XTM-DITA structure at Human-Computer Interaction Service

Piedad Garrido[1], Jesús Tramullas[2], Manuel Coll[1], Francisco Martínez[1], Inmaculada Plaza[3]
[1] Computer Science, Universidad de Zaragoza
44003 Zaragoza, Spain
[2] Document Management Systems, Universidad de Zaragoza
55009 Zaragoza, Spain
[3] Electronic Engineering and Communications, Universidad de Zaragoza
55009 Zaragoza, Spain

{piedad, tramullas, iplaza}@unizar.es, mkhollv@gmail.com

Abstract. This work describes a software engine which works with textual documents containing historical information. The purpose of this work three-fold: firstly to show the validity of the developed engine to correctly identify and label the entities of the universe of discourse with a labelled-combined XTM-DITA model. Secondly to analyze the improvements achieved in the interaction between people (users) and computers with a practical application of the designed methodology to a real-world problem in the semantic web area and thirdly to plan its future integration into a traceability system.

Keywords: Topic Maps, XTM, DITA, JDO, Lucene, Freeling, object-oriented database, historical information, human computer interaction.

1 Introduction

Textual documents containing historical information are a required source of reference for socio-cultural and historical research development. Besides, their importance grows exponentially as they are increasingly used in learning systems and virtual learning environments, and in the diffusion of the digital environment culture which addresses all audience types. The success of the Perseus Digital Library Project (http://www.perseus.tufts.edu) is a clear example of this. The growing number of historical document collections, both textual and multimedia, available at digital libraries will demand the development of more advanced data access tools and information presentation tools where Topic Maps standard [6, 7 and 8] will play a central role.

One of the main features of textual documents with historical information is their poor level of data structure. The wide range of document types that can be used in history research shows a lack of easy definable patterns, and when they do exist, they correspond to a highly specific series of documents. Then there is the additional problem of adding the irregular appearance of significant historical entities (people's names, locations, dates, acts, etc.), and what this entails within the textual documents themselves. Identifying these entities is one of the problems we face when we wish to integrate these documents into semantic web environments. These problems have been resolved thanks to embed DITA (Darwing Information Typing Architecture) [13] in the knowledge architecture.

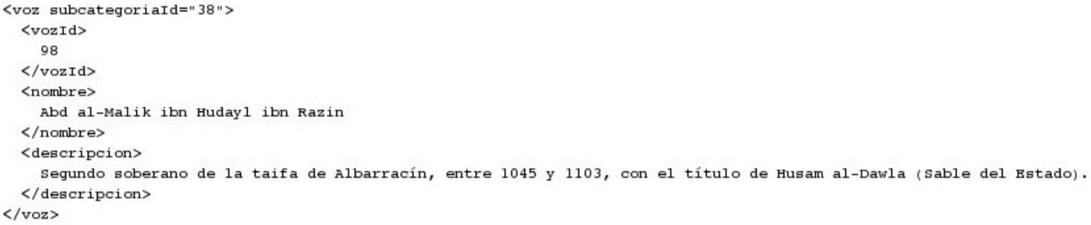
The aims of this work are on the one hand to show the engine identifies and translates Spanish language entities into XTM[1]-DITA scheme correctly (see figure 2). This way to label textual information improves the interaction between people's thoughts and computers and it allows getting better information visualization of the universe of discourse by applying heuristics [5] instead of restricting it to a textual knowledge representation (see figure 3). On the other hand to prove its versatility because of the fact that as information treated as similar (animal identifier, locations, dates, acts, etc.), we are going to test the idea in a traceability management system project for a pig slaughterhouse using RFID tags [9].

[1] XTM is the Topic Map specification. Available at: http://www.topicmaps.org/XTM

2 Engine Description

The processing sequence that the engine must follow is as so:

- Initialize and start the reading of the introduced XML source (see figure 1), provided by Dicom Medios[2], for testing the software with real-world application.



```
<voz subcategoriaId="38">
  <vozId>
    98
  </vozId>
  <nombre>
    Abd al-Malik ibn Hudayl ibn Razin
  </nombre>
  <descripcion>
    Segundo soberano de la taifa de Albarracín, entre 1045 y 1103, con el título de Husam al-Dawla (Sable del Estado).
  </descripcion>
</voz>
```

Fig.1. Example of XML source registers

- Use a Spanish language analyzer, called Freeling[3], to generate the necessary objects for each XML label in a way that is automatic and clear.
- Interpret the body of each entry of the source document to fill the generated objects according to an intelligent system based on rules and pattern matching techniques.
- Start the XML document generation to the XTM-DITA format from the data of the generated objects.

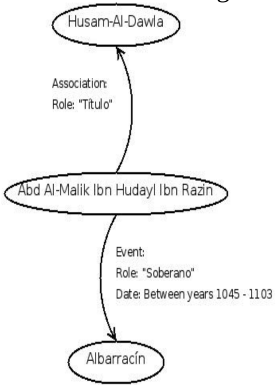


Fig.2. XTM-DITA registers structure

- Establish the persistence of these objects by means of an object-oriented database.
- Index the textual content of the objects in a Lucene[4] database.

All these proposed steps must be modular so that they can be externally modified by other people. The purpose of this characteristic is to refine these mechanisms so that they specialize in specific domains, thus providing certain versatility and to improve them with new techniques without having to recompile the whole system.

Our approach is based on the use of heuristics designed for this field of application. Working with these heuristics in an intelligent agent environment, which performs searches at an internal and external level of application, helps to provide a high level of independence which will allow this application to be used in various fields of knowledge. This approach has also verified the reliability of using XTM-DITA combined as proxies for semantic content [3]. Finally, we wish to point out that the semantic information extraction has been worked by parsing an XML document obtained through a search performed by a user in natural language.

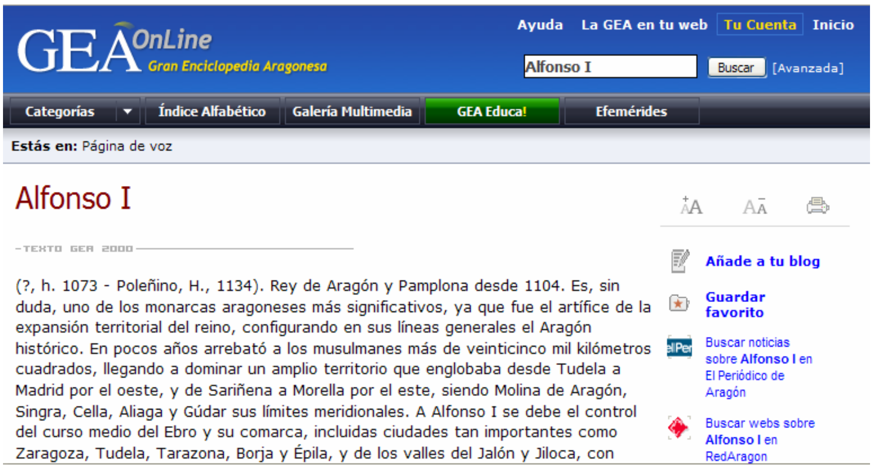
[2] The enterprise in charge of GEA (http://www.enciclopedia-aragonesa.com/)

[3] Freeling Home Page: http://garraf.epsevg.upc.es/freeling/

[4] Lucene Home Page: http://lucene.apache.org/

3 HCI Improvements

The Human Computer Interaction improvements achieved with this superimposed information approach to textual documents are on the one hand, the combination of DITA and XTM. DITA has many powerful visualization features: topic-based, localization-friendly, usability, consistency and minimalist. Several of them clearly oriented to user-centered design and, XTM allows end users to experience an interface that for once makes it possible to find the information they are looking for, quickly, easily, and intuitively (compare figure 3 and 4). On the other hand, the integration of mature software applications such as Lucene and Freeling are very important for the information retrieval process [1] and very useful for the usability versus usefulness aim to create future user interfaces [4]. Besides, technologies like JDO[5] whose benefits are portability, database independence, high performance, integration with EJB (Enterprise JavaBeans) and ease of use[6] helps to accomplish three of the seven user interface design principles: feedback, structure and tolerance.



Estás en: Página de voz

Alfonso I

TEXTO GEA 2000

(?, h. 1073 - Poleñino, H., 1134). Rey de Aragón y Pamplona desde 1104. Es, sin duda, uno de los monarcas aragoneses más significativos, ya que fue el artífice de la expansión territorial del reino, configurando en sus líneas generales el Aragón histórico. En pocos años arrebató a los musulmanes más de veinticinco mil kilómetros cuadrados, llegando a dominar un amplio territorio que englobaba desde Tudela a Madrid por el oeste, y de Sariñena a Morella por el este, siendo Molina de Aragón, Singra, Cella, Aliaga y Gúdar sus límites meridionales. A Alfonso I se debe el control del curso medio del Ebro y su comarca, incluidas ciudades tan importantes como Zaragoza, Tudela, Tarazona, Borja y Épila, y de los valles del Jalón y Jiloca, con

Fig.3. Current Graphical User Interface

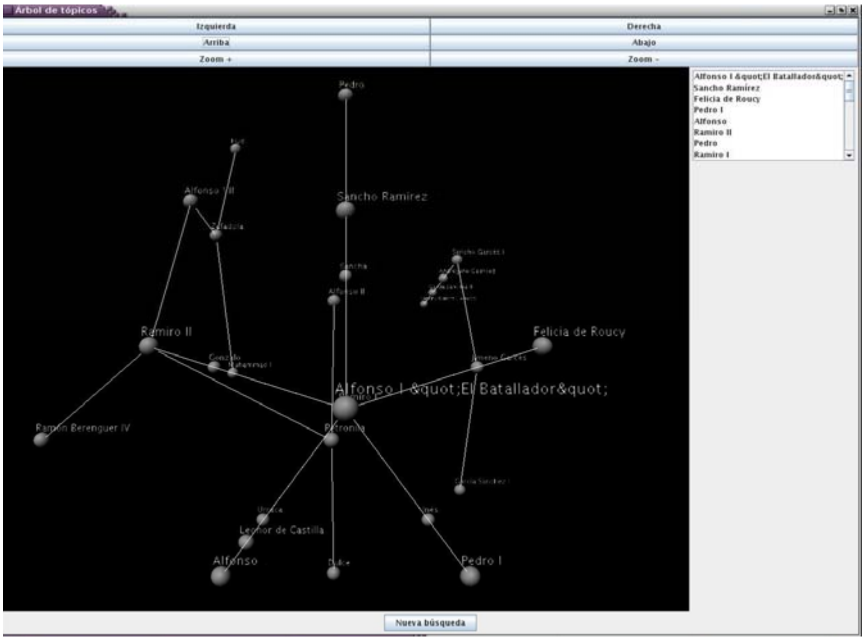


Fig.4. Graphical User Interface Prototype

[5] Java Home Page: http://java.sun.com/jdo/

[6] Application programmers can focus on their domain object model and leave the details of persistence (field-by-field storage of objects) to the JDO implementation.

4 Conclusions and Future Developments

In this paper, we advocated several ideas: that good full-text information processing is essential, especially in areas with poor levels of information structure such as textual documents containing historical information or traceability system using RFID tags [10, 11 and 12]; that modularity in software design is extremely important and allow the developed engine to run in any knowledge area without the need to recompile the whole system or to adapt it to different environments; and that combining object-oriented databases, mature software applications such as Lucene APIs, Freeling, Latent Semantic Indexing (LSI) and metadata is extremely useful to future information retrieval processes.

If all of these ideas are integrated into an intelligent agent framework, their potential is multiplied because intelligent agents offer many advantages in information retrieval processes and leave the system ready for the future development of recommender systems.

The object-oriented database management system, JPOX, and the Lucene database enable XTM-DITA tags to be suitable managed because the relationships among objects are more flexible and enriched in comparison with other systems developed under relational database management systems [5].

And finally, XTM-DITA visualization is a promising technique for both enhancing users' perception of structure in large information spaces and providing navigation facilities [14]. It also enables people to use a natural tool of observation and processing – their eyes as well as their brain – to extract knowledge more efficiently and find insights, as you can observe at figure 4 and some authors has used in topic maps-based digital course libraries [2] in e-learning environments. Nowadays, the usability and quality evaluation of the software prototype, in which the engine is built-in, is in charge of a research group of the University of Zaragoza called EduQTech (Education, Quality and Technology)[7].

References

1. Baeza-Yates, R. y Ribeiro-Neto, B. Modern Information Retrieval. Addisson Wesley, 1999.
2. Dicheva, D.; Dichev, C.; Dandan, W. Visualizing Topic Maps for e-learning. ICALT 2005, Fifth IEEE International Conference on Advanced Learning Technologies, pp. 950-951
3. Gelb, J. DITA and Topic Maps: Bringing Pieces Together. International Conference on Topic Maps, Oslo 2-4 April, 2008. Available at: http://www.topicmaps.com/
4. Green, P. Iterative Design. Lecture presented in Industrial and Operations Engineering 436 (Human Factors in Computer Systems, University of Michigan, Ann Arbor, MI, February 4, 2008.
5. Grossman, D. y Frieder, O. Information Retrieval: algorithms and heuristics. Springer, 2005.
6. ISO 13250: 2003. Information Technology—SGML Applications—Topic Maps
7. ISO 13250-2: 2006. Topic Maps—Part 2: Data Model.
8. ISO 13250-3: 2007. Information Technology—Topic Maps—Part 3: XML Syntax
9. ISO 14223/1:2007. Radio frequency identification of Animals, advanced transponders – Air interface
10. ISO 15961:2004. Information technology -- Radio frequency identification (RFID) for item management – Data protocol: application interface.
11. ISO 15962: 2004. Information technology -- Radio frequency identification (RFID) for item management -- Data protocol: data encoding rules and logical memory functions
12. ISO 19762-3: 2005. Information technology -- Automatic identification and data capture (AIDC) techniques -- Harmonized vocabulary -- Part 3: Radio frequency identification (RFID)
13. Linton, J. y Bruski, K. Introduction to DITA: A Basic User Guide to the Darwin Information Typing Architecture. Comtech Services, 2006
14. Sharp, H.; Rogers, Y., Preece, J. Interaction Design. Beyond Human-Computer Interaction. 2ª Ed. Chichester, Hoboken, NJ: Wiley, 2007.

[7] EduQTech Home Page: http://www.unizar.es/eduqtech/