# BUILDING A CULTURAL HERITAGE ONTOLOGY FOR CANTABRIA

Francisca Hernández, Luis Rodrigo, Jesús Contreras, Francesco Carbone
Fundación Marcelino Botín, http://www.fundacionmbotin.org ; Intelligent Software Components, www.isoco.com
E-Mail: hdz@fundacionmbotin.org; {lrodrigo, jcontreras, fcarbone}@isoco.com
URL: www.fundacionmbotin.org; www.isoco.com

**Abstract**

**The paper describes, probably the largest initiative for the construction of an ontology that compiles the knowledge around the cultural heritage related to a specific region — Cantabria in Spain. As a basis for the schema, we chose the RDF representation of CIDOC CRM (v. 4.2), and integrated complementary models, such as FRBRoo, Dublin Core, or project specific extensions to reach a generic schema to be able to represent information concerning different domains while keeping a sufficient level of detail to be representative and useful. Once a stable version of the schema was achieved, reliable information sources were identified, each of them having a different format to represent the data, sometimes standardized (MARC21, or EAD), sometimes proprietary. Specific tools were designed and built to import all the information in the different formats into the ontology.**

**Unstructured high quality information is prepared by subject matter experts and formatted into data sheets for their automated importation process. The objective was to fully rebuild or to update the ontology contents at any moment, whenever updated resources become available. The ontology, being itself one of the objectives of the project, is also the foundation for a series of applications that are built on top of it.**

## INTRODUCTION

The objective of this paper is to expose the project Cantabria's Cultural Heritage Ontology [CCHO]. In 1998, the IST project *Contemporary Virtual Archives in XML*

(COVAX)[1] intended to prove the advantages of XML as a common format for interchanging data across the boundaries of archives, libraries and museums. This project showed some fruitful achievements; the feasibility of a distributed information systems built on databases from different ALM, making use of standards such as MARC 21, EAD, AMICO and Dublin Core as metadata schemas and the Z39.50 searching protocol, translated to XML via XER[2] (XML Encoding Rules)[3].

At that point, there was an opportunity to enlarge the spectrum to all kinds of cultural heritage and to build a system to locate and collect it, and make it useful and browsable. The objective was a system gathering all the information in one site, avoiding the users to spend effort discovering sources in directories, web sites, catalogues, finding aids, inventories, etc.

Nowadays we are talking not only about references and primary sources but also about the knowledge incorporated to different systems as a way to overcome current limits that the information structure applies to search and retrieval. As an example, none of the usual databases, at least those build upon MARC 21, can solve queries like 'authors born in Cantabria', or 'Cantabrian printers in the XIX century'.

This was the main objective of the Cantabria's Cultural Heritage Ontology, using semantic web technologies to make an explicit and intelligent integration of all the information about the cultural heritage of Cantabria.

The Marcelino Botín Foundation[4] is the institution that promotes, finances and sustains the project, in its strategy of contributing to the protection and dissemination of the Cultural Heritage of Cantabria.

---

1 Hernández, Francisca [et al.] (2003) "XML for Libraries, Archives, and Museums: The Project Covax". *Applied Artificial Intelligence*, 17 (8-9), pp.797-816.
This project reflected the working lines expressed in Hernández, Francisca; Agenjo, Xavier (1999). "¿Tres vías al conocimiento?: La información de archivos, bibliotecas y museos y el derecho de los ciudadanos a los documentos primarios". *Boletín de la ANABAD*, 49 (n. 3-4), pp. 559-568.

2 http://xml.coverpages.org/xer.html

3 It seems that these aims could be reached by the application of two protocols, SRU and OAI-PMH

## THE PROJECT

The Cantabria's Cultural Heritage is a very large and heterogeneous domain, composed of different fields like, industrial, ethnographic, scientific, natural, artistic, bibliographic and documentary heritage. It is represented in many kinds of manifestations (man made objects, biological objects) and carried by a big variety of information carriers (primary documents, digital copies, web sites, etc.). It has been studied, analysed, protected by different actors throughout the history (people and institutions), each of them documenting their actions and knowledge in all kinds of records, monographs, articles, legal texts, collections, etc. This knowledge comes from different sources (encyclopaedias, reference books, finding aids, government publications, databases, web sites, etc.) that can be accessed only in a fragmented way, without any or with a weak interconnection. Besides that they are in different stages of formalisation, standardisation or, even, digitisation. All of the above made very difficult to access this information.

Based on the Semantic Web technologies, the project has built an ontology referred to the scientific and technical fields related with the cultural heritage (history, geography, art, literature, library and information management, museums and archives, etc.), which have been generated by people and institutions in the development of their activities (among them universities, non profit institutions, foundations, government bodies, etc.).

This way, the Cantabria's Cultural Heritage Ontology considers the different kinds of heritage and their components as they are defined in the Cantabria and Spanish current legislation (or international dispositions as in UNESCO World Heritage Sites). It also considers the publications that have been edited until now, even before

[4] http://www.fundacionmbotin.org/

the first Spanish legislation for the protection of the cultural heritage was established, to analyse[5], study, assess, manage, restore, protect and disseminate this heritage.

The project also intends to reach the power and versatility of the semantic web in retrieving information. The *Semantic Web Education and Outreach Interest Group: Case Studies and Use Cases*[6], which includes the Cantabria's Cultural Heritage Ontology[7], shows the variety of procedures, methodologies and applications in the semantic web. For the project we have followed the techniques that bibliography offers for the categorization and organisation of documents that exist about a particular field. This way, in the first stage we found the information published about Cantabria in reference sources (paper, digitised or databases) like bio-bibliographies, general bibliographies, special bibliographies, biographies, dictionaries, encyclopaedias, directories, catalogues, bibliographic databases, authority records, etc. That is equivalent to say, that is meaningful, what has been studied about Cantabria. This is the reason why the bibliography *Historia de Cantabria : un siglo de historiografía y bibliografía (1900-1994)*[8] has been chosen as the foundation of the ontology content. In the second stage we have considered the descriptions of the items (bibliographic materials, archive records, artefacts, buildings, pieces etc.) that shape the Cantabria's Heritage. In this stage we have used libraries and museums catalogues, finding aids and guides etc., as fundamental sources. It must be said that some of these sources are in databases, with proprietary structures, and many of them don't have even a digital version. Since the aim of the ontology is the integration of information it is very important to locate standard sources (which means sources structured following standards as MARC 21, EAD, etc.) that could be converted to the structure of the ontology. Throughout the project it has been necessary to convert some of these sources to standard formats. It is important to say that, at the present stage of the ontology technology, the project has not considered the use of its RDF

[5] The discovery of the paintings in the Altamira Cave in 1879, and declared Unesco Worl Heritage Site in 1985), motivated an international controversy about their authenticity reflected in speciallised reviews and conference papers.
[6] http://www.w3.org/2001/sw/sweo/public/UseCases
[7] http://www.w3.org/2001/sw/sweo/public/UseCases/FoundationBotin
[8] Suárez Cortina, Manuel, ed. **(1995)** *Historia de Cantabria : un siglo de historiografía y bibliografía (1900-1994)*. Santander : Fundación Marcelino Botín.

structure as a way to convert legacy data. The idea is first to convert legacy data to standard (and in use) formats or schemas and then to integrate into the ontology by specific conversion rules. The same situation can be found in the third stage, the knowledge about heritage management. In this section, we have included legal texts, government bodies at every level (national, regional, and municipal); and their programmes, projects, etc. Finally, we can locate general knowledge (and information structures) composed of geo-referenced data[9], chronologies, official statistics, etc.

## Standards

The project has developed a common structure for the data mentioned, having in mind, as a principle, the importance of standardisation in itself for the future evolution of the project and for its extension. Because of this we chosen the models defined by CIDOC CRM and FRBRoo rather than create a new structure. In first position we have followed the standard *ISO 21127 Information and documentation-A reference ontology for the interchange of cultural heritage information*, that is the Conceptual Reference Model (CRM)[10] developed by the CIDOC (ICOM). This standard provides a formal structure to define concepts and relationships used in cultural heritage. It is a common and extensible semantic frame to which any archive, library or museum structure of information can be converted. It is also a guide to model the entities that shape the information about cultural heritage. This standard also assures the interoperability[11] of the Cantabria's Cultural Heritage Ontology with other ontologies or systems that could make use of CIDOC CRM and as a way to facilitate the technological transference. In other words, CIDOC CRM allows for the transformation of different information sources to a common and interoperable ontology.

---

[9] This kind of data, very poor in general for Cantabria, has been obtained from geographical servers and must be necessary to create them in many cases.
[10] http://cidoc.ics.forth.gr
[11] Europeana/EDLnet project has as one of its objectives to investigate the use of CIDOC-CRM and FRBR as way to promote the interoperability between systems.
http://europeana.eu/public_documents/4_FrameworkIssues_WP2_Stefan.ppt

Undoubtedly, it has been necessary to add new classes and relationships, as ISO 21127 foresee, more specially to detail professional relationships between people or between people and institutions, occupations and professions, or family relationships, or to add new relationships to 'actors' to expand the possible actions beyond the museum activity. The conceptual model of CRM had, since the beginning, a very strong influence from the museums field, but other aspects as bibliographic needed to be expanded. One of the used models, with a big influence in the new international cataloguing code known as RDA[12], is the proposed by the Functional Requirements for Bibliographic Records (FRBR)[13]. Since 2003, the Working Group on FRBR/CRM Dialogue has the aim of designing an object oriented extension for the CRM. The latest version is FRBRoo 0.9[14] published in January 2008. This extension follows the ISO 21127 standard as FRBRoo classes (i.e., F1.Work, F2.Expression, F3.Manifestation_Product_Type and F5.Item) are subclasses of classes already defined in CRM or are equivalent classes. The Cantabria's Cultural Heritage Ontology has incorporated FRBRoo, covering the modelling of bibliographic information, that is one of the most important sources for the ontology, and for every ontology in cultural heritage.

In the conception of CRM there was the possibility to convert data from archives, libraries and museums and there are some proposals to transform Dublin Core, Encoded Archival Description, MDA Spectrum, etc. Dublin Core represents a special case in the Ontology, it is not only used as the structure to convert legacy data (is the mandatory schema used by OAI-PMH repositories and for metadata harvesting) but also the structure to manage the ontology as it describes each of the classes, properties and instances as information resources.

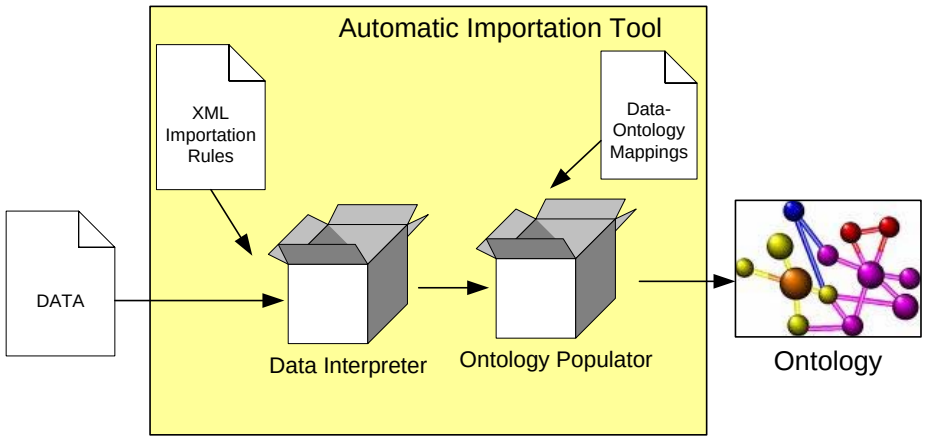
## POPULATING THE ONTOLOGY

[12] Joint Steering Committee for Development of RDA. Resource Description and Access. http://www.collectionscanada.gc.ca/jsc/rda.html
[13] http://www.ifla.org/VII/s13/frbr/frbr.pdf
[14] http://www.ifla.org/VII/s13/wgfrbr/FRBR-CRMdialogue_wg.htm

Once the ontology schema has been defined, the main challenge has been the automatic importation of data from heterogeneous sources. The first task consisted in selecting the relevant information sources from a variety of on-line sources, such as institutional archives, web pages, relational databases, etc. If data are considered not accurate or incomplete, some Excel templates which cover the most relevant instances of the ontology were used. Domain experts manually filled these templates in order to complement existing knowledge or contribute some new one

Handling different data sources meant handling different input formats: excel sheets, structured web pages, bibliographic records using MARC21 XML, Dublin Core, EAD and EAG[15] formats, generic XML documents and relational Databases were considered. For each of them we implemented automatic acquisition tools 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.

Each tool for automatic importation is divided in two main modules: data interpretation/extraction and data insertion in the ontology.




[15] EAG (Encoded Archival Guide) is a DTD for a guide to the Spanish Archives: Censo-Guía de los archivos españoles.
http://aer.mcu.es/sgae/jsp/censo_guia/seg_nivel/presentacion_proyecto.html

Information about the semantics of each field of the input data is defined in a XML file thanks to the help of domain experts: for data in MARC 21 format several XML rules files describe in a more friendly way the meaning of each tag. In this way we provide the Data Interpreter the information about how extract data from the input The Ontology Populator contains the knowledge for populating the ontology using the data received from the previous module.

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.

Information sources sometimes provide data about the same subject but describing different aspects. For instance a bibliographic record offers basic information about the author of a book, while a relational database contains biographical information about the same person. Specific tools have been developed in order to match and merge this data.
It has also to be considered that information coming from different sources could have conflicts: 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.

## CANTABRIA'S CULTURAL HERITAGE PORTAL

All the contents automatically imported from the different sources are integrated into the project ontology. Up to now the ontology has over 160 classes, more than 450 properties and +110.000 instances, before applying a set of deduction rules that infer some more knowledge from the explicit contents that are in the ontology.

The ontology is managed through a Sesame (http://www.openrdf.org) repository, configured to use SwiftOwlIm (http://www.ontotext.com/owlim) as a high-performance storage and inference layer.

The portal is the main application of the project, and is conceived to offer easy access to all the contents gathered in the ontology. It should be aimed both to specialised users looking for a concrete piece of information, and to general users, that just want to spend a while navigating without a particular objective.

In order to fulfil 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, 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.

Although ontologies are a step forward in the sense that their content is machine readable, most existing semantic portals have decreased their human readability. This is due to the fact that current approaches try to visualise the content of the ontology as it is, meaning that navigation strictly has to follow the ontological structure, and if not, the deviation from the ontological structure is hard coded in the user interface (e.g. JSPs or ASPs). In our approach, we introduce the notion of visualisation ontology to decouple the ontology structure from its visualisation (including

navigation). The visualisation ontology allows us to separate what we see from how we see it. Moreover, it allows choosing which content should be visualised, or not, from all the information contained in the ontology. We call this a visualisation ontology. The visualisation instances define how information should be collected from different classes and attributes in the original domain ontology. Once the information is collected, it can be published in any desired format, thanks to the use of velocity templates. HTML and WML have been used so far, but any further format would be easily included.

Thanks to the usage of the visualisation ontology we exploit the concept of serendipity, showing not only the requested information, but also information that is semantically related to it.

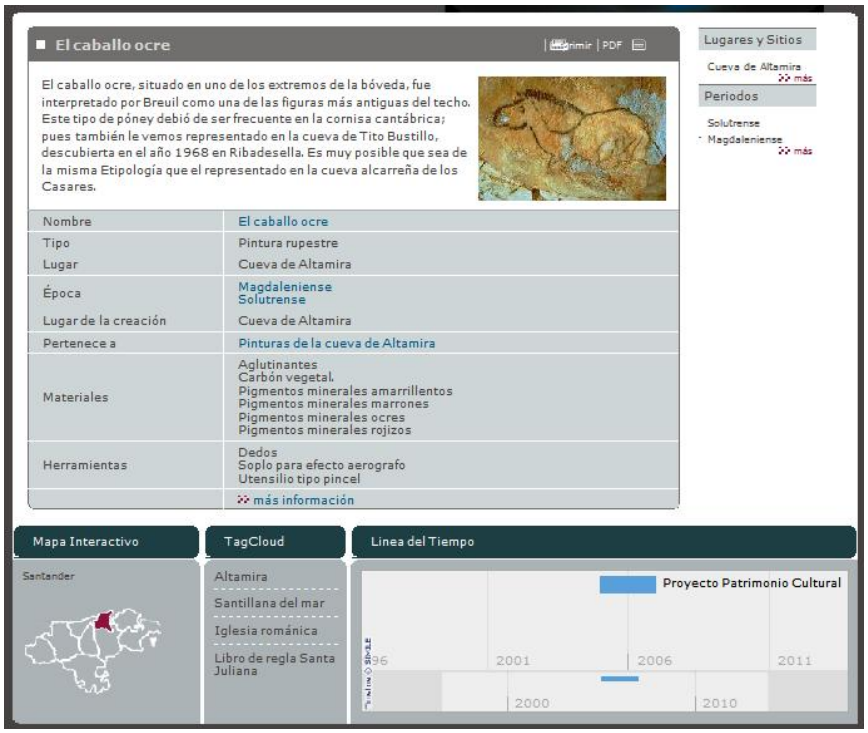



Figure 1: Ontology publication

As seen in Figure 2, the user is presented the information he has requested, this time "*El caballo ocre*", one of the most famous prehistoric paintings in Spain, which is displayed in the main section of the screen, surrounded by information that she has

not explicitly asked for, such as related places, or historical periods (in the right hand column), or temporal and geographical data, in the bottom. This set of contextual information is dynamically generated depending on the central chart the user is visualising.

The portal also includes three alternative approaches to access the information apart from the standard static menu and search box that is contained in the portal.

## Timeline

To visually represent events on time, we have chosen to adapt a component developed in the SIMILE project[16], 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 have the possibility of restricting the searches to those events that happened in a certain period, by selecting it in the timeline.

## Tagcloud

A cloud with the most frequently visited charts has been included in the portal , what aids the user selecting the pieces of information that most repeatedly catches his attention. This also constitutes a rapid entry point to the system for those users that do not have a specific navigation target in mind, but just want to have a look at the contents.

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

## 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.

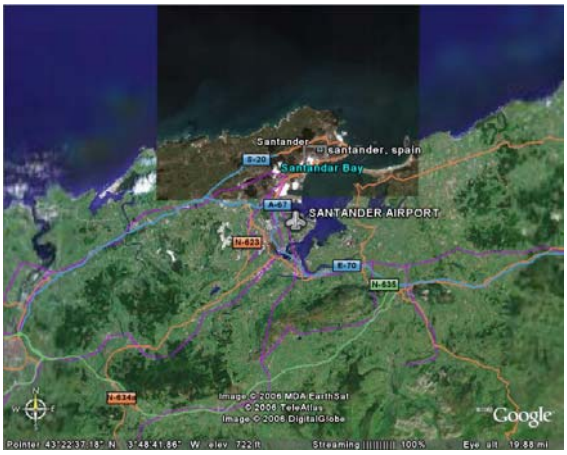



Figure 2: Satellite map for Cantabria region

# ADVANCED SEMANTIC APPLICATIONS

The ontology allows not only for Web publishing of relevant and related information, but also for advanced applications building. Thanks to the underlying structure, the possibility of querying, inference or intelligent information integration a set of compelling application is being deployed on top of the cultural heritage ontology.

## Semantic Search Engine

The semantic search engine allows for searching of information from the ontology using precise queries. The search engine retrieves data instead of documents as a traditional search engine would do. For instance, the query:

*"All books published between 1907 and 1917, and written by authors from Santander"*

will return a list of instances representing books rather than a list of links to documents containing keywords.

## Tourist Application

Some tourism operators offer an added value for their travel products including cultural experiences and information on the visited place. The cultural ontology application is ready to offer added value information, via web services, for online tourist applications:

- Cultural routes built following a given topic (famous author, work, period, etc.)
- Additional information for GPS points of interest including history, related information, etc.


Figure 3: Routing Application showing "Romanic Route"

## Semantic Wikipedia

One of the most important issues when using ontology in large industrial applications in the ontology maintenance and evolution. Wikipedia software has proven to be an efficient and usable method for collaborative knowledge management. We have included a semantic extension to the traditional software. The semantic Wikipedia allows for new semantic data acquisition and modification by expert users, not necessarily with semantic web technology background. With an user management extension, authorized users can modify ontology content, suggesting modifications on specific data. These modification are then included into the ontology by a supervised workflow process.


Figure 4: Wikipedia front page