Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 147 (2014) 345 – 348

Procedia
Social and Behavioral Sciences

ICININFO

XTM-based Metadata Exploration and Retrieval: The History of Spanish Civil Engineering Project

Ricardo Eito-Brun*

Universidad Carlos III de Madrid c/Madrid 124, Getafe (Madrid) , 28030, Spain.

Abstract

This paper describes the History of Spanish Civil Engineering Project. The purpose of this project is to create a web repository where users can access the description of archival materials related to Spanish civil engineers. Personal fonds are being added to this repository based on the initial work completed on the Carlos Fernández Casado and E. Torroja pesonal fonds. From a technical perspective, the project makes use of the RDF and XTM standards to automatically generate topic maps that can be used to browse the different fonds and collections.
The system architecture keeps the navigational aids (based on XTM) independent of the descriptive metadata used as a source. The XTM topic map works as an intermediate layer to ensure uniform access to information resources spread across different servers. The paper discusses the differences between RDF and XTM, and how these standards are being applied in the generation of the access layer.

Selection and peer-review under responsibility of the 3rd International Conference on Integrated Information.

Keywords: XTM; RDF;, Metadata Aggregation; Engineering; EAC-CPF

* Corresponding author. Tel.: +34 669796498 fax: +34 669796498
E-mail address: reito@bib.uc3m.es

Selection and peer-review under responsibility of the 3rd International Conference on Integrated Information.
doi:10.1016/j.sbspro.2014.07.110

1. Introduction

One of the major issues in data retrieval today is providing users with the necessary tools to access metadata spread across different, remote repositories. To accomplish this objective, an integrated metadata registry (IMR) is proposed. IMR is based on the use of sematic metadata and controlled vocabularies (thesauri encoded in SKOS and authority records) to record the different kinds of semantic relationships between metadata and information resources.

This tool is part of the results obtained after a two year collaboration period between CEHOPU (Centre for Historical Studies of Public Works and Town Planning- CEDEX-Ministerio de Fomento) and the Library and Documentation Department of the Universidad Carlos III de Madrid.

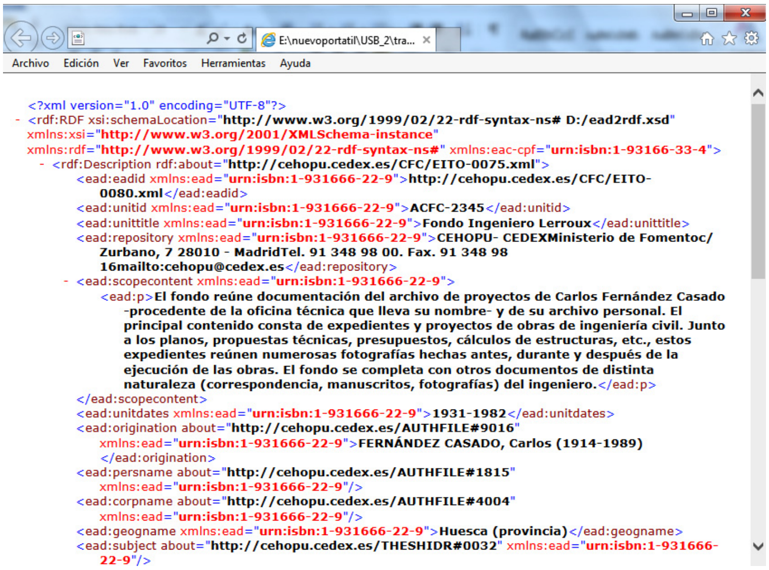
The initial scope of this project was the creation of two separate websites to make accessible on the Internet the personal fond of the most important Spanish civil engineers. As a result of analyzing the access methods proposed for these sites, the research team identified the need of building a technical framework where additional documents coming from a distributed network of information services could be located. One of the critical aspects of this vision was the fact that the different kinds of relationships between information resources and their metadata should be made explicit to enable end-users an easy exploration of the metadata repository. The solution combines the advantages of traditional controlled vocabularies for information retrieval with the benefits of full-text and qualified information retrieval technologies.

2. Conceptual architecture

The proposed system is based on the use of standard metadata schemas (EAD, Dublin Core, MODS) and controlled vocabularies (thesauri and authority records). Information professionals working on the description and cataloguing of materials can asign access points (keywords, descriptors) from shared, web-accessible vocabularies accessible through the SRU (Search/Retrieve URL) standard protocol (Reiss, K. 2007).

Once the metadata records are completed, they can be automatically collected by an unattended process. Metadata records are collected as RDF records containing a subset of the whole record metadata: main descriptive metadata, like title, the creator, dates, a abstract and keywords that correspond to different aspects treated in the document: subject, persons, corporations, geographic locations, etc. As long as the values for these subjects are taken from different controlled vocabularies and authority files, the RDF record keeps the URI (Uniform Resource Identifier) for the concepts and entities represented by these values. The usage of URIs is necessary to ensure that the different information centers providing materials to the IMR use the same identifiers to refer to the same concepts, even if they opt to make any change on the labels displayed to the end users.

The resulting RDF file contains an <rdf:Description> element for the information resource itself, and for the different entities (persons, corporations, geographic locations, etc.) and subjects that are assigned to the finding aid as access points. The <rdf:Description> element works as an envelope where the minimum metadata for the referred entity or subject is included (basically the URI, a descriptive label and the existing dates in the case of persons and institutions). The <rdf:Description> element corresponding to the finding aid contains references to the URIs of the entities and subjects included within the same RDF file, in order to assure the correct interpretation and consistency of the metadata.



```
  <?xml version="1.0" encoding="UTF-8"?>
- <rdf:RDF xsi:schemaLocation="http://www.w3.org/1999/02/22-rdf-syntax-ns# D:/ead2rdf.xsd"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns:eac-cpf="urn:isbn:1-93166-33-4">
  - <rdf:Description rdf:about="http://cehopu.cedex.es/CFC/EITO-0075.xml">
        <ead:eadid xmlns:ead="urn:isbn:1-931666-22-9">http://cehopu.cedex.es/CFC/EITO-
          0080.xml</ead:eadid>
        <ead:unitid xmlns:ead="urn:isbn:1-931666-22-9">ACFC-2345</ead:unitid>
        <ead:unittitle xmlns:ead="urn:isbn:1-931666-22-9">Fondo Ingeniero Lerroux</ead:unittitle>
        <ead:repository xmlns:ead="urn:isbn:1-931666-22-9">CEHOPU- CEDEXMinisterio de Fomentoc/
          Zurbano, 7 28010 - MadridTel. 91 348 98 00. Fax. 91 348 98
          16mailto:cehopu@cedex.es</ead:repository>
      - <ead:scopecontent xmlns:ead="urn:isbn:1-931666-22-9">
          <ead:p>El fondo reúne documentación del archivo de proyectos de Carlos Fernández Casado
            -procedente de la oficina técnica que lleva su nombre- y de su archivo personal. El
            principal contenido consta de expedientes y proyectos de obras de ingeniería civil. Junto
            a los planos, propuestas técnicas, presupuestos, cálculos de estructuras, etc., estos
            expedientes reúnen numerosas fotografías hechas antes, durante y después de la
            ejecución de las obras. El fondo se completa con otros documentos de distinta
            naturaleza (correspondencia, manuscritos, fotografías) del ingeniero.</ead:p>
        </ead:scopecontent>
        <ead:unitdates xmlns:ead="urn:isbn:1-931666-22-9">1931-1982</ead:unitdates>
        <ead:origination about="http://cehopu.cedex.es/AUTHFILE#9016"
          xmlns:ead="urn:isbn:1-931666-22-9">FERNÁNDEZ CASADO, Carlos (1914-1989)
          </ead:origination>
        <ead:persname about="http://cehopu.cedex.es/AUTHFILE#1815"
          xmlns:ead="urn:isbn:1-931666-22-9"/>
        <ead:corpname about="http://cehopu.cedex.es/AUTHFILE#4004"
          xmlns:ead="urn:isbn:1-931666-22-9"/>
        <ead:geogname xmlns:ead="urn:isbn:1-931666-22-9">Huesca (provincia)</ead:geogname>
        <ead:subject about="http://cehopu.cedex.es/THESHIDR#0032" xmlns:ead="urn:isbn:1-931666-
          22-9"/>
```

Fig. 1. Sample XML data used by the system.

The RDF records are regularly collected by the IMR using HTTP requests and responses. The information from the different RDF files are processed and merged into an XML Topic Map (XTM) that keeps all the subjects, person, organization and geographical names as well as the relationships between them and the URLs of the information resources where these relationships can be traced. This topic map contains a separate topic for each entity (person, corporation, geographic location, and subject) used as an access point in any of the processed RDF records. The references to the indexed documents and records are treated as occurrences that provide information about these entities, following the topic maps philosophy.

In this way, the IMR highlights the relationships between the different entities involved in the creation and custody of the indexed materials. IMR, based on a global XTM, offers a flexible way to capture the different types of relationships. The use of topic maps to facilitate retrieval of different types of data has been analysed by different authors (Yi, M. 2008), (Tramullas, J. S.; Garrido, P. 2006). In addition to the information about entities and occurrences, the XTM topic map also records the different relationships between entities, and the context in which these relationships are identified. For example, if an engineer has worked for a specific company, this relationship is included in the topic map by means of a specific <association> element. This element is used to record the relationship between entities, as well as the type of relationship between them. The context in which this relationship is valid corresponds to the finding aids where this relationship is documented in the <scope> XTM element. The <scope> element is used to indicate "in which context" a relationship between entities exists. In this project, the <scope> element refers to the finding aids describing the documents where the relationship between entities is documented. In case there are different finding aids providing evidence of the same relationship between a pair of entities, the <scope> element shall be repeated as many times as needed. Both the entities and the relationships are typed, as requested in the Topic Maps specification.

The processing of the RDF (W3C (2004). RDF/XML Syntax Specification (Revised)) (W3C (2004). RDF Vocabulary Description Language 1.0) files to generate and increment the contents of the XTM file is done by means of a Visual Basic program. This software completes different steps: check whether the entity being processed already exists in the XTM file by means of the URI, check whether the relationship with the other entity already exists, etc. These controls ensure the integrity of the data in the IMR and avoid the creation of duplicated entries or relationships.

The XTM file is the actual metadata registry. It may be seen as a single, big XML database containing all the data needed to retrieve and access the distributed finding aids and explore the relationships between entities from a single location. The XTM file is indexed with the Oracle Berkeley DB XML to provide users with full-text and qualified searching facilities, so they can combine an exploratory approach based on browsing with direct access to the information via the indexes built with that tool.

IMR administrator can execute a process that automatically generates a web-based publication from the data in the XTM topic map. The process generates HTML pages for browsing by means of successive XSLT (W3C (2007). XSL Transformations (XSLT) Version 2.0) transformations run in an unattended way. The resulting HTML files contain hypertext links to enable users navigate across the data repository.

3. Conclusions

The IMR has proven to be a valid approach to ensure metadata aggregation and discovery in a network of distributed information centers. Although the activity has been initially planned for information centers managing metadata encoded in EAD, EAC-CPF, Dublin Core and MODS, the technical solution is fully compatible with other metadata schemas.

The proposed usage of XTM is interesting to demonstrate the potential of this specification for metadata discovery. XTM gives the choice of going beyond traditional integrated indexes, providing researchers with the possibility of discovering relationships between the entities and subjects treated in different documents. The combination of full-text XML indexing with a semantic-based browsing of the whole repository of metadata offers an appropriate balance to enable the navigation through large information spaces.

References

Reiss, K. (2007). "SRU, Open Data and the Future of Metasearch." Internet Reference Services Quarterly 12(3/4): 369-386.

Tramullas, J. S. ; Garrido, P. (2006). "Constructing Web subject gateways using Dublin Core, the Resource Description Framework and Topic Maps." Information Research 11(2): 1-1.

W3C (2004). RDF Vocabulary Description Language 1.0: RDF Schema, W3C Recommendation, 10 February 2004. Retrieved April 09, 2011, from http://www.w3.org/TR/rdf-schema/

W3C (2004). RDF/XML Syntax Specification (Revised), W3C Recommendation, 10 February 2004. Retrieved April 09, 2011, from http://www.w3.org/TR/rdf-syntax-grammar/

W3C (2007). XSL Transformations (XSLT) Version 2.0: W3C Recommendation 23 January 2007. Retrieved April 10, 2011, from http://www.w3.org/TR/xslt20/(consultada 03/01/2009).

Yi, M. (2008). "Information organization and retrieval using a topic maps-based ontology: Results of a task-based evaluation." Journal of the American Society for Information Science & Technology 59(12): 1898-1911.