# XML Annotation of Historic Documents for Automatic Indexing *

Cristina Ribeiro[12], Gabriel David[12], André Barbosa[12]

[1] FEUP—Faculdade de Engenharia da Universidade do Porto
[2] INESC — Porto
Rua Dr. Roberto Frias s/n, 4200-465 Porto, Portugal
{mcr,gtd,andre.barbosa}@fe.up.pt

**Abstract.** This paper describes the process and tools used in the development of a digital documentation center for historic documents. A team of archivists has handled document organization and description according to archival standards. The document repository and the access interface are supported on a prototype multimedia database. The focus here is on the treatment of the textual contents.

One of the goals of the project is to provide a search mechanism for the documents. Plain full-text search is of limited use in this kind of documents, where only archaic Portuguese or Latin versions of the text are available. The designations used for places, people and even common objects have diverse forms and are hard to recognize for non-specialized users.

The first step of this work has been the definition of a set of tags to mark relevant aspects of the documents, followed by the specification of a compact XML dialect. An XML editor and a set of associated tools were then developed. The editor helps archivists to enrich documents with content annotations that have high potential as search terms. This set of tools is currently in use and the annotated documents are indexed by a search tool sensitive to the annotations.

**Keywords:** XML Edition, XML dialects, Cultural Heritage Applications.

## 1 Introduction

Digital technologies are raising the public expectations with respect to access to cultural heritage. The traditional custodians of historic materials (archives, libraries, museums) are being prompted for providing access to various items in digital form. The diversity of available assets requires concern with their organization and search.

A researcher who deals with historic information sources needs access to the original documents or, at least, to good quality reproductions. That is one of the reasons for the extensive digitization programs going on. But images are

* Partially supported by FCT under project POSC/EIA/61109/2004 (DOMIR)

hard to search and the skills to read ancient documents are confined to a group of specialists who do not always agree on a given interpretation. A document image can be complemented by a transcription of its contents in text format. Depending on the language of the original document and the intended readers, a translation to a modern language may also exist. Summaries and comments by the historian may also condense or explain the document. All these forms are searchable and correspond to different reading levels.

Understanding a historic document requires knowledge of its content, of what it says, who is involved, persons or institutions, which are the places mentioned, and so on. But it also depends on who has created the document, when, for what purpose, aspects related to the context of the document, that explain its relevance. Archivists usually deal with this second aspect in a relatively structured way, following description rules according to international standards [1,2]. Storing such information is compatible with a relational database approach [3]. However, the work done by historians on the analysis of the document content is by nature a lot more dependent on the document itself and thus inherently semi-structured. It may contain the identification of certain words in the document as representing persons, places, events, etc. A summary or comments by the historian may be included in a more or less systematic way. This kind of information is better dealt with using an XML approach, using the document text and suitable annotations.

A second reason for using XML is its ability to express relationships between different elements through the use of XML attributes. Combined with authority files, this is an efficient way not only to mark, for instance, a certain expression as a person's name, but also to associate it to the person in the authority file, regardless of the specific name form used in the mention.

The kind of documents at stake is the textual document subject to digitization. But, as soon as the database is able to store images, why not storing sound or video? The whole world of multimedia databases becomes an issue. In this world content descriptions are a lot more varied than in the textual content. The relevant international standards [4] are based on XML and this constitutes the third reason for choosing XML in databases of historic documents.

Finally, to index the combination of database fields and XML documents under an approach of full text indexing but assigning different relevance factors to different tags or database fields, first a transformation of the latter into XML is performed. After that, it is possible to apply XML enabled indexing and querying tools. XML thus plays the role of a glue tying the document content and the several metadata components together.

The paper is organized as follows. Section 2 justifies the importance of having wide computer based access to historic documents. Then an architecture to improve their retrieval is presented in section 3. The component that is the subject of the paper is the annotation editor for historic documents presented in section 5, after the discussion of its schema in section 4. Some conclusions are put together in the final section 6.

## 2 Access to Cultural Heritage assets

Cultural heritage assets are complex objects: they may include for instance historic texts, photographs, songs, tales and performances, along with various texts, which convey the reflections that specialists have produced when analyzing them.

For a cultural heritage institution, it is more and more the case that its objects have been subject to some sort of treatment. The most common is description, and in museums, archives and libraries it is expected that objects only begin their life in some form of custody after they have been described. But other manipulations are also frequent. Digitization is one of them, performed either to safeguard especially valued items from frequent manipulations, to explore the digital content with specific automatic tools or simply to ease the access and expose the available collections. Another one is content analysis in several perspectives and to different depths.

Online public access to cultural heritage has grown as a consequence of the availability of digital versions of objects, the expansion of networking and generalization of systematic use of databases in the management of object collections. The huge numbers of items in these collections require some form of organization. The experience of archives points to a hierarchical organization with flexible number of levels and level designations but with a uniform set of description attributes.

Access to cultural heritage repositories must allow for a spectrum of different user profiles, ranging from the technical staff that describes objects to the manager of the objects or collections who chooses the structure of the collection, from the lay user that may engage in browsing digital representations to the scholar who can contribute valuable annotations on the history or features of the objects.

When digital versions of the objects are available, and specially if they already have specialized metadata attached, it becomes obvious that they might be made available for browsing and searching. Due to the diversity of objects, the different metadata standards and the different modes of use, there is currently no uniform solution for this task [1,4,5].

## 3 Searching Historic Documents

Historic documents are usually written records testifying facts considered important for historic studies. They are kept in public or private archives where they are preserved and put at the disposal of the users, according to specified access rules. Due to their uniqueness, the preservation concerns have always conflicted with the goal of providing access to this memory of the society. Reproduction has been the answer for it, in microfilm and, nowadays, by digitization and subsequent computer storage. Very high quality digitization is also seen as a way to ensure long time preservation as it can be subject to lossless reproduction, even if the original decays.

The approach implied in this paper has been developed in two previous projects, both the model and an implementation. In the first one, Archivum [3], a hierarchical model for the storage of contextual metadata has been designed, thus addressing the problem of dealing with large sets of documents. In this model, the description of a document is a leaf in a tree whose root is the fonds, or structured collection. Implemented in a relational database, it supports searches on the description attributes, returning references to the place in the archive where the actual document can be found.

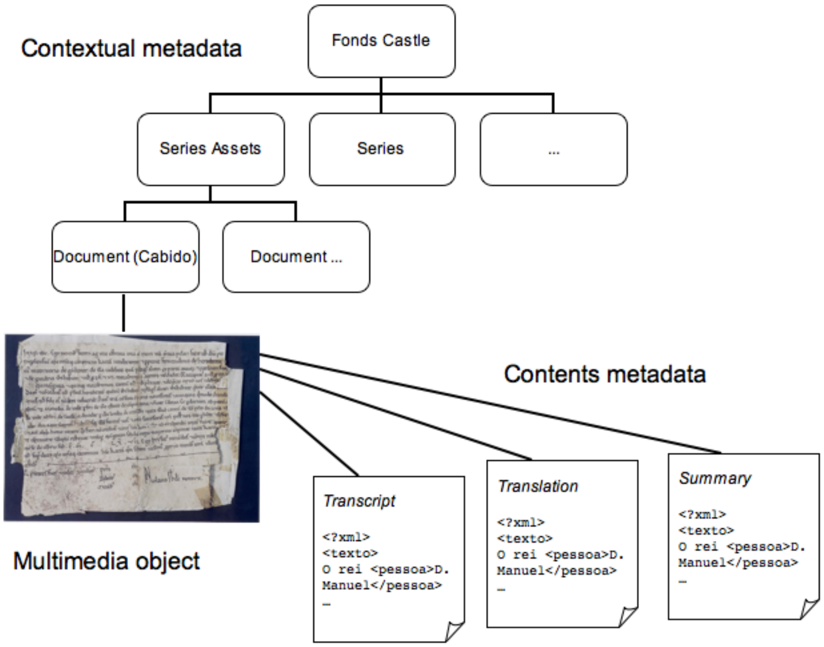


**Fig. 1.** Information architecture: contextual metadata, multimedia object and content metadata

In the second project, Metamedia, the goal has been to extend the ideas to deal with also storing the "documents", taken in an enlarged sense encompassing multimedia objects. For instance, a news program on a TV broadcaster could be seen as a document, as well as the digitization of a judicial process, or the digitally signed message of an e-commerce transaction. The goal has been to build a multimedia database able to display the object along with its description.

Soon after it was recognized that the richer nature of the objects called for a new part in the model. It is called the contents metadata and is primarily concerned with adding to the multimedia object descriptors related to its contents, like for instance the color distribution of a picture, the identification of scene cuts or the subtitles in a video or the melody in a music record. There is a large number of such descriptors in international standards [4,5] and more could be devised. Due to the variability of their structure, it is not practical to force them into a relational model. Most of them are defined in XML and an arbitrary number can be associated to the object they describe. The search paradigm supported by these descriptors is no longer the satisfaction of a boolean expression, with no meaning in many cases, but some form of similarity search.

These ideas are fed back to the representation of historic documents, which are basically textual but where the image containing the scanned version plays such an important role, substituting for many purposes the actual document. In this setting (see example in Figure 1), the information relevant for a single document may be located in a few different places:

- contextual metadata: description at the level of Fonds (relational record; inherited by the document);
- contextual metadata: description at the level of Series (relational record; inherited by the document);
- contextual metadata: specific document description (relational record);
- multimedia object: high quality digitization (plays the role of information source; picture);
- multimedia object: low resolution digitization (for preview; picture);
- content metadata: transcription of the original into an annotated text format (XML descriptor);
- content metadata: translation to modern Portuguese (XML descriptor);
- content metadata: summary (XML descriptor).

The previous sections describe the query model for the end-user. However, to organize the production of metadata in particular content metadata consisting in text and markup., a different approach is more suitable: the documents are classified by fonds and a directory is created for each fonds; transcriptions and translations are then put together in the corresponding directories; the specialized XML Editor, before opening a single document, analyzes the directory in search for common definitions, relevant to the specific document.

## 4 A Dialect for Document Annotation

The project goals impose concern with both the technical description of documents and the exploration of the meaning of its textual content. The technical description is handled by archivists and uses the ISAD/ISAAR archival description standards [1,2], comprising items such as title, summary, creator, owner and dates. For the textual content, processing is required. The first step,transcription of the document text, is not an automatic task due to the nature of the documents but had already been produced for the documents in the collection. The result is in most cases a text in Latin or archaic Portuguese. The History experts have identified several annotations that can be associated to words or phrases in the text, clarifying their meaning. Examples of such annotations are those marking the name of a person or of an institution, a title of nobility or a place.

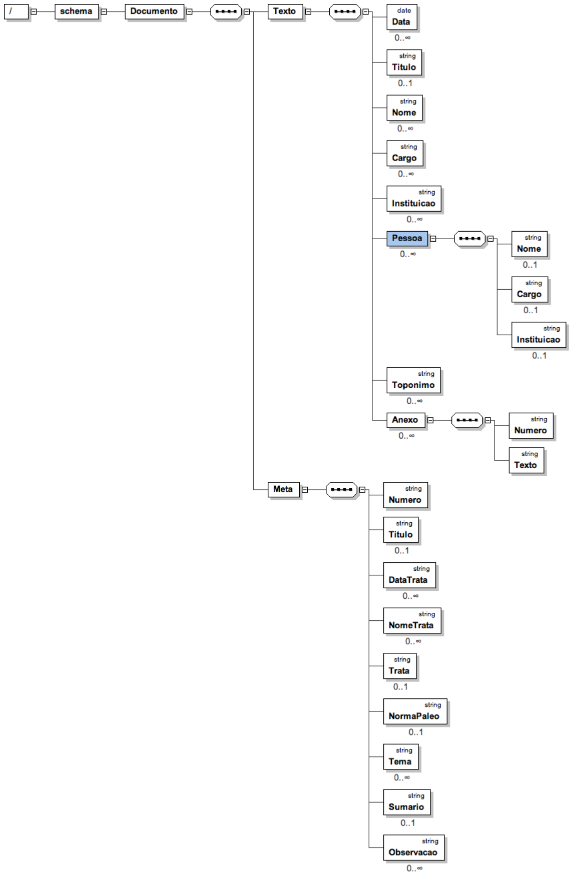


**Fig. 2.** XML Schema for the annotation dialect

The MetaMedia platform [6] has been proposed for supporting the descriptive metadata, providing the environment for archivists to edit and structure the documents in the collection and for regular users for browsing and searching. After collecting annotations on document contents it is possible to have search performed on a combination of descriptive metadata, full text and annotations. In the case of medieval documents, annotations are fundamental for taking advantage of the document content in search.

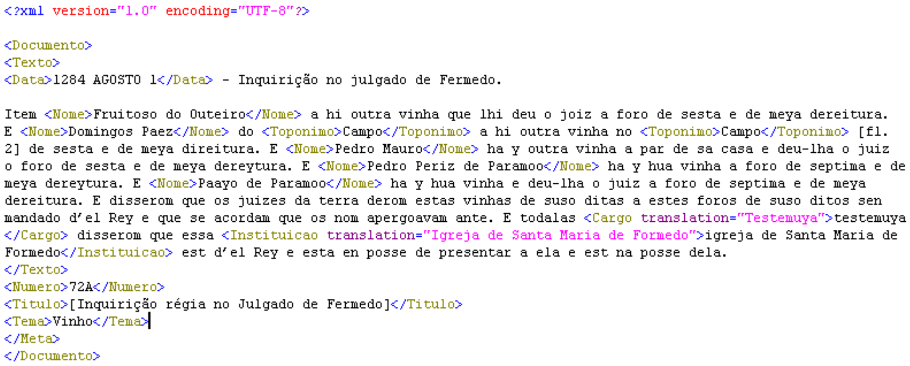
XML has been the obvious choice for the annotation of documents. The very specific nature of the documents excluded the use of a standardized dialect. The annotations are included in the documents by the domain experts in the process of analyzing the documents. A simple and intuitive interface was required for their task.

The first challenge of the project was to specify an appropriate dialect that could be applied to annotate all the documents of the documentation center.

The issue has been discussed in the development team and the result is an XML dialect addressing the concepts identified for this kind of collection and with a focus on flexibility in order to encompass different kinds of documents.

The dialect is described with the XML Schema graphically represented in Figure 2. The structure of the annotated documents includes a set of metadata descriptors and annotations on parts of the text. The XML instances basically have a block of descriptors (document number and title, information on the scholars who treated the text, themes, a summary) followed by a block of text interspersed with annotations on expressions identifying people, places, nobility titles, among others. The Schema clearly distinguishes elements that qualify a piece of text from those that apply to the whole document. Annotations occur inside the *Texto* tag (text) and the so-called external annotations are inside the *Meta* tag (metadata). Annotations where translation occurs are used to specify the contemporary equivalent of some expression in the original text.

The root element is *Documento* (document) representing the document. Three elements are always present in the document: *Numero* (number), *Data* (date) and *Titulo* (title). The first is an identifier used in the documentation center, the second is a complex element containing information on meaningful dates related with the document that can also have a local (Toponimo) associated with the date and the third is the title of the document. Several other elements are defined such as *Nome* (name), *Cargo* (job), *Instituicao* (instituion), *Pessoa* (person), *Toponimo* (place), *Sumario* (summary), *Tema* (theme), *Observacao* (observation) and *Anexo* (annex). *Pessoa* is a complex element and contains optional information about name (*Nome*), job (*Cargo*) and institution (*Instituicao*). The other complex element is *Anexo*, containing informations about the identifier of an attached document (*Numero*) and its text, the content of a full document (*Documento*). Some of the elements are repeatable. Some can either be part of one complex element or just appear alone in the text.



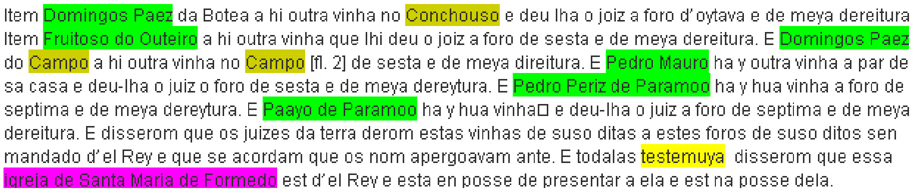
```
<?xml version="1.0" encoding="UTF-8"?>

<Documento>
<Texto>
<Data>1284 AGOSTO 1</Data> - Inquirição no julgado de Fermedo.

Item <Nome>Fruitoso do Outeiro</Nome> a hi outra vinha que lhi deu o joiz a foro de sesta e de meya dereitura.
E <Nome>Domingos Paez</Nome> do <Toponimo>Campo</Toponimo> a hi outra vinha no <Toponimo>Campo</Toponimo> [fl.
2] de sesta e de meya direitura. E <Nome>Pedro Mauro</Nome> ha y outra vinha a par de sa casa e deu-lha o juiz
o foro de sesta e de meya dereytura. E <Nome>Pedro Periz de Paramoo</Nome> ha y hua vinha a foro de septima e de
meya dereytura. E <Nome>Paayo de Paramoo</Nome> ha y hua vinha e deu-lha o juiz a foro de septima e de meya
dereitura. E disserom que os juizes da terra derom estas vinhas de suso ditas a estes foros de suso ditos sen
mandado d'el Rey e que se acordam que os nom apergoavam ante. E todalas <Cargo translation="Testemuya">testemuya
</Cargo> disserom que essa <Instituicao translation="Igreja de Santa Maria de Formedo">igreja de Santa Maria de
Formedo</Instituicao> est d'el Rey e esta en posse de presentar a ela e est na posse dela.
</Texto>
<Numero>72A</Numero>
<Titulo>[Inquirição régia no Julgado de Fermedo]</Titulo>
<Tema>Vinho</Tema>
</Meta>
</Documento>
```

**Fig. 3.** Example of Document Annotations (XML source)

### 4.1 Example Documents and Annotations

Figures 3 and 4 contain two views of an instance document conforming to the developed XML Schema. The first is the textual source document where the separation between the enhanced full text (element *Texto*) and the document metadata section (element *Meta*) are visible.

The view in Figure 4 comes from the Annotation Editor interface described in what follows. Only the base text is visible here, and the parts where content annotations were introduced are highlighted (each kind of annotation corresponds to a different color). The annotations inside the *Meta* element are not visible—they are assigned by choosing their menu entries in the editor.



Item Domingos Paez da Botea a hi outra vinha no Conchouso e deu lha o joiz a foro d' oytava e de meya dereitura Item Fruitoso do Outeiro a hi outra vinha que lhi deu o joiz a foro de sesta e de meya dereitura. E Domingos Paez do Campo a hi outra vinha no Campo [fl. 2] de sesta e de meya direitura. E Pedro Mauro ha y outra vinha a par de sa casa e deu-lha o juiz o foro de sesta e de meya dereytura. E Pedro Periz de Paramoo ha y hua vinha a foro de septima e de meya dereytura. E Paayo de Paramoo ha y hua vinha□ e deu-lha o juiz a foro de septima e de meya dereitura. E disserom que os juizes da terra derom estas vinhas de suso ditas a estes foros de suso ditos sen mandado d' el Rey e que se acordam que os nom apergoavam ante. E todalas testemuya disserom que essa igreja de Santa Maria de Formedo est d' el Rey e esta en posse de presentar a ela e est na posse dela.

**Fig. 4.** Example of Document Annotations (Editor view)

## 5 The Annotation Editor

The editor gives specialists an easy way to annotate the documents, generating files in the XML dialect specified in the XML-Schema. The editor has been designed with a focus on flexibility, so that changes in the underlying schema can be easily accommodated.

The editor is implemented as a Java applet, making it easy to execute in a web browser in any platform. Figure 5 shows the editor window, with a set of buttons at the top giving access to the generic file-manipulation functionalities and a side pane where the list of annotations for the main text are available.

For the collection of historic documents, content metadata was already available concerning aspects such as authorship or themes. The editor has been coupled with a set of tools to deal with the conversion of these metadata items into document annotations.

A controlled vocabulary has been adopted for describing document contents. The vocabulary includes two parts: a predefined thesaurus and the lists of terms created when editing a group of documents.

The features provided by the annotation editor therefore comprise file management, controlled vocabulary visualization and management, annotation visualization and management, and access to the conversion tools.

In the file management area we can find the general options like creating, opening, closing and saving a document. Management of the controlled vocabulary (*Vocabulario Controlado*) is performed according to the context of the

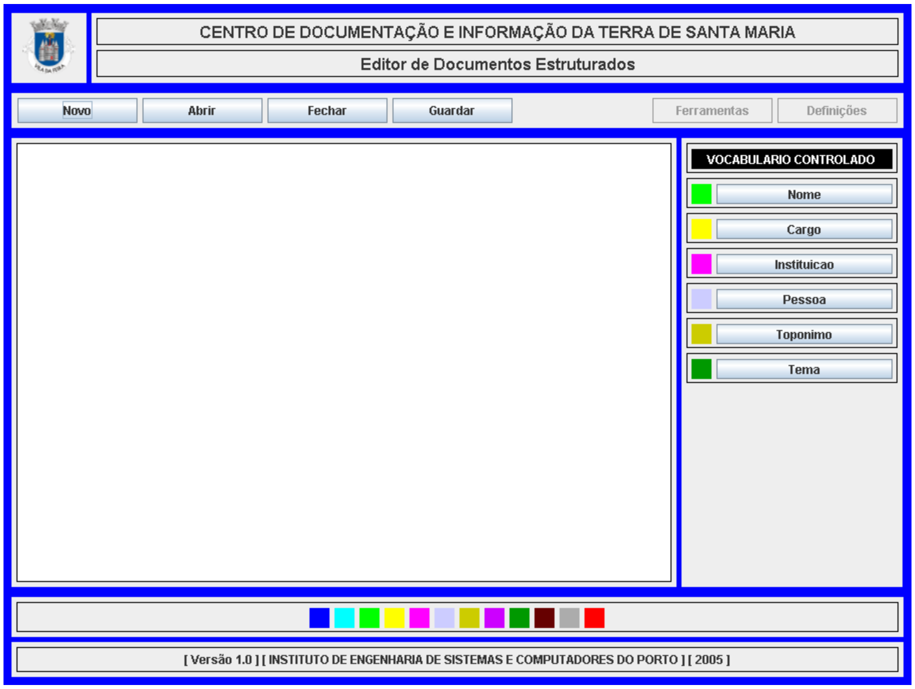

CENTRO DE DOCUMENTAÇÃO E INFORMAÇÃO DA TERRA DE SANTA MARIA
Editor de Documentos Estruturados
Novo
Abrir
Fechar
Guardar
Ferramentas
Definições
VOCABULARIO CONTROLADO
Nome
Cargo
Instituicao
Pessoa
Toponimo
Tema
[ Versão 1.0 ] [ INSTITUTO DE ENGENHARIA DE SISTEMAS E COMPUTADORES DO PORTO ] [ 2005 ]

**Fig. 5.** The Annotation Editor

document being edited. The annotations from documents belonging to the current working directory are available in the controlled vocabulary for the current document. New terms created when annotating the current document will be added to this contextual vocabulary and become available for the edition of documents grouped under the same directory.

To add new annotations to the document the user can use either the controlled vocabulary menu or the contextual menu (right mouse button). It is possible to consult all the annotations made in the document using the legend area with the color signs at the bottom of the editor window.

At any time it is possible to run another tool or change the definitions of the editor. In this case it is necessary to close the document edition first. After this the options tools (*Ferramentas*) and definitions (*Definições*) are enabled.

### 5.1 Conversion Tools

The conversion tools have been created to incorporate in the XML documents information already created by the domain specialists and for allowing the use of an existing thesaurus for controlling themes to be associated to the documents.

The Documents Converter is a simple tool to convert previous annotations made by the documentation center and stored in Microsoft Excel format to XML in the annotation dialect. This tool only requires the path of the original file and the path for the new file as shown in Figure 6.

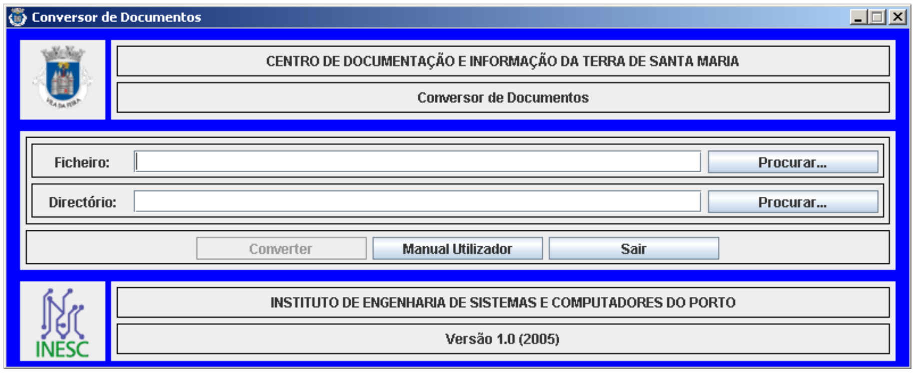


**Fig. 6.** Documents Converter

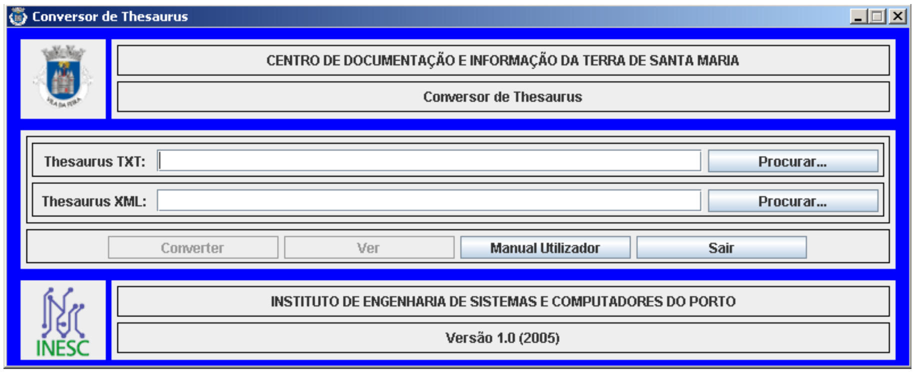


**Fig. 7.** Thesaurus Converter

The existing thesaurus is a hierarchic theme structure and has been received in Microsoft Excel format. The Thesaurus Converter creates a new representation used by the editor, converting the old format to XML. Just as in Documents Converter, this tool only requires the original and the destination paths. The interface is shown in Figure 7.

The Controlled Vocabulary Editor has been created to manage the annotations for the set of documents saved in a common directory. This vocabulary is saved as a binary file for efficiency. The tool is intended to read and edit it allowing the user to change this set of words and also to export the vocabulary

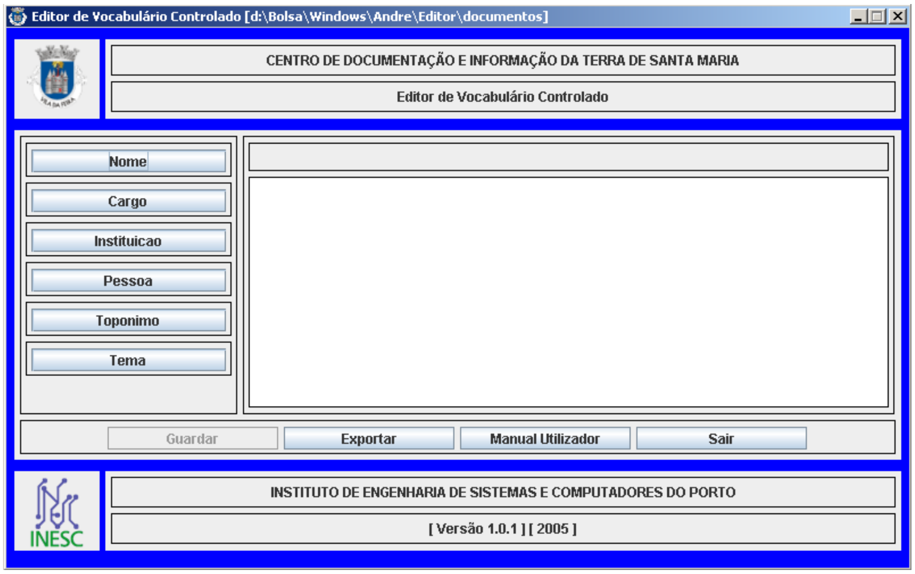

Editor de Vocabulário Controlado [d:\Bolsa\Windows\Andre\Editor\documentos]
CENTRO DE DOCUMENTAÇÃO E INFORMAÇÃO DA TERRA DE SANTA MARIA
Editor de Vocabulário Controlado
Nome
Cargo
Instituicao
Pessoa
Toponimo
Tema
Guardar
Exportar
Manual Utilizador
Sair
INESC
INSTITUTO DE ENGENHARIA DE SISTEMAS E COMPUTADORES DO PORTO
[ Versão 1.0.1 ] [ 2005 ]

**Fig. 8.** Controlled Vocabulary Editor

in XML format for later use or just for presentation purposes. Figure 8 shows its interface.

## 6 Conclusions

Text content annotation is a natural use for XML, that can easily handle the need for a flexible set of annotations appearing at arbitrary spots in the text while keeping generic descriptors associated to the document as a whole.

This work has been motivated by the collaboration in a project for a digital documentation center managing a collection of historic documents. The original documents belong to several fonds on the National Archives, and the collection includes high-resolution digitizations and textual transcriptions. For some documents translations are also available. The purpose of the application is to make the documents available to an audience ranging from scholars to the general public, providing search on the textual document contents. The straightforward approach of using the full text of the documents is not an effective solution, as most documents are not available in current-day Portuguese.

The Annotation Editor is currently being used by the History and Archives specialists in the team to create the annotated versions of the documents. Existing controlled vocabularies and a thesaurus for themes are integrated in the editing facilities. The document model captured in XML Schema supports the definition of specialized indexes. Indexing on separate document descriptors allows more flexible querying and browsing. The final application will use both the "enhanced full text" documents and their archival description records in the search interface.

The application supporting the work described here is a result of a line of research on multimedia databases and multimedia retrieval, seeking representations and methods for effective retrieval on structured multimedia objects. The current collection is being explored mainly on its textual content, but the availability of digitized versions of the documents will allow experimentation with image features as indexing dimensions and their use in the retrieval strategies.

## References

1. ISAD(G): International Standard for Archival Description (1999) http://www.ica.org/biblio/isad_g_2e.pdf (1.12.2005).
2. ISAAR(CPF): International Standard Archival Authority Record for Corporate Bodies, Persons, and Families (1995) http://www.ica.org/biblio/isaar_eng.pdf (1.12.2005).
3. Archivum: System of Objects with Temporal Support for Archival Description. Technical report INESC (1998)
4. ISO/IEC: JTC1/SC29/WG11—MPEG-7 Overview (2004) http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm (1.12.2005).
5. TV-Anytime: TV-Anytime Forum Website (2005) http://www.tv-anytime.org/ (1.12.2005).
6. Ribeiro, C., David, G.: A Metadata Model for Multimedia Databases. In Bearman, D., Garzotto, F., eds.: Proceedings of ICHIM01: International Cultural Heritage Informatics Meeting- Cultural Heritage and Technologies in the Third Millennium. Volume 1., Politecnico di Milano and Archives & Museum Informatics (2001)