DCPAPERS
Proc. Int'l Conf. on Dublin Core and Metadata Applications 2011

Metadata Aggregation in Historical Engineering Archives: Building an Integrated Metadata Registry

Ricardo Eito Brun
Universidad Carlos III de
Madrid, Spain
reito@bib.uc3m.es

Abstract

This communication describes a prototype project completed by a research team of the Universidad Carlos III de Madrid to build an integrated metadata registry (IMR) for historical engineering archives. This registry was developed as part of a project completed for the Ministerio de Fomento of Spain. The proposed solution offers a way to collect and aggregate metadata from a network of archives that hold historical fonds of civil engineering documents. To enable metadata aggregation and ensure metadata compatibility, the archives participating in the network are requested to share authority records encoded according to the definitive version of EAC-CPF and use descriptors from a set of thesauri published by the Spanish Ministry.

The developed prototype makes use of automated remote calls through HTTP to collect metadata from EAD (Encoded Archival Description) and EAC-CPF records created by the different archives and then process them to build an XML Topic Map (XTM) on XML format. The registry itself consists of an XTM file that is later processed to build the different pages that the end-users and researchers use to navigate the registry and discover the information spread through the different fonds and collections.

Keywords: metadata aggregation; Encoded Archival Description; Encoded Archival Context; EAC-CPF; EAD; engineering archives; RDF; Topic Maps; XTM.

1. Engineering archives. Need of Metadata Integration

Engineering archives are usually spread across different archival institutions. Usually, each archival institution describes their fonds as separate entities. Even if these descriptions are stored within the same database, current approaches to archival description do not make visible to researchers the relationships that exist between the engineers, architects, government agencies, companies, locations and subjects (engineering and construction techniques, materials, etc.) used by the engineering community in a specific period. This constraint of traditional approaches to archival description was identified during a collaboration project 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 fonds of two of the most important Spanish civil engineers: Eduardo Torroja and Carlos Fernández Casado. As a result of analysing the access methods proposed for these sites, the research team identified the need to build a technical framework where additional fonds and documents coming from a distributed network of archival institutions could be searched, and where the relationships between the different entities and subjects used as access points in the different finding aids could be explored to identify additional knowledge.

https://doi.org/10.23106/dcmi.952135919

2. Sharing EAC-CPF Authority Records and Controlled Vocabularies to ensure semantic compatibility

Archives providing metadata to the IMR were provided with an editing tool they can use to create and edit their finding aids and descriptive records. Anyway, they could use any other editing tool as long as it provides the capability of working as an SRU client (Reiss, 2007).

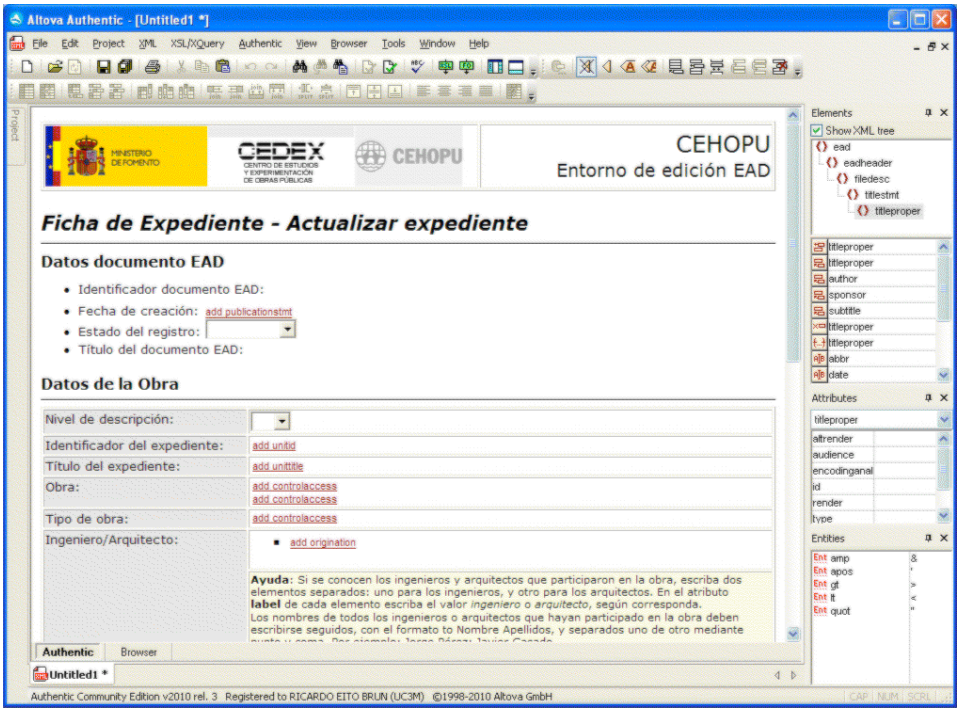

FIG. 1. Sample RDF file generated from an EAD finding aid

The EAC-CPF and the SKOS repositories are native XML databases stored in a single server. The SRU interface to search these databases from the EAD/ISAD(G) editor is implemented by means of XML web services on the PHP and VBScript programming languages. Archivists cataloguing materials can select the remote, controlled vocabulary they want to search, enter the terms and restrict the search to different choices: just preferred terms, any term in the vocabulary, terms with a meaning more specific than the proposed one, etc. In the case of the authority records, searches can be restricted to the different elements within the EAC-CPF specification. The technical communication between the EAD editing tool and the EAC-CPF and SKOS repositories is implemented by means of SRU requests and messages.

Figure 2 shows the interaction between the EAD editing tool and the SKOS/EAC-CPF remote repositories:

https://doi.org/10.23106/dcmi.952135919

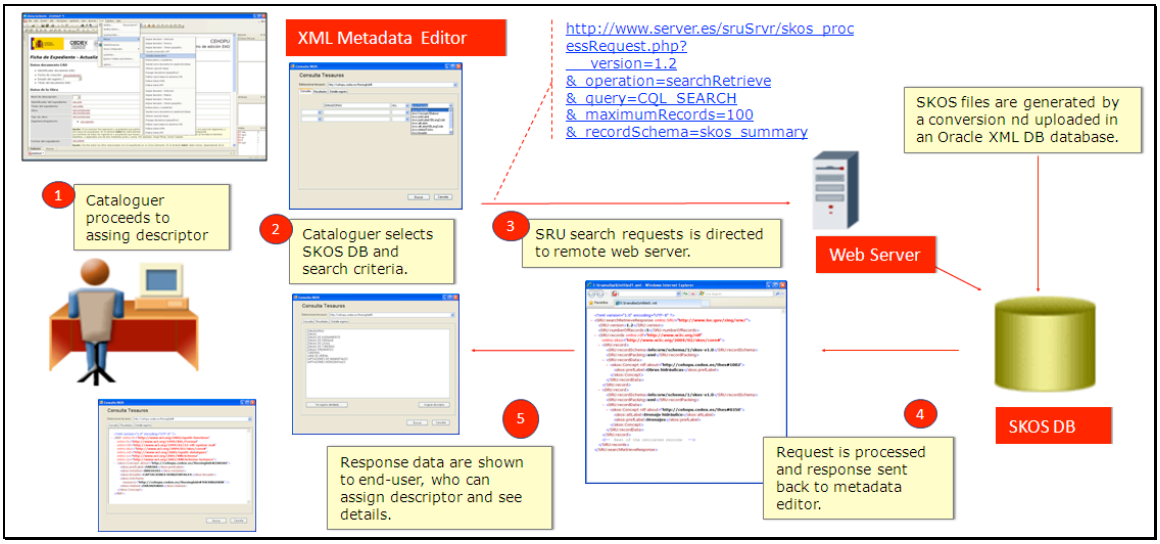

FIG. 2. Sample RDF file generated from an EAD finding aid

3. Capturing Metadata from EAD records

Once the EAD record is completed, the EAD editor automatically generates an RDF record with a subset of its data. This subset corresponds to the main descriptive metadata, like the unit title, the originator, unit dates, a summary description and the different keywords assigned by the cataloguer. These keywords correspond to different EAD elements like <subject>, <persname>, <corpname> or <geogname> (that is to say, the elements proposed in EAD as access points and usually grouped into the <controlaccess> element. As the values for these elements are taken from different controlled vocabularies and authority files, the RDF record keeps the URI (Uniform Resource Identifier) for the concepts or entities represented by these values. The usage of URIs is necessary to ensure that the different archives 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 finding aid itself, and for the different entities (persons, institutions, 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.

https://doi.org/10.23106/dcmi.952135919

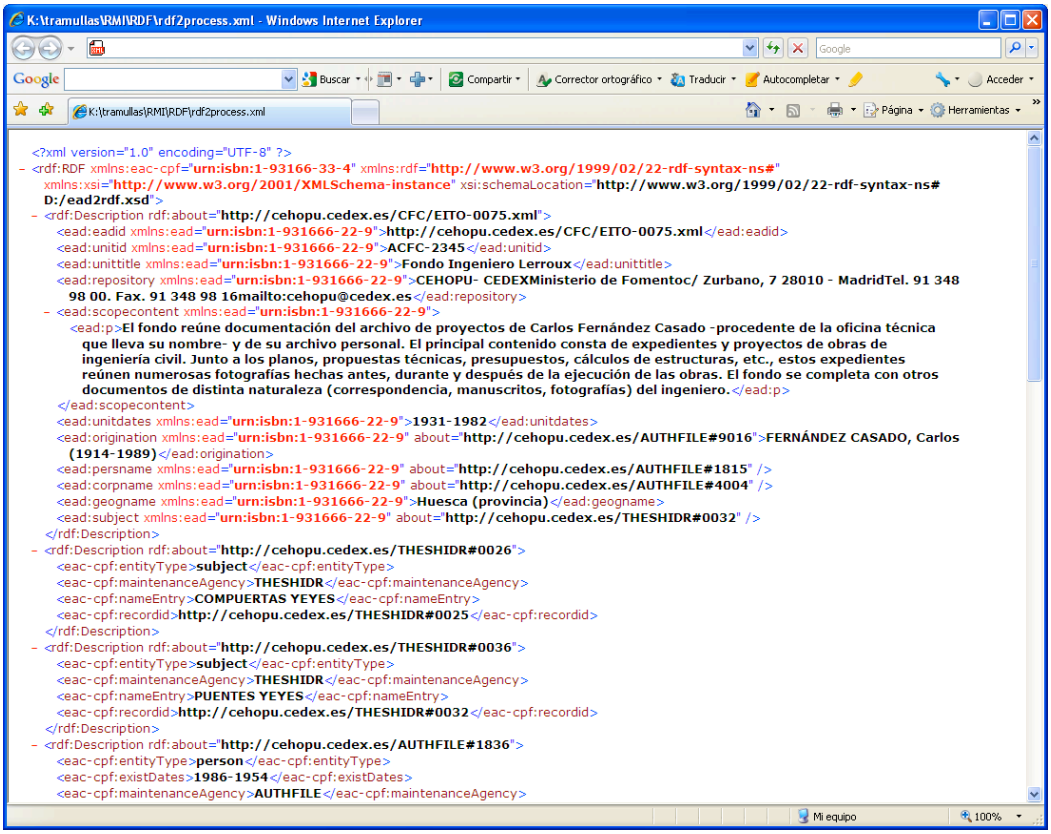

FIG. 3. Sample RDF file generated from an EAD finding aid

From a technical point of view, the RDF records are generated by means of an XSTL stylesheet that transform each EAD file into a RDF record. The resulting RDF record is stored in a specific folder of the local web site where it may be harvested by the IMR.

3. Metadata Harvesting and Integration in the XML Topic Map

The IMR regularly collects the generated RDF records from the list of participant centers using the HTTP protocol. Although the choice of using the OAI-PMH standard was initially considered, it was finally discarded, and a simpler, custom protocol based on http requests was put in place to collect these files. RDF was selected as the format to expose subsets of metadata taken from the EAD descriptions, as this is a well-known standard from the W3C and its usage gives high opportunities for reuse across different projects. The RDF files are collected and processed by the IMR server to build the integrated metadata registry that consists of an XML Topic Map (XTM). Data are converted from the RDF representation used for harvesting metadata into XTM, as the IMR purpose is to highlight the relationships between the different entities involved in the creation and custody of the engineering works and their related documents. The authors believe that XTM offers a more flexible way to capture these relationships than RDF. The possibility of using RDF as the standard format for storing the IMR data using RDF tools like Sesame® is the subject of future research. The use of topic maps to facilitate retrieval of different types of data has been analysed by different authors (Yi, 2008), (Tramullas, 2006). The topic map created 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 EAD finding aids are treated as occurrences that provide information about these entities, following the topic maps philosophy.

https://doi.org/10.23106/dcmi.952135919

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 may 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 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.

4. Generation of the IMR Publication

The XTM 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. It is important to note that finding aids are locally published by each archive, and that the IMR only contains the URL and the basic metadata for them; in addition, recording and aggregating the relationships between entities regardless the fonds, archive or finding aid where they are documented gives researchers a powerful tool to explore the distributed collections, as users can browse and identify relationships beyond the scope of a specific, individual archive.

To enable end-user navigation and browsing, the IMR also incorporates a module for the automatic generation of a web-based publication. This process starts making an split of the XTM file into separate XML files (one file per entity). This intermediate XML file contains the basic metadata about the entity, the list of entities to which it is related and the context (i.e. references to the finding aids) that provides evidence of these relationships. The process to generate the HTML pages for end-users is implemented in this way (using XSLT transformations running in an unattended way) to avoid implementations based on server-side pages and components that require a more complex infrastructure.

Once the split is completed, a second step is to generate automatically an HTML file from each intermediate XML. The resulting HTML file contains all the necessary hypertext links to enable users to navigate across the publication. These hyperlinks point to other HTML pages within the IMR web site that correspond to related entities. Figure 4 shows the HTML page for an entity (an engineer), and at the right hand side of the page there are different links to access the persons, institutions, companies, places and subjects related to the works of this engineer. By clicking on these links a list of related entities and the corresponding context is displayed. In the case of links pointing to the finding aids (this happens with the occurrences and with the context of the relationships), the link heads the user to the file maintained by the archives in their local web sites. Full text search is implemented on the collection of HTML pages by means of the Google (using the customization capabilities provided by this search engine).

https://doi.org/10.23106/dcmi.952135919

FIG. 4. HTML page generated for an entity (person).

5. Conclusions

The IMR has proven to be a successful approach to ensure metadata aggregation and discovery in a network of distributed archives. Although the activity has been initially planned for historical archives containing civil engineering documentation and using archival standards like EAD or EAC-CPF, the technical solutions developed as part of this project are fully compatible and may be easily deployed in other collaboration scenarios working with other metadata standards like Dublin Core, MODS or other kind of materials. IMR also shows the possibility of applying Semantic Web standards like RDF to facilitate metadata exchange and integration. The proposed usage of XTM is interesting to demonstrate the potential of this specification for metadata discovery. XTM gives the option of going beyond traditional integrated indexes, providing researchers with the possibility of discovering relationships between entities recorded on distributed documents accessible through the network. For researchers, especially in engineering archives, their interests go beyond the discovery of "related documents", and they also need to explore which are the techniques, methods or materials applied by engineers, their relationships with other colleagues and companies, or the places where they worked. This kind of information may be extracted from the metadata initially assigned to the descriptive records to create a richer information space for researchers as those requested by relevant authors like Pitti (2006).

References

Pitti, D. V. (2006). "Technology and the Transformation of Archival Description." Journal of Archival Organization 3(2): 9-22.

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.

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.

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)

https://doi.org/10.23106/dcmi.952135919

W3C (2004). Resource Description Framework (RDF): Concepts and Abstract Syntax, W3C Recommendation, 10 February 2004. Retrieved April 09, 2011, from http://www.w3.org/TR/rdf-concepts/.

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 (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/.

https://doi.org/10.23106/dcmi.952135919