Chapter 11

Using Semantics in XML Information Access: Application to the Portuguese Emigration Museum

Flavio Xavier Ferreira
University of Minho, Portugal

Pedro Rangel Henriques
University of Minho, Portugal

Alda Lopes Gancarski
Institut Telecom, France

ABSTRACT

This chapter presents an ongoing work in the context of the Portuguese Emigration Museum about information access in XML collections associated with semantic information. The museum asset is made up of documents of more than 8 kinds, ranging from passport records to photos/cards and building-drawings. In this chapter, the authors discuss the approach used to create the exhibition rooms of the virtual Web-based museum. Each room consists of the information contained in those single or interrelated resources. The information exhibited in each room is described by an ontology, written in OWL. The authors also discuss the approach used to take advantage of a combination of structural and semantic information to efficiently retrieve documents from the MEC collection. Both approaches can be automatised to allow a very systematic way to deal with the huge and rich museum assets.

INTRODUCTION

This chapter introduces a system that addresses the task of accessing information contained in XML documents of the Portuguese Emigration Museum (“Museu da Emigração e das Comunidades,” MEC, in www.museu-emigrantes.org).

This system has two approaches. One approach uses ontologies to specify different views over the information, and then offer navigation functionalities to explore them, using context sensitive behaviour. The second approach uses a query system that takes advantage of a combination of structural and semantic information contained in the museum archive.
Fafe, as many other Portuguese towns and villages, mainly at the north, has a huge cultural heritage characterising the social phenomena of emigration (especially to Brazil) along the nineteenth and first half of twentieth centuries.

In this context, Miguel Monteiro, supported by the staff of Fafe’s Town Hall (via Cultural Department), started some years ago collecting information from passport governmental records into a database and soon aroused the idea to gather all sorts of documentation into a repository and create a Web-based virtual museum that makes easily accessible this rich cultural heritage. This kind of museum is important and interesting for Emigrants and their descendants as well as for Historian researching in that area, and of course for the general public. The Museum was born in 2001; its material was inherited mainly from official documents or personal writings reporting on the departure, travel, and stay abroad, but there are also a large number of assets bearing witness to the less usual phenomena of emigrants’ return. Besides the documents, a large set of buildings (private or public, professional or philanthropic, and other non-physical evidences) left by the emigrants around the country can be also considered assets.

At the moment MEC virtual rooms are handmade, making them difficult to maintain and to add new information, and most important, lacking a systematic way for information acquisition, treatment and exhibition. In addition, some inconsistencies are evident from room to room but even inside the same room.

This fact gave the motivation for the project here reported, that aims to build a systematic approach for the acquisition, archiving, treatment, and exploration of the Museum’s documentary resources. In our perspective, each room is seen just as a specific view over a common information repository. The repository should be a digital archive (in database format or as a collection of XML files) of all the information resources referred above as museum’s assets. Each view (the knowledge enclosed in the respective room) can be specified by an ontology, as traditionally done by philosophers to organise the discourse over a certain closed-world. The extraction process can be automatized adopting a standard notation for the ontology description; moreover, the Web page that implements the user-interface in each room can also be automatically built.

Along the chapter, we will describe the research work exploring RDF/OWL to define the semantics associated to the XML information of the MEC. The W3C proposal for semantic descriptions is the Resource Description Framework (RDF) (Manola & Miller, 2004). RDF Schema (RDFS) (Brickley & Guha, 2004) is an extension to RDF that contains the basic builders for the description of ontologies. RDFS has some limitations as a standalone ontology language (does not allow for example the equivalence between two objects or two classes). To overcome these deficiencies, the W3C defined the Web Ontology Language (OWL), which is now the standard for ontologies definition.

Structured query languages are currently being used to retrieve information from both XML documents and semantic descriptions. To query XML documents, XPath (Berglund, et al., 2007) and XQuery (Boag, et al., 2007) were proposed by the W3C. In XQuery, the user can base his query not only on the textual contents of documents, but also from their structure. The result of a query is the set of structural elements that satisfy all the restrictions of the query.

To access semantic information three methods can be used: a simple expression in natural language (Finin, et al., 2005); through a navigation structure (Fluit, et al., 2005), or using a dedicated query language like SPARQL (Corby, et al., 2004). The SPARQL query language was defined by the W3C to find information in RDF descriptions (Prud’hommeaux & Seaborn, 2007).

Accessing information over resources is generally made directly on the documents or indirectly on semantic descriptions to find the associated