Cantabria Cultural Heritage Semantic Portal

Francisca Hernandez Carrascal[1], Luis Rodrigo[2], Jesus Contreras[2], Francesco Carbone[2]

[1]Fundación Marcelino Botín, http://www.fundacionmbotin.org
fhcarrascal@gmail.com
[2]Intelligent Software Components, www.isoco.com
{lrodrigo, jcontreras, fcarbone}@isoco.com

Abstract. This document describes an ongoing commercial project that aims to model the cultural heritage domain in an ontology, containing data from eleven types of heritages, from bibliographic items, to industrial patrimony, or to prehistoric excavations for a Spanish region, Cantabria. Besides, the project will develop a series of applications that get the most out of the ontology, beginning by a semantic portal, which is the application presented in this paper. The portal will provide easy and intuitive access to a huge volume of information that comes from heterogeneous sources selected by domain experts. People interested in Cantabria will be able to find with as few clicks as possible, not only the information they are interested in, but also some contextual information that complements the one the user is watching, suggesting new navigation directions, and uncovering hidden relations. The ontology has been designed based on the international standards for cultural models, such as CRM, FRBR or Dublin Core.

1. Introduction

Cultural heritage information, especially artistic or historical domains, includes a dense net of interrelations. Understanding the circumstances of certain artistic work creation or its influences improves user experience. From the historical or political period, to personal influences or relations, to the material or technique used is relevant information for users. This information is usually elaborated manually in the form of documents accompanying works in exhibitions and publications.

The existence of international standards for cultural information representation and need for advanced and precise information querying and searching suggests the semantics can help. One of the possible evolution paths of the Semantic Web until reaching the mainstream market is similar to how the WWW has exploited. While the traditional WWW has start from academic HTML islands and then was extended to its current dimensions. Cultural heritage domain offers very good conditions to constitute semantic islands that can be interconnected since it is a relatively stable domain, with no frequent changes, very strong interrelations and a set of accepted standard models to encode its information.

The aim of this project, called "Ontology for Cantabria's Heritage", is to use Semantic Web technologies for intelligent integration of information about Cantabria region and heritage. It includes all kind of cultural heritages, from bibliographic items, to industrial patrimony, or to prehistoric excavations for this Spanish region. To achieve this objective, the project will design, construct, feed and exploit a cultural heritage ontology. It aims also to develop a methodology for the ontology population and exploitation for other regions of Spain, or even outside Spain.

The project is funded and led by the Marcelino Botín Foundation[1] and is performed in collaboration the University of Cantabria[2], domain experts and iSOCO[3] , a semantic technology company. The Marcelino Botín Foundation aims at transferring research and scientific discoveries to the whole society by funding several cutting edge initiatives. These initiatives are then transferred to public or private entities, as would be the case of this project.

The portal has been constructed with the idea of reaching as many users as possible. However, the expected average profile corresponds to domain experts, namely librarians, archivists, curators, or people with equivalent backgrounds.

2. Ontology

The ontology is the core element of the project. It gathers data from different sources into a model that is the combination of several standard models, namely:

- The recently approved standard for cultural information management called Conceptual Reference Model (CRM)[4]. Developed by the International Committee for Documentation of the International Council of Museums (ICOM-CIDOC), allows for representing the relevant information and its relations in an explicit and complete way. This standard became an ISO[5] norm in 2006: "Information and documentation - a reference ontology for the interchange of cultural heritage information ISO 21127", is used for cultural information repository creation.
- Dublin Core[6] : Meta data for online resource descriptions. ISO Standard 15836-2003
- Functional Requirements for Bibliographic Records (FRBR)[7] : is a conceptual entity-relationship model developed by the International Federation of Library Associations and Institutions (IFLA) that relates user tasks of retrieval and access in online library catalogues and bibliographic databases from a user's perspective.

[1] http://www.fundacionmbotin.org/
[2] http://www.unican.es
[3] http://www.isoco.com
[4] http://cidoc.ics.forth.gr/
[5] http://www.iso.org/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=34424
[6] http://dublincore.org/
[7] http://www.ifla.org/VII/s13/wgfrbr/index.htm

For those areas of the model that did not have enough development, a joint team of domain experts and knowledge engineers have devoted the initial steps of the project to reach an ontology schema as complete as possible for all the kind of data relevant for the potential users. The whole ontology has been split into four levels in order to facilitate its design and maintenance.

- Control Level: represents the meta-model information as well as meta-data information about data acquisition sources, annotation data, comments, etc.
- General Level: Upper level ontology for representing concepts such as: Agents, Objects, Processes, etc.
- Basic Type Level: includes concepts for representing values of some attributes. It defines the types or ranges for: Temporal Periods, Distances, Geo-positions, Dimensions, Familiar relations, Etc.
- Domain Level: Each kind of heritage will derive from these upper levels. E.g.: natural, industrial, bibliographic, etc.

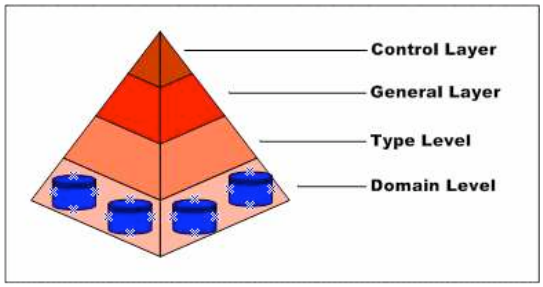

Fig. 1. Level approach to Cultural Ontology

Regarding the contents, the team of domain experts selects the relevant information sources from a variety of on-line sources, such as institutional archives, web pages, relational databases, etc… If a data gap is detected in any area, there are also some Excel templates which cover the most relevant instances of the ontology, and can be manually filled by the domain experts to complement existing knowledge or contribute some new one. Automatic acquisition tools have been built to process each kind of input, this is, excel sheets, structured web pages, bibliographic records using MARC21 XML, Dublin Core, EAD and EAG formats, generic XML documents and relational Databases, so that the whole process of populating the ontology with information can be done completely automatic and, thanks to the Dublin Core properties used, is also fully traceable and repeatable. At this moment, the ontology is populated from:

- 211 EAD files.
- 184 EAG files.
- 9478 MARC21 records.
- 912 Dublin Core records.
- 2 relational databases.
- 4 official web pages from the Cantabria government.
- 45 excel sheets.

At this moment, when conflicts are detected on information coming from different sources, Dublin core properties are used to locate the source of the different data, and one of them is manually corrected to ease the process of information integration. Nevertheless, an application for semi-automatic data integration has already been designed, and will be incorporated in the near future.

As more and more sources are discovered, the ontology is in a continuous process of growth. At the moment of writing this article, the ontology has 153 classes, 410 attributes and more than 110.000 instances.

To load the ontology in the portal, and handle all the accesses and queries, a Sesame[8] repository with SwiftOWLIm[9] backend is used.

3. Access

Objectives

The portal has been designed to offer access to all the information that has been (and will be) gathered along the project development. Due to the nature of the information that will be collected, a very high volume of very specialized data will be available. If the portal is not designed to ease as much as possible the location of a concrete fact, or number, it will become completely unusable. However, at the same time, considering that the portal will be available to the general public, it should not underestimate the importance of providing interesting contents for them. With this objective also in mind, the portal was also designed to supply a pleasant navigational experience for someone who is not seeking for an answer, but just taking a look and surfing without a specific objective in mind.

In order to fulfill the first requirement, easy access to every piece of information in the portal, the system gives the user the possibility of using different criteria to search for information. Besides a standard search box, and an advanced search facility, instances in the ontology are timely and geographically referenced, an information that can be exploited using the map and the timeline in the portal. There is also a tagcloud providing direct access to the most popular elements of the ontology. And finally, the menus provide direct access to information in two versions: a fixed menu with lists of most important elements in the ontology, and a series of side menus that provide contextual suggestions depending on what the user is watching every time. All these facilities are better described in the following subsections.

The second main objective, being an attractive site to navigate through for the general public is addressed by including interactive elements that catch the users attention,

[8] http://www.openrdf.org/
[9] http://www.ontotext.com/owlim/

such as the tagcloud, the map, or the timeline, and by making available a good amount of contextual information, that stimulates the navigation around the portal.

Interactive Map

Since the ontology contains geo-positions for resources like *Site* or *Place*, and most of the instances have some relation to those concepts, they can be represented over a map picture. The Interactive Map application allows for placing different domain layers, such as singular buildings, churches, caves, etc. on the map in an interactive way and connected to the ontology.

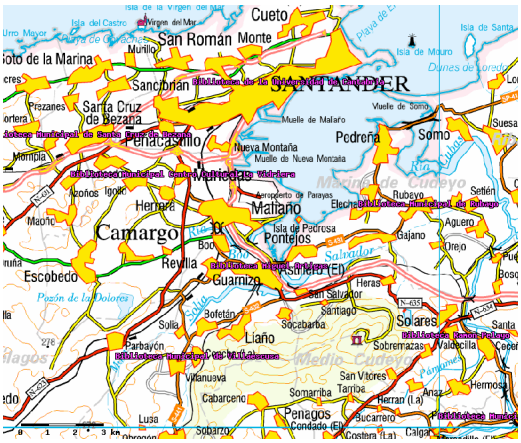

Fig. 2. Map for Cantabria Region

The map server has been developed with technology from iSOCO, and in a near future the map will be substituted by a satellite photographic version.

The map is tightly related to the search boxes in the portal, being possible to show the results of a search as points in the map, but also to restrict the result set of a search to those located in a region of Cantabria previously selected in the map.

Timeline

To visually represent events on time, we have chosen to adapt a component developed in the SIMILE project[10], Timeline. This allows representing all the events present in the ontology in an easy and attractive fashion.

The timeline allows the user to have a clear view of how events are distributed on time, and accessing events knowing when they occurred. In a future, the user will

[10] http://simile.mit.edu

have the possibility of restricting the searches to those events that happened in a certain period, by selecting it in the timeline.

Navigation components

While the user navigates in the portal, he goes through a collection of charts with information focused on a specific topic. Meanwhile, the system is permanently showing a side bar with links that take to related topics, that change as the chart the user is watching changes, therefore exploiting the context to give the user new navigation directions.

Depending on what kind of chart the user is reading, i.e., one of a person, an institution, an event, etc… the kind of links suggested are also different. When there are too many candidates to be shown, the ones that have more relations that connect them in the ontology are the ones selected. The user can always ask to be given the complete list of, for example, the people related to an institution.

Normal/Advanced search

As in many other web sites that contain a good amount of information, the portal provides the possibility of searching for specific items. With this objective, we index not the ontology, but the charts that are presented to the user, and provide two mechanisms for searching.

First of all, a standard search box, but with the capability of managing morphological variations, orthographical accents and spelling corrections is always present is the top right position.

Secondly, if the user knows well the domain, and what she is looking for, she can choose to use the advanced search, which provides a set of forms, specially designed to look for specific kind of entities (people, events, places, etc…) based on their main relations with the rest of instances.

Alfresco repository

Apart from the knowledge in the ontology, the portal also contains a collection of documents, stored in al Alfresco server. While navigating, the user is always offered the set of documents in which the instance she is watching is mentioned.

Nevertheless, not all the information about documents is stored in the Alfresco server. To take advantage of the frbr model integrated in the project ontology, useful for bibliographic information, this kind of information about the documents stored in the documental repository will be stored in the ontology, allowing for more complex criteria when searching.

While at present the repository is used only for textual documents, the infrastructure separation of data and metadata allows the usage of any multimedia file, something that will be explored in further phases of the project.

4. Visualization

As aforementioned, the portal should provide both high usability and visual attractiveness in order to achieve a pleasant navigational experience. With this objective in mind, the design of the portal has been created by design professionals. Five different designs where created, and a team including computer scientists, librarians and domain experts selected the one that is in use now.

Instance Charts

The main element in the screen while navigating is the chart in charge of presenting the user the information she has requested (see Fig. 3).

Fig. 3. Example of chart in the portal.

The generation of these charts follows an innovative process, which facilitates its generation and adaptation to the user needs. The description of the process to generate these charts for each of the main topics in the domain is stored in a so called "visualization ontology". This splits completely the data representation and data visualization, allowing the choice of the most adequate mechanisms for each of them. For example, the chart in Fig. 3 was generated using the standard chart defined in the visualization ontology for artistic items, which compiles information from several instances in the domain ontology, and arranges them in a suitable manner.

This visualization ontology is easily generated from a domain ontology, and only needs the adequate adjustment to be useful and satisfactory for the users.

Regarding the technical details, the publication ontology contains "publication instances", which are templates to publish information about the most important elements in the domain, and are implemented with a combination of SeRql queries and apache velocity templates, that automatically collect and organize the required information from the domain ontology and present it in a visually friendly manner.

User profiles

The publication ontology paradigm also allows the creation of different publication entities for the same entity in the domain. The selection of the most adequate one for each situation can be done based on configurable criteria. In this case, three different publication instances have been generated for each domain item, which contain appropriate information for three different kind of users, such as, a basic user, a domain expert and someone interested in the raw ontology contents (usually developers or very advanced users).

The user does not need to identify himself to choose the adequate profile for him. Instead, the system starts with the most basic charts, and includes a link to ask for "More Information". If pressed, the portal changes to the advanced user chart generation, and shows links to change the information detail to "Basic" and "Raw ontology contents". Once the user has selected the desired information detail level, the portal keeps using it until a new profile is selected, allowing easy and intuitive configuration.

5. Conclusions

The paper has presented the Cultural Heritage portal for one of the regions in Spain, Cantabria. The application is centered on an ontology that gathers information from a bunch of different reliable sources, aggregating it to exploit the undercovered relations that emerge from the aggregation by providing contextual suggestions to the user interests.

The portal is addressed to both casual users that want to spend some time navigating and domain experts that seek for a very concrete piece of information. Different criteria can be used to access every item in the portal.

While still not in use, the design allows the integration of multimedia files to complement the information in the ontology and improve the user experience.

The portal is included in a commercial project that aims to build seven additional applications that exploit the knowledge captured in the ontology, and which has already captured the attention of different regions in Spain.