The CPDOC Semantic Portal: Applying Semantic and Knowledge Organization Systems to the Brazilian Contemporary History Domain

Abstract
Presents the semantic portal project of the center for teaching and research in the Social Sciences and Contemporary History (CPDOC) of the Fundação Getúlio Vargas, Rio de Janeiro. This project involves the use of semantic and visualization technologies and natural language processing techniques to allow enhanced ways to access the CPDOC collections.

Introduction
CPDOC is a major center for teaching and research in the Social Sciences and contemporary History located in Rio de Janeiro, Brazil. CPDOC is also the leading research institute in the country and holds a major collection of personal archives, oral histories and audiovisual sources pertaining to Brazilian contemporary history. It is part of Getúlio Vargas Foundation (FGV), a prestigious Brazilian research and higher education institution founded in 1944.

This paper describes the semantic portal project being developed at CPDOC and its partner, the Applied Mathematics School from FGV, along with all the initiatives that are being undertaken in order to achieve the final goal. Among those initiatives we can highlight the development of two ontologies; one for the field of Brazil’s contemporary history and another one that aims to standardize all document metadata throughout the CPDOC archives. The final goal is to implement the descriptive information about the archives in a RDF triplestore, applying semantic technologies, and connecting the digitized archives to the Linked Open Data (LOD) hub, granting access to the World Wide Web users via a single and unified interface. Inspiration for this initiative was the STAR project (Binding et al, 2010), that used similar techniques in the archeology domain.

The Archives
The CPDOC archives are multimedia in nature, being composed by many collections comprising journals, books and documents in audio, text, video, pictures, drawings and photographs. The main collections are organized as such:

- Personal Archives (AP): About 200 archival files, summing up to 1.8 million documents including text, images and videos.
- Oral History Program (PHO): A huge set of testimonies (in audio and video) consisting of more than 1000 interviews, which add up to five thousand hours of recordings.
- Brazilian Historical Biographic Dictionary (DHBB): the current version includes 7553 entries, of which 6584 are of biographical nature and 969 relate to institutions, events and concepts of interest in Brazilian history after 1930.
In 2008, the CPDOC initiated a huge digitization project, converting most of the archives to electronic format. In 2010 the archives had 300,000 textual documents, 65 film rolls, 106 tapes (VHS, Beta and U-MATIC), 350 Long Plays, 187 K7 tapes, 85 roll tapes and about 32000 photographs from the personal archives files. Besides that, 5000 hours of interviews from the oral history program were also digitized and all these documents are available for consultation by researchers locally in the institution. When the digitization is completed approximately 80,000 photographs will be available for Web access, comprising the whole set of images donated to the Center. Additionally, all the DHBB are already in the digital format.

The problem

The main problem to be addressed is the heterogeneity and multimodality of the archives, and the lack of descriptive metadata standards for all types of document. For a semantic portal to work, we do need a single and unified standard to describe each item in the archives, and also a domain ontology - in this case, the Brazilian contemporary history. The three main collections, or ”systems”, as they are called internally – AP, PHO and DHBB – have different sets of controlled vocabularies of their own. As they were built independently over time and by various different groups of researchers and analysts, each one of the three presents a great degree of semantic idiosyncrasy. The lack of links between concepts – internally to the systems, not to mention external relationships with other collections outside FGV – undermines the exploitation of the myriad of relations among the items, making every collection insulated. Besides that, and in spite of the single search interface it shows today on the website, the technologies used are independent too, impeding cross search among the collections. The main search tool is a full text index, that yields high recall but low precision, and the net effect is that the collections are under-utilized.

The proposed solution

The presented scenario is a complex one, and requires a number of reasonably independent initiatives. In the following sections, we will present the solution's goals, the main set of tasks to be fulfilled and the initiatives taken so far. The main goals are:

- To offer unified and integrated access to the whole set of CPDOC systems (AP, PHO and DHBB);
- To give the users a rich set of navigational options: concept or hierarchy based navigation, and also full-text search;
- To enable Searches that traverse different systems, finding all relevant results – achieving good balance between precision and recall;
- To be able to present the search results in a faceted manner by the most relevant metadata elements (e.g. media format, collection name, etc)
- Homogenize concept terminology – no more syntactic idiosyncrasies;
- Adopt widely used metadata formats and ontologies (e.g. DCMI, FOAF, Bibliographic Ontology, COMM, EAD);

1 http://cpdoc.fgv.br/acervo/arquivospessoais
2 http://cpdoc.fgv.br/acervo/historiaoral
3 http://cpdoc.fgv.br/acervo/dhbb
4 http://dublincore.org/
5 http://www.foaf-project.org/
6 http://bibliontology.com/
7 http://comm.semanticweb.org/
8 http://www.loc.gov/ead/
• Development of specific ontologies for the domain of the Brazilian contemporary history and enhance documents description by the addition of uncovered characteristics / attributes;
• Building a RDF triples database – each triple will connect a document and its metadata, allowing integration with other cultural heritage resources via the Linked Open Data (LOD)
• Migration of the current knowledge organization systems (KOS), as glossaries, thesaurus and controlled vocabularies to the SKOS \(^9\) format and integration to the ontologies. All resources will have an URI;
• Promote enhanced visibility to search engines, through revamping the portal interface and eliminating the current need to login to access;
• Facilitate the process of document tagging with the new metadata standards; as there will be a huge amount of documents that will need to be analyzed and described;
• Integrating with the Learning Objects Databases and the FGV Digital Library.

These goals are the main guidelines to actions and to realize these there are several ongoing projects and initiatives. These initiatives are described below:

1. Face and character recognition project: Has as its main goal to offer an optimized process for tagging the images in the AP archive, by means of computational facial recognition of characters. A software was developed to apply facial recognition techniques to associate faces with characters described in the image caption – an idea pioneered by Google's Picasa \(^11\) software. This software allows for a friendly way to browse the AP photography archives, together with the identification of the characters in each photograph. As the faces are recognized, each photograph would generate a RDF triple linking the photograph URI to the KOS that contains the names of relevant personalities in Brazilian contemporary history. The main interface of the developed solution is shown in Fig. 1.

![Fig. 1: The VIF (Very Important Faces) software main window – character's tagging](image)

\(^9\) http://linkeddata.org/
\(^10\) http://www.w3.org/2004/02/skos/
\(^11\) http://picasa.google.com/
2. Sound and text alignment project: Its goal was to produce automatic alignments and, therefore, transcriptions of voice and video recordings for the Portuguese language, in the scope of the PHO archives. The recorded audio consists of interviews. Some of the recordings have already been transcribed by humans while others are available as audio only. This process would link the excerpts of interviews to the transcriptions, and these transcriptions would be harvested using natural language processing techniques to extract common concepts – allowing it to be linked to the existing ontologies.

3. Text mining project: Consists of a set of initiatives conceived to give support to both the “Face and character recognition” and “Sound and text alignment project”. Here, important descriptors are mined from image captions, associated documents, and transcriptions, thus generating RDF triples associating to the domain ontologies.

4. DHBB “Wikification” project: Was conceived to promote a deeper interconnection of CPDOC’s internal databases with external ones, such as Wikipedia. The benefits of this interconnection are improvements in public visibility and the establishment of social collaborative networks to contribute/correct the collections. It is being implemented with the open source semantic wiki software MediaWiki ¹² with its semantic extensions ¹³. Entries are being migrated from the current system to the wiki to demonstrate the capabilities of the functionalities of the tool. This project benefits from the ontologies that are being created, that will serve as semantic annotations to the entries.

5. Contemporary History Ontologies: after realizing the lack of formal KOS resources regarding the domain, it is being developed as a master’s dissertation, the definition of a conceptual model for contemporary history ontologies, covering competency questions that all ontologies must deal with, and the main facets: Agent (person, group or organization), Role, Event, Place, Time and Document (source of historical information). The huge (10,000+), non-hierarchical list of concepts currently used for the description of the archives, along with the entries of the DHBB, will be good sources for the creation of domain ontologies. The first light ontology developed under this model deals with the “military government years”, a phase in Brazilian history also called as the “dictatorship years”.

6. Description Ontology: Another ontology, the CPDOC description ontology is being developed as part of a doctoral research. This research includes analysis of the descriptive metadata used in each archival file nowadays; the harvesting of existing metadata standards for all kinds of documents and media (Dublin Core, Bibliographic Ontology, FOAF, COMM), and issues of compatibility with archival description standards (NOBRADE, ISAAR(CPF), ISAD(G)).

7. Implementation of a RDF triplestore: in order to store the triples built – all the triples associate a document URI with an ontology concept URI. There will be set a Sesame ¹⁴ database to serve as the triplestore. We will have to generate URIs for each document item stored in the current Oracle database that holds all the digitalized documents.

All the projects are independent, in spite of the fact that they converge and aid the idea of a Semantic Portal, as depicted in Fig.2.

The main activities described earlier are shown in the framework above. In the leftmost we have the CPDOC archives, that is being digitized and, in this process, natural language processing techniques are being used to find all the access points they will share, and the

---

¹² http://www.mediawiki.org/wiki/MediaWiki
¹³ http://semantic-mediawiki.org/
¹⁴ http://www.openrdf.org/
concepts for linking are brought from the domain ontologies that are being developed. The RDF triplestore connects those ontologies to each item in the collections, and allows the portal interface to offer concept based and hierarchical search, along with a SPAQRL query builder for advanced users. The portal architecture is still being discussed.

**Fig. 2: The Semantic Portal Framework**

**Current Developments**

The whole project began in 2010, with character recognition tasks. Nowadays, many tasks are being developed in parallel. The researchers’ academic backgrounds involve mathematics, history, engineering, information science, computer science archival science, librarianship and even movie science and biology. Those are present in the main groups, with the tasks of developing KOS (ontologies and thesaurus), NLP tasks, image processing, sound processing and automatic RDF triples creation. In spite of some budgetary constraints, there are good results so far, and we do expect that the portal can be fully operational by the end of 2013. As independent projects, each one will offer functionalities and features that can enhance access to the archives. We hope that the project can offer a new resource for a huge community of researchers in the field, and also add one of the first Brazilian contributions to the Linked Open Data information hub.

**References**