Semantics.gr: A self-improving service to repositories and aggregators for massively enriching their content

Haris Georgiadis, Agathi Papanoti, Maria Paschou, Alexandra Roubani, Dimitra Pelekanou, Despoina Chardouveli, Evi Sachini

National Documentation Centre / National Hellenic Research Foundation
Athens, Greece
{hgeorgiadis, apapano, mpasxo, arouba, pelekanou, dxardo, esachin}@ekt.gr

Abstract: Most aggregators face challenges regarding searchability, discoverability and visual presentation of the content due to metadata heterogeneity. Fully automated enrichment in most cases is not sufficient. We developed semantics.gr, a tool for massive semantic enrichment and contextualization of content that combines a curator/expert metadata mapping suite with a self-improving automatic suggestion mechanism. We subsequently used this tool in order to enrich the content of searchculture.gr, the cultural heritage aggregator of National Documentation Centre (EKT). The results were substantial allowing us to improve drastically the quality of search and navigation of the aggregator's portal.

Keywords: aggregator, semantic enrichment, contextualization, linked data, automatic categorization, vocabularies, thesauri, cultural heritage.

1 Introduction

National Documentation Centre (EKT) has created an open cultural digital content infrastructure [1], a platform that promotes content quality in digital repositories by validating and certifying their content, aggregate their certified collections and publish them in a central aggregator/portal, *searchculture.gr*. During the first in production year of searchculture.gr, the main way for someone to discover the aggregated content was by using the portal's search engine which supported advanced key-based searching and facet filtering on search results against several metadata fields including dc:type [6]. However, the original documentation of the dc:type field was so much heterogeneous across the different repositories that limited the potential of further exploiting dc:type field for searching, filtering and navigating.

We can divide the heterogeneity of the values of a metadata field into two categories, representation-related and documentation-related. The former involves using different term variations for describing the same concepts, such as different languages, synonyms, mixing plural and singular numbers and using of different case styles ("all caps", "all lowercase"). The documentation-related heterogeneity involves different levels of quality or the use of different methodologies in the documentation of a metadata field. For dc:type this may range from using extremely general terms (for example "exhibit" instead of more precise types such as “sculpture”) to using very specialized terminology (for example “oenochoe” which is a specific type of “vase”). Heterogeneity has a negative impact on the following three important properties that constitute challenges to every aggregator: 1) Searchability & multilingual search 2) Discoverability & multilingual navigation and 3) Visual presentation.

Most aggregators use semantic enrichment techniques aiming in improving the first of these three properties. Europeana [3] and MorE [2] use automatic enrichment to terms of established vocabularies and thesauri. Complete straightforward automation adopts an ‘enrich-if-you-can’ strategy, horizontally, without taking into account special particularities of collections or hidden opportunities, which would inevitably involve some level of curation. Homogenization is not achieved and this is why aggregators usually avoid offering alternative ways of content exploration, such as hierarchical navigation or facet filtering on item types that would increase the discoverability and the visual presentation of the aggregated content.

Aiming in improving all the three afore mentioned properties, we initialized a pilot project aiming in enriching semantically and homogenizing the entire content of searchculture.gr, against a compact hierarchical bilingual SKOS [5] vocabulary of cultural item types developed by EKT which was linked to Getty AAT [4]. In order to achieve that we created *semantics.gr*, a platform for publishing vocabularies and thesauri that includes a tool for content enrichment and contextualization.

The platform was initially created as a platform where EKT and other institutions can create and publish vocabularies and thesauri. It can host vocabularies of virtually any RDF-based schema, such as SKOS [5]. An authorized user can model vocabulary schemata by creating owl classes (for example skos:Concept) that group parametric owl properties. The creation and configuration of owl classes and their properties are done via a user friendly web UI. When institutions create new vocabularies they first have to choose one of the registered owl classes. After that, they can start creating *vocabulary entries* using a parametric dynamic form that embeds all the properties of the respective owl class as form components whose functional and validation behavior reflects the respective owl property parameters. When an institution completes a vocabulary it can choose to publish it through semantics.gr open portal as Linked Data. This paper focuses on the enrichment tool of semantics.gr which is thoroughly described in Sec. 2.

2 The tool for semantic enrichment of Semantics.gr

The enrichment tool of semantics.gr is a tool for massively enriching metadata records (items) with references to vocabulary entries that are published in the platform. The target *repository* has to be first registered in the system. The tool does not execute the actual enrichment of the repository content. Instead, it provides with a GUI environment that embeds advanced automated functionalities that help institutions to easily define enrichment *mapping rules* for their repository metadata. The enrichment mapping rules are defined per distinct value of a predefined metadata field (for example dc:type) rather than per item. The tool accesses repository metadata via OAI-PMH harvesting in order to run count aggregations on specific metadata fields. Note that the tool only stores metadata field values and not the entire metadata. It can be used by repositories or aggregators to enrich their content. Particularly for an aggregator, it is recommended that the enrichment rules are set per collection in order to handle separately the documentation particularities of each provider/repository.

Table 1. 4 repositories with different documentation qualities.

Repository	Quality class	Documentation quality class description
R1	*A*	Good documentation of dc:type.
R2	*B*	Extremely specialized documentation of dc:type.
R3	*C*	Insufficient documentation on dc:type, useful dc:subject.
R4	*D*	Insufficient documentation on dc:type, useful dc:title.

After the phase of setting enrichment mapping rules is completed, the enrichment rules can be provided on request via a REST API in json format which can then be used by the digital repository or aggregator system to enrich their content or collection in a bulk and straightforward one-pass fashion. Note that we used the tool having the role of an aggregator, which means that each aggregated collection was enriched (according to the respective mapping rules served by semantics.gr in json format) while being re-ingested (re-indexed) in searchculture.gr (without affecting repositories).

In order to better illustrate how the enrichment tool works we are going to use an example. Suppose an aggregator-institution that has aggregated collections from 4 repositories and wishes to enrich/homogenize their item types (dc:type). Some documentation qualities of these repositories are summarized in Table 1. Each repository represents a different documentation quality class, *A*, *B*, *C* and *D*. The institution wishes to enrich these 4 collections based on their original dc:type values with references to a SKOS bilingual hierarchical vocabulary of types named *V* already published in semantics.gr. Vocabulary *V* contains the following 5 entries:

```
➔ http://scs.gr/sculpture
  skos:prefLabel  “Sculpture”@en | “Γλυπτό”@el
     ➔ http://scs.gr/figurine
       skos:prefLabel  “Figurine”@en | “Ειδώλιο”@el
➔ http://scs.gr/Jewellery
  skos:prefLabel  “Jewellery”@en | “Κόσμημα”@el
➔ http://scs.gr/vessel
  skos:prefLabel  “Vessel”@en | “Σκεύος”@el
     ➔ http://scs.gr/vase
       skos:prefLabel  “Vase”@en | “Αγγείο”@el
```

In their simplest form, enrichment rules are simple mappings from distinct values of a specific metadata field that we call *primary field* (in our example dc:type) to vocabulary entries. The enrichment tool supports automatic suggestion of mapping rules which by default is based on string similarity matching between metadata field values and indexed labels of vocabulary entries (in our example, skos:prefLabel values). The automatic mapping suggestion is very effective and efficient leveraging the indexing system that semantics.gr uses for its search engine, particularly, Apache Solr.

In our example, repository *R1*, a class *A* repository, falls into this average case. The curator first initializes a new mapping form dedicated to repository *R1*, sets metadata field dc:type as the primary field and chooses *V* as the target vocabulary. Then, the enrichment tool harvests metadata records from the repository and creates a list of 3 distinct dc:type values with their cardinalities (1st column of Table 2). Next, the curator triggers the auto-suggestion functionality which successfully maps all distinct dc:type values to the correct vocabulary entries. The curator then confirms these suggestions and the mapping phase is completed. The mapping rules are illustrated in Table 2. Label 'auto' indicates that the mapping rule was automatically created.

Table 2. Primary field: dc:type (Class A example)

dc:type value	**Entry from vocabulary *V1***	
sculpture art *(120 items)*	**http://scs.gr/sculpture**	**auto**
greek vases *(230 items)*	**http://scs.gr/vase**	**auto**
jewelleries *(135 items*)	**http://scs.gr/Jewellery**	**auto**

Repository *R2* represents class *B* that includes repositories that use very thorough documentation for the primary field. As shown in Table 3, all dc:type values of *R2* are narrower terms than those of vocabulary *V*. Since the auto-suggestion functionality wouldn't help in this case, a specialized curator, for instance an archaeologist, will manually assign the correct vocabulary terms to the narrower dc:type values (the GUI offers components such as drop down lists with auto-complete mechanism or modal windows for advanced vocabulary entry searching).

Table 3. Primary field: dc:type (Class B example)

dc:type value	**Entry from vocabulary *V1***	
amphora *(100 items)*	**http://scs.gr/ vase**	**manual**
oenochoe *(110 items)*	**http://scs.gr/ vase**	**manual**

We realized that the set of manual assignments of vocabulary entries to these metadata values which are usually *similar*, *narrower or instantiation terms* constitute valuable knowledge that we could leverage to improve effectiveness of auto-suggestion in future enrichments reducing manual assignments. We achieved this by simply storing those terms in a hidden field called *keywords* inside the respective vocabulary entries which is also indexed by the search engine. Therefore, the auto-suggestion mechanism which is based on the search engine will inheritably suggest vocabulary entries with matching keywords as well. The keywords that will be created after the enrichment of *R2* for entry http://scs.gr/vase are "amphora" and "oenochoe".

Repositories of class *C* are repositories that have insufficient documentation of the primary field (either for all or for some of the items) but have another metadata field that can contribute in the enrichment process. We call this metadata field *secondary field* and its values *filters*. For example, a metadata record may have a dc:type value "folklore object" but a dc:subject value "Jewel". To use secondary fields, the user must specify the metadata field that plays this role. This time, when the tool starts harvesting, it keeps for each distinct value of the primary field (for example for each dc:type value) a set of all values from the secondary field that was found (for example dc:subject values) which can be used as filters to route assignments to different vocabulary entries.

Table 4. Primary field: dc:type, Secondary field for filters: dc:subject (Class C example)

dc:type	**Filters (dc:subject)**	**Entry from vocabulary *V1***	
ceramic objects *(101 items)*	amphora (↗) , vase (↗), statuette (↗) …	http://scs.gr/vase if filter in [“vase”, “amphora”]	**auto** **auto**
		http://scs.gr/figurine if filter in [“statuette”]	**auto** **auto**
exhibits *(55 items)*	earing (↗), amphora (↗), …	http://scs.gr/Jewellery if filter in [“earing”]	**auto** **auto**
		http://scs.gr/vase if filter in [“amphora”] & NOT in [“earing”]	**auto** **manual**

Returning to our example, repository *R3* falls in this category. The secondary field is set to be dc:subject and the final mapping rules are shown in Table 4. For now, let’s focus on the first mapping rule for dc:type value “ceramic objects”: a metadata record with this dc:type value will be enriched with the reference http://scs.gr/vase only if it has one of the following dc:subject filters: “vase” or “amphora” or with the reference http://scs.gr/figurine if it has a dc:subject value “figurine”. The auto-suggest mechanism can easily suggest this rule as well as long as there are vocabulary matches (on labels or keywords) for these filters. The form has hyperlinks for every filter, denoted as “↗” in Table 4, that the user can use to search the repository/aggregator for items having the specific values on primary and secondary fields (we used deep linking for accessing searchculture.gr acceptance portal). The user can create complex logical expressions on the filters - including the logical NOT operator for setting exceptions - in order to create finer and more precise rules and avoid false positives. Returning on Table 4, items with dc:type value “exhibits” will be enriched with http://scs.gr/vase if they have a dc:subject “amphora” but they do not have a dc:subject “earing” (suppose that a digital file shows an earring whose shape is of an amphora).

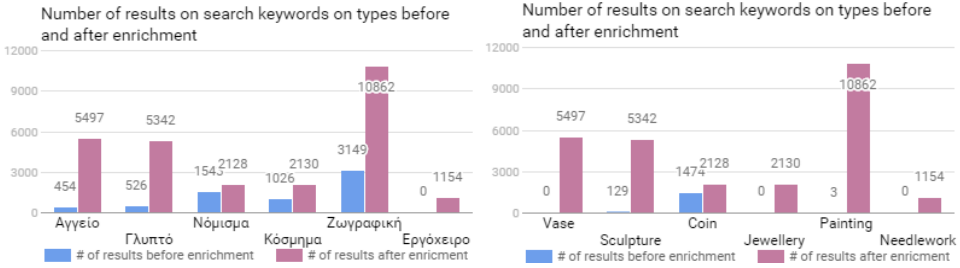
Finally, repositories of class *D* are repositories that have a very insufficient documentation of the primary field and no secondary fields to be used directly as filters. There are though highly selective fields (whose number of distinct values approaches the number of all items), such as dc:title or dc:description, that may contain words that reveal the appropriate vocabulary entry. For example a dc:title “An amphora from the Mycenaean period” implies that the item is a vase. Clearly using entire dc:title values as filters is not practical. However, what we can do is searching inside the dc:title values for specific words and then using the matching words (instead of their entire values) as filters. The problem is that we do not know a priori which terms to search. Our solution is searching for all terms that index vocabulary entries, being the labels (as skos:prefLabel) and their *keywords*. These words constitute a very useful large set of simple terms to be searched in such secondary field values. The rest of the mapping process is identical with the one described for class *C* repositories

3 Enriching the content of SearchCulture.gr – The results

More than 150K items of searchculture.gr were classified into a compact and balanced set of 130 types using semantics.gr. Table 5 illustrates the number of repositories per documentation class as introduced in Sec. 2 and the total of enriched items per class. Metadata records are enhanced with a separated field ‘EKT Type’ that holds the references to vocabulary terms (the original dc:type values remained un-touched). The enrichment directly increased the searchability of the content as illustrated in the experiment shown in Fig. 1(a) where we compared the number of search results returned by searchculture.gr for 6 search keys in Greek before and after the enrichment. The improvement was remarkable. We then repeated the same experiment but this time using the same 6 search keys in English, as shown in Fig. 1(b). Since the majority of the items were documented in Greek, the improvement was even more impressive. After the enrichment of the content, searchculture.gr was enhanced with new navigation functionalities that leverage the enrichment in improving discoverability including a uniform tag cloud and hierarchical navigation on EKT types. Our future plans focus on repeating the same process in order to enrich and homogenize the spatial and temporal fields as well as the subject headings.

Table 5. Repositories and number of items per documentation class

Documentation Class		# of repositories	# of items
A:	sufficient existing dc:type values	20	30764
B:	extremely specialized dc:type values	5	11102
C:	insufficient dc:type values – useful dc:subject	24	60181
D:	insufficient dc:type – resorting to dc:title values	4	55912
Total		**53**	**157959**



(a) (b)

Fig. 1. Improve in searchability of 6 keywords in Greek (a) and in English (b)

References

[1] Georgiadis, H., Banos, V., Stathopoulou, I.O., Stathopoulos, P., Houssos, N., Sachini, E.: Ensuring the quality and interoperability of open cultural digital content: System architecture and scalability. IISA: 178-183 (2014)

[2] Gavrilis, D. and Ioannides, M. and Theofanous, E.: Cultural Heritage Content Re-Use: An Aggregators's Point of View, ISPRS. II-5/W3: 83-87 (2015)

[3] Stiller, J., Petras, V., Gäde, M., Isaac, A.: Automatic Enrichments with Controlled Vocabularies in Europeana: Challenges and Consequences. EuroMed: 238-247 (2014)

[4] The Getty Art & Architecture Thesaurus, http://www.getty.edu/research/tools/vocabularies/aat/

[5] SKOS Simple Knowledge Organization System, https://www.w3.org/TR/skos-reference/

[6] Dublin Core Metadata Element Set, Version 1.1, httpW://dublincore.org/documents/dces/