Retrieving Structured Information from (Semi-)/(Un-)Structured Cultural Object Documentation

Stella Markantonatou, Athena Research Centre, Athens, Greece
Panagiotis Minos, Athena Research Centre, Athens, Greece
George Pavlidis, Athena Research Centre, University Campus at Kimmeria, Xanthi, Greece

ABSTRACT

In the course of developing facilities for integrating cultural heritage in the everyday education practice, highly structured information was retrieved from both the structured and the unstructured Europeana documentation contributed by the Greek cultural institutions (~480K entries); Modern Greek is the working language. Satisfactory results were obtained by using in-house developed medium sized Getty/AAT compatible vocabularies and simple heuristics. The paper reports on the development of controlled vocabularies and the retrieval of structured information from the unstructured Europeana documentation. Retrieval results show the importance of controlled vocabularies and thesauri as regards the exploitation of digital library content.

KEYWORDS

Controlled Vocabularies, Digital Libraries, Education, Europeana, Multimedia Databases, Serious Games, Thesauri

INTRODUCTION

This work is about facilities for integrating cultural heritage into the everyday teaching practice; more particularly, it is about the development of platforms for creating serious games that take advantage of the cultural information in the web.

Two terms will be used throughout this paper: ‘learning object’ and ‘cultural object’ and they both denote how the respective digital objects have been documented rather than referring to their content. Thus, in the framework of this research, a learning object is a digital object retrieved from a repository that uses the international standards LOM1/LRE2 to document the objects it contains. Similarly, a cultural object is a digital object retrieved from a repository that uses a CIDOC-CRM3 (or some other international standard for cultural object documentation) compatible documentation. The underlying idea is that a learning object has been formulated and documented to address educational needs and it may contain one or more cultural objects, or even other learning objects for this matter of fact, that have not been developed or documented with education needs in mind (Markantonatou, Minos, Tzortzi & Pavlidis, 2016).

In addition to retrieval requirements, education software imposes quality control restrictions especially because it is interactive and open to the younger ages. A database that contains all the objects used by an educational system facilitates quality control, as opposed to free web access. Such a database has to ensure communication with international repositories of both cultural and learning objects;
therefore it has to be compatible with cultural object documentation standards such as CIDOC-CRM (Doer, 2003; Crofts, Doerr, Gill, Stead & Stiff, 2009) and learning object documentation standards, such as LRE-MAP. A database designed in this way would accommodate both cultural objects and learning objects and would support the principled documentation and storage as well as the flexible search and retrieval of learning objects that contain cultural objects, for example serious games that contain pieces of music and 3D representations of statues and paintings. The authors of this paper have developed a database that satisfies these requirements and have populated it with cultural objects together with their standards compatible documentation that they retrieved from Europeana. All the contributions to Europeana by Greek cultural institutions (>480K objects) were retrieved and stored.

This database required information that was more structured than the information available in Europeana. The quality of the structure of the information in Europeana varies with the provider. Quite often Europeana provides access to unstructured textual data – unstructured because they contain units of information that (1) from a standardization point of view, should have been codified under different rubrics (2) occur in unpredictable format and order – see the example cases (1)-(3) in the next section. Since structured information was required, the unstructured Europeana documentation had to be subject to some semantic analysis in order to make sure that the right information was accommodated in the right database slot. This is not the typical task of retrieving Europeana objects relevant to some description – such as the tasks discussed in (Petras, Ferro, Gäde et al., 2012; Petras, Bogers, Toms et al., 2013). Instead, for each object in Europeana, a new object was developed in the database; the new object had standardized metadata that were retrieved from both the standardized and the unstructured metadata of Europeana. The method applied drew on a combination of controlled vocabularies/thesauri and simple heuristics; satisfactory results were obtained.

It must be noted that neither conceptually organized lexica of considerable size and coverage, such as the WordNet, nor controlled vocabularies/thesauri of cultural terms nor authority lists of some considerable size are openly available for Modern Greek; therefore the resources used were developed from scratch. The Greek vocabulary of names of objects and materials that was developed was mapped on the Getty/AAT; the mapping has assigned a structure to the Greek vocabulary and at the same time it constituted a first step towards a principled linking of Greek data with other Europeana data (de Boer, Isaac, Schreiber, van Ossenbruggen, Wielemaker & Stiller, 2011). AAT was chosen because it is a thesaurus for cultural documentation that covers a wide variety of objects, materials and techniques (Harping, 2010; Lanzi, 1998) and also, it has been used in multilingual data linking efforts within Europeana (de Boer et al., 2011).

In what follows, first the problem of retrieving structured information from the unstructured Europeana documentation is illustrated with examples. The retrieval method makes heavy use of controlled vocabularies and of mappings between vocabularies of different languages. As the different cultures have adopted somewhat different conceptualizations of the world, attempts to ‘link’ or ‘map’ vocabularies used to describe the cultural heritage of different societies run into correspondence problems; such is the case of mapping Greek controlled vocabularies extracted from cultural documentation texts with the AAT that more or less represents the American-English conceptualization of the world. Some indicative problems and the proposed solutions are discussed. Lastly, the method used to retrieve information from Europeana is presented and evaluated; the overall conclusion is that the proposed, relatively cheap yet resource-dependent, method of retrieval of structured information from the unstructured documentation of Europeana returns satisfactory results, at least in the case of the Greek data in Europeana.

**STATEMENT OF THE PROBLEM**

Consider the following Europeana entries:

Intertwining Culture With Education Through Gamified Storytelling: The Case of “Myth Trek”
www.igi-global.com/article/intertwining-culture-with-education-through-gamified-storytelling/202455?camid=4v1a

An Insight Into the Scientific Datasets in Archaeological Research Through the Pottery Provenance Case
www.igi-global.com/article/an-insight-into-the-scientific-datasets-in-archaeological-research-through-the-pottery-provenance-case/188814?camid=4v1a