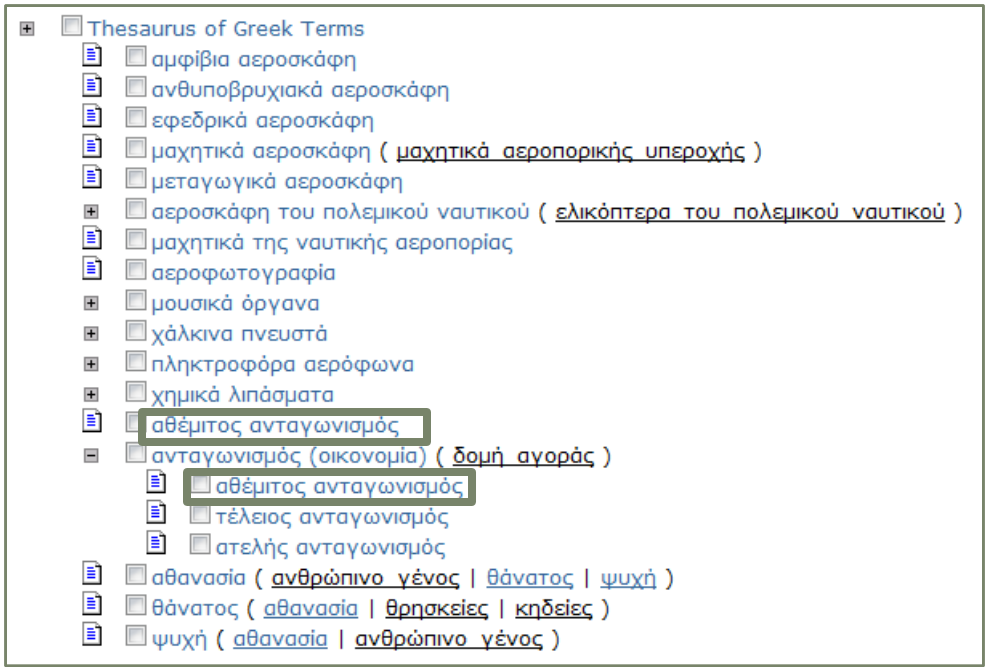
- **Incorrect** rendering in the tree hierarchy
  - Some terms may appear in the wrong level/depth
- **Incomplete** rendering in the tree hierarchy
  - Some terms may be missing
- Why?
  - Provided XLST does not handle every case
  - EKT implementation is not exhaustive
    - Not every possible relation is explicitly asserted
    - (but semantically consistent)

EKT Thesaurus in DSpace – Problems in the Hierarchical Form

1. **Repetition of terms**
   - Some terms appear both **stand-alone** as well as **sub-terms** of other terms in the tree hierarchy

Reason

⇨ Each **<node>** in the node schema will appear at the top level of the hierarchy, regardless of its possible reference as a sub-(or super-) term, by another concept



EKT Thesaurus in DSpace – Problems in the Hierarchical Form

2. **Missing terms**

   - Thesaurus **top concepts** are not present in the node tree

Reason

⇨No separate concept description is provided for these concepts (→*no separate* `<node>` *element exists*)

   - Terms that appear *only as* ***broader*** terms are not included

Reason

⇨No handling of broader terms during parsing

EKT Thesaurus in DSpace – Problems in the Hierarchical Form

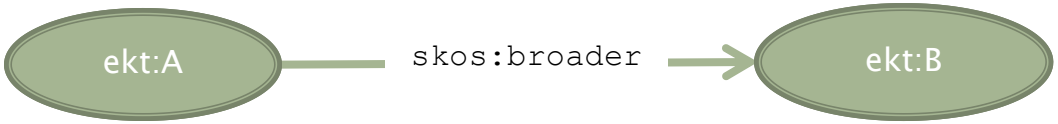
3. **Wrong place of some terms in the node tree**

Reason

⇨ Handling for only narrower (and not broader) terms

Example

asserted relationship:



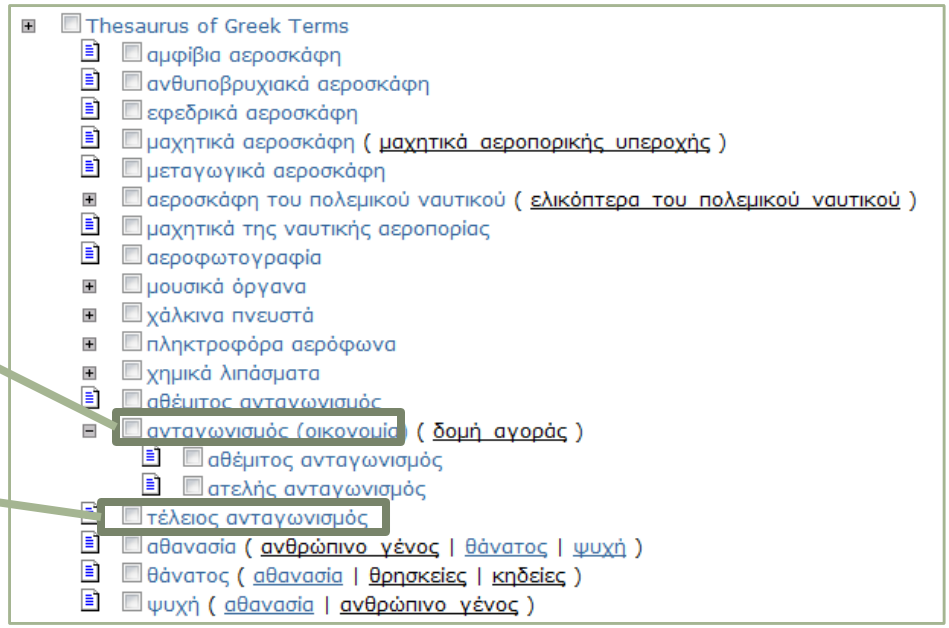
non-asserted relationship:

Term A is rendered as a top term and not under term B!

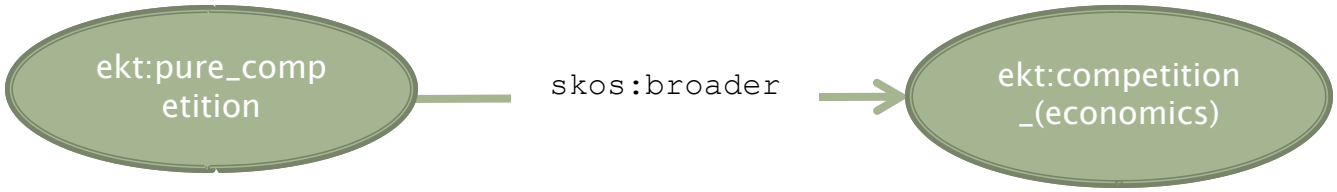
Wrong Place of a Term in the Hierarchical Form

Term:
"ανταγωνισμός (οικονομία)"@el
"competition (economics)"@en

Term:
"τέλειος ανταγωνισμός"@el
"pure competition"@en



asserted relationship:



non-asserted relationship:

HPCLab
University of Patras

Possible Solution: OWL Ontologies

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

    ⇨ No repetitions
- 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

Inference-based Classification and Rendering: Example 1

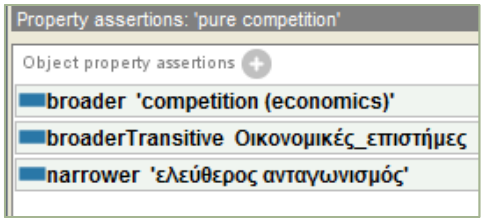
- Construct the top-level hierarchy
  - Possible algorithm

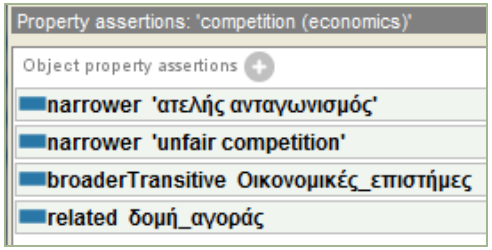
```
find every skos:hasTopConcept term TC;
for-each TC {
   find every skos:narrowerTransitive term NT;
     for-each NT {
        find a skos:broader term BT;
        if no such BT exists then
            add TC skos:narrower NT;
     } //for-each NT
} //for-each TC
```

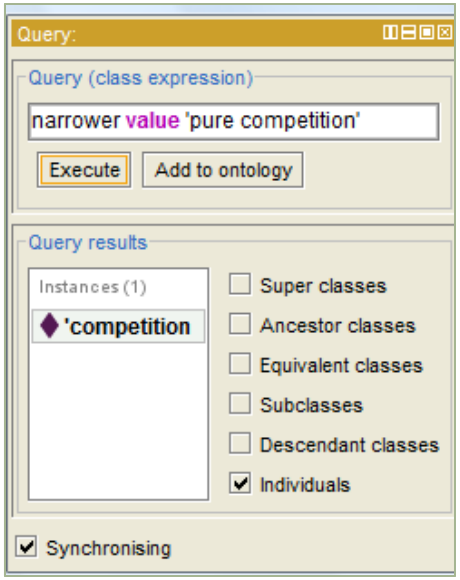
- Result
  - Top concepts appear (correctly)
  - Direct descendants of top concepts appear in their right place

Inference-based Classification and Rendering: Example 2

- Handling of all types of relationships (even of broader)
  - ⇨Terms in their right place (under their broader ones)
  - ⇨No missing terms







Inferenced-based query through Protégé 4: the term 'pure competition' is inferred *to be narrower of 'competition' even though this is not explicitly asserted*