A self-improving service to repositories & aggregators for massively enriching their content

DHC 2016 workshop
22.11.2016

Haris Georgiadis PhD
Senior Software Engineer/Researcher
National Documentation Centre | EKT
EKT is the main Greek scientific content provider

We aggregate, collect, document, preserve and disseminate authoritative digital content & data, produced and used by the Greek and international scientific, research and cultural communities.

- Digital repositories
- openABEKT (proprietary Integrated Library System)
- Large scale aggregators of authoritative Greek scientific and cultural content
  - openarchives.gr
  - searchculture.gr
Semantics.gr: Information system for vocabularies & semantic enrichment

- A pilot system where EKT and other authoritative institutions will create, establish & link their own semantic vocabularies and thesauri
  - hosts vocabularies of virtually any schema
  - supports hierarchies and multilingualism
  - supports links to external vocabularies & thesauri
- Bilingual Open portal
  - Published vocabularies as Open Linked Data
  - Search engine
  - Hierarchical navigation
- Metadata enrichment tool
Creating a vocabulary in semantics.gr

- Authorized users can create RDF based vocabulary schemata.
  - Create parametric owl properties
  - Group owl properties in owl classes
  - Successfully modelled skos:Concept, edm:Place, edm:Timespan, edm:Agent
- Institutions will obtain user accounts and create their own vocabularies and thesauri.
  - The system provides with a dynamic form to create vocabulary entries.
  - The form reflects the parameters set.
- Institutions can publish their vocabularies. Those vocabularies are accessible through the portal as Open Linked Data.
Searchculture.gr: aggregator of cultural heritage digital content

- Aggregates digital content from registered digital repositories
- Public portal
  - Modern & effective search engine
  - Dissemination as Open Linked Data
  - Multilingual (English / Greek)
  - Rights statements per item, mostly CC, defining copyright, accessibility and reusability status
  - Presents the participating institutions & their collections

150,000+ items | 53 collections | 42 institutions
Heterogeneity in Metadata

- Heterogeneity in **dc:type** metadata field:
  - **Representation-related** ex. different languages, synonyms, plural / singular numbers, different case styles
  - **Documentation-related** ex. use of very broad or very narrow (specialized) terms
- Negative impact of heterogeneity in user experience:
  - **Searchability & multilingual search**
  - **Discoverability & multilingual navigation**
  - **Visual presentation**
## Four different types of repositories we came across

<table>
<thead>
<tr>
<th>Documentation quality class</th>
<th>Documentation quality class description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td><strong>Good</strong> documentation of <code>dc:type</code></td>
</tr>
<tr>
<td>B</td>
<td><strong>Extremely specialized</strong> documentation of <code>dc:type</code></td>
</tr>
<tr>
<td>C</td>
<td><strong>Insufficient</strong> documentation on <code>dc:type</code>, useful <code>dc:subject</code></td>
</tr>
<tr>
<td>D</td>
<td><strong>Insufficient</strong> documentation on <code>dc:type</code>, useful <code>dc:title</code></td>
</tr>
</tbody>
</table>
Semantics.gr enrichment tool

- **Mapping rules** from metadata field values (ex. dc:type) to vocabulary terms per collection/repository
- **Automatic mapping suggestion** enhanced by a self-improving mechanism
- Use of **secondary metadata field** values (ex. dc:subject) as **filters**
  - Supports searching for **key-terms** inside **descriptive fields** (ex. dc:title) as **filters**
- Curated **complex expressions on filters** in order to create finer and more precise rules and avoid false positives
- A mapping rule set assigned to a repository/collection is available through a **REST API in json format**
Class A enrichment: automatic mapping metadata values to vocabulary entries

Vocabulary V1

<table>
<thead>
<tr>
<th>dc:type value</th>
<th>Entry from vocabulary V1</th>
</tr>
</thead>
<tbody>
<tr>
<td>sculpture art (120 items)</td>
<td><a href="http://scs.gr/sculpture">http://scs.gr/sculpture</a></td>
</tr>
<tr>
<td>greek vases (230 items)</td>
<td><a href="http://scs.gr/vase">http://scs.gr/vase</a></td>
</tr>
<tr>
<td>jewelleries (135 items)</td>
<td><a href="http://scs.gr/Jewellery">http://scs.gr/Jewellery</a></td>
</tr>
</tbody>
</table>
Class B enrichment:
the self-improving mechanism

Vocabulary V1

- http://scs.gr/sculpture
  skos:prefLabel "Sculpture"@en | "Γλυπτό"@el
- http://scs.gr/figurine
  skos:prefLabel "Figurine"@en | "Στатυήσ"@el keywords: "statuette"
- http://scs.gr/Jewellery
  skos:prefLabel "Jewellery"@en | "Κοσμήμα"@el keywords: "earring"
- http://scs.gr/vessel
  skos:prefLabel "Vessel"@en | "Σκυτός"@el
- http://scs.gr/vase
  skos:prefLabel "Vase"@en | "Αγγείο"@el keywords: "amphora", "oenochoe"

<table>
<thead>
<tr>
<th>dc:type values</th>
<th>Entry from vocabulary V1</th>
</tr>
</thead>
<tbody>
<tr>
<td>amphora (100 items)</td>
<td><a href="http://scs.gr/vase">http://scs.gr/vase</a></td>
</tr>
<tr>
<td>oenochoe (110 items)</td>
<td><a href="http://scs.gr/vase">http://scs.gr/vase</a></td>
</tr>
<tr>
<td>earring (55 items)</td>
<td><a href="http://scs.gr/Jewellery">http://scs.gr/Jewellery</a></td>
</tr>
<tr>
<td>statuette (15 items)</td>
<td><a href="http://scs.gr/figurine">http://scs.gr/figurine</a></td>
</tr>
</tbody>
</table>
## Class C enrichment:
Using secondary fields as filters

### Vocabulary V1

- **http://scs.gr/sculpture**
  - skos:prefLabel "Sculpture"@en | "Γλυπτό"@el **keywords:** “statue”

- **http://scs.gr/figurine**
  - skos:prefLabel "Figurine"@en | "Στάσιμο"@el **keywords:** “statuette”

- **http://scs.gr/Jewellery**
  - skos:prefLabel "Jewellery"@en | "Κοσμήμα"@el **keywords:** “earring”

- **http://scs.gr/vessel**
  - skos:prefLabel "Vessel"@en | "Σκύφος"@el
  - http://scs.gr/vase
    - skos:prefLabel "Vase"@en | "Αγγείο"@el **keywords:** “amphora”, “oenochoe”

<table>
<thead>
<tr>
<th><strong>dc:type values</strong></th>
<th><strong>Filters: dc:subject</strong></th>
<th><strong>Entry from vocabulary V1</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>ceramic objects</td>
<td>amphora (↗) , vase (↗), statuette (↗), figurine (↗), ...</td>
<td><a href="http://scs.gr/vase">http://scs.gr/vase</a></td>
</tr>
<tr>
<td>(101 items)</td>
<td></td>
<td>if filter in [&quot;vase&quot;, &quot;amphora&quot;] auto</td>
</tr>
<tr>
<td></td>
<td></td>
<td><a href="http://scs.gr/figurine">http://scs.gr/figurine</a></td>
</tr>
<tr>
<td></td>
<td></td>
<td>if filter in [&quot;statuette&quot;, &quot;figurine&quot;] auto</td>
</tr>
<tr>
<td>exhibits (55 items)</td>
<td>earing (↗), amphora (↗), ...</td>
<td><a href="http://scs.gr/Jewellery">http://scs.gr/Jewellery</a></td>
</tr>
<tr>
<td></td>
<td></td>
<td>if filter in [&quot;earing&quot;] auto</td>
</tr>
<tr>
<td></td>
<td></td>
<td><a href="http://scs.gr/vase">http://scs.gr/vase</a></td>
</tr>
<tr>
<td></td>
<td></td>
<td>if filter in [&quot;amphora&quot;] auto</td>
</tr>
<tr>
<td></td>
<td></td>
<td>but <strong>NOT</strong> in [&quot;earing&quot;] manual</td>
</tr>
</tbody>
</table>
**Class D enrichment: Searching for terms inside descriptive fields**

Vocabulary V1

- **http://scs.gr/sculpture**
  - skos:prefLabel "Sculpture"@en | "Γλυπτό"@el **keywords:** "statue"
- **http://scs.gr/figurine**
  - skos:prefLabel "Figurine"@en | "Σταυρί"@el **keywords:** "statuette"
- **http://scs.gr/Jewellery**
  - skos:prefLabel "Jewellery"@en | "Κοσμήμα"@el **keywords:** "earring"
- **http://scs.gr/vessel**
  - skos:prefLabel "Vessel"@en | "Σκύρος"@el **keywords:** "amphora", "oenochoe"

Example of dc:title value: "An [amphora](http://scs.gr/vase) from the Mycenaean period"

<table>
<thead>
<tr>
<th>dc:type values</th>
<th>Filters (terms found in dc:title values)</th>
<th>Entry from vocabulary V1</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D objects (240)</td>
<td><img src="http://scs.gr/vase" alt="amphora" />, <img src="http://scs.gr/vase" alt="vase" />, <img src="http://scs.gr/figurine" alt="earring" />, <img src="http://scs.gr/Jewellery" alt="jewellery" /></td>
<td><strong><a href="http://scs.gr/vase">http://scs.gr/vase</a></strong> if filter in [&quot;vase&quot;, &quot;amphora&quot;] but <strong>NOT</strong> in [&quot;statue&quot;]</td>
</tr>
<tr>
<td>art items (85)</td>
<td><img src="http://scs.gr/sculpture" alt="sculpture" />, <img src="http://scs.gr/figurine" alt="statue" />, <img src="http://scs.gr/figurine" alt="figurine" /></td>
<td><strong><a href="http://scs.gr/sculpture">http://scs.gr/sculpture</a></strong> if filter in [&quot;sculpture&quot;, &quot;statue&quot;]  <strong><a href="http://scs.gr/figurine">http://scs.gr/figurine</a></strong> if filter in [&quot;figurine&quot;]</td>
</tr>
</tbody>
</table>
The GUI of the enrichment tool
### Enriching the content of Searchculture.gr: The results

<table>
<thead>
<tr>
<th>Documentation Class</th>
<th># of repositories</th>
<th># of items</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A:</strong> sufficient existing dc:type values</td>
<td>20</td>
<td>30.764</td>
</tr>
<tr>
<td><strong>B:</strong> extremely specified dc:type values</td>
<td>5</td>
<td>11.102</td>
</tr>
<tr>
<td><strong>C:</strong> insufficient dc:type values – useful dc:subject</td>
<td>24</td>
<td>60.181</td>
</tr>
<tr>
<td><strong>D:</strong> insufficient dc:type – resorting to dc:title values</td>
<td>4</td>
<td>55.912</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>53</strong></td>
<td><strong>157.959</strong></td>
</tr>
</tbody>
</table>
dc:type values before and after the enrichment process for a specific repository (Documentation class C)

Homogenization based on dc:type and dc:subject (secondary field for filtering)
Improved searchability & search multilingualism

Enriching the content of Searchculture.gr: The results

Number of results on search keywords on types before and after enrichment
New ways for navigation: improved **discoverability** of content

Enriching the content of **Searchculture.gr**: The results
Thank you!

Haris Georgiadis PhD
Senior Software Engineer/Researcher
National Documentation Centre | EKT
e: hgeorgiadis@ekt.gr

Agathi Papanoti
Archeologist
National Documentation Centre | EKT
e: apapano@ekt.gr

Special Thanks to

Dimitra Pelekanou
Graphic Designer
National Documentation Centre | EKT
e: pelekanou@ekt.gr

National Documentation Centre
a: 48, Vas. Constantinou Av. GR-11635, Athens
e: ekt@ekt.gr | www.ekt.gr