Application of semantic tagging to generate superimposed information on a digital encyclopedia

Piedad Garrido, Jesus Tramullas, and Francisco J. Martinez

University of Zaragoza, Spain
{piedad, tramullas, f.martinez}@unizar.es

Abstract. We can find in the literature several works regarding the automatic or semi-automatic processing of textual documents with historic information using free software technologies. However, more research work is needed to integrate the analysis of the context and provide coverage to the peculiarities of the Spanish language from a semantic point of view. This research work proposes a novel knowledge-based strategy based on combining subject-centric computing, a topic-oriented approach, and superimposed information. It subsequent combination with artificial intelligence techniques led to an automatic analysis after implementing a made-to-measure interpreted algorithm which, in turn, produced a good number of associations and events with 90% reliability.

Key words: Topic Maps, DITA, XTM, electronic encyclopedia, automatic processing, artificial intelligence, superimposed information

1 Introduction

An encyclopedia is a work that covers all aspects of human knowledge. Basically, it is a comprehensive and complete reference work of which there are two main types: (i) in alphabetical order, used like dictionaries, but with much more information. Each volume contains the terms included between the two words that appear on its spine, and (ii) arranged as themes, known as systematic classification. Each volume deals with a different theme. To locate certain information, it is necessary to resort to the contents and the alphabetical index in each volume.

In recent years, many electronic encyclopedias have been published. At the beginning of this electronic era, they were diffused on optical devices like CD-ROMs and DVDs. Presently, their online format has spread and led to very complete multimedia documental products with which users may interact with thanks to their hypertext characteristics. Indeed, the diffusion of online electronic encyclopedia contents has led to such as an increase in the production of contents to be processed by documentation units that their manual processing proves practically impossible, at least by traditional techniques.

The online Gran Enciclopedia Aragonesa (GEA, the Great Aragonese Encyclopedia) [1] is a combination of online alphabetically arranged encyclopedias and those classified into categories. This format affects the way users act because once they have tried it, they prefer the real-time self-service format to the traditional encyclopedias to search for specific information.

GEA was published in print in 1981 but it was at the end of 2001 when it began its scanning, starting off the first version of GEA online in September 2003. The potential users of the systems are professional scholars, researchers and the general public. GEA is structured in XML voices, sorted into categories and subcategories and can be accessed through an index or a common integrated form on the website. This structure favours the interchange of information and communications on the one hand but it is still 'static' and 'rigid', which hampers the work of renovation of the voices, and their customers must resort too many levels of depth until they find the information they need. Furthermore, their searches per topic are limited to either the categories offered (art, biographies, sciences, geography, heraldry, history, humanities and entertainment), or a search done in alphabetical order.

The analysis of GEA is important because: (i) the process of updating the encyclopedia never ceases, (ii) the analysis of the Spanish language is very interesting and there are fewer initiatives regarding this, and (iii) it promotes the technology transfer between universities and enterprises. The aims of the project were: (i) to get a 'dynamic' version of the product ables to automate processing of textual documents with historic information in Spanish using free software technologies, and (ii) to start a research, development and innovation project between Dicom Medios and the University of Zaragoza.

This article analyses how superimposed information with XTM [2] and DITA [3], the knowledge-based strategy adopted, and artificial intelligence (AI) technologies may apply to improve textual documents with previously digitalised historic information for the purpose of assisting the preservation of these materials and of providing solutions to overcome the existing technical difficulties to make historic heritage perpetual. This proposal, which is based on a knowledge-based strategy combining superimposed information, subject-centric computing and topic-oriented approach, will: (i) facilitate the task of automating the documental analysis of content (semantic description) with this kind of textual documents, (ii) permit a more thorough representation of the contents, (iii) increase the possibilities of retrieving requested information, and (iv) adapt their use to each user's needs.

This document is arranged as so: Section 2 introduces some key concepts we are working with, such as superimposed information, ISO standard 13250 (Topic Maps) [2] and its XTM specifications, as well as the Darwin Architecture to transcribe information (DITA). Section 3 includes some works related to our proposal. Section 4 presents algorithms designed to automatically analyse textual documents containing historical information. The results obtained have been analysed as Section 5 explains. Finally, Section 6 offers the most important conclusions.

2 Main Concepts: superimposed information, subject-centric computing, and topic-oriented approach

The philosophy of work using a knowledge-based strategy which blends superimposed information [4], subject-centric computing [5], and topic-oriented approach [6] is based on documents needing a logical representation structure with which access to them may be facilitated. This logical structure is not normally deduced from textual information, rather from the contexts in which the documents were created and implemented, and in which these documents are used. Normally, the semantic description of a document is not limited to a single model that has been adapted to a discipline. This is because many metadata models exist with various forms of implementation that makes them all useful for solving a specific information representation problem. For this reason, combining several models helps accomplish richer semantics to subsequently enable simpler indexing and, in turn, allow more efficient search processes.

ISO standard 13250:2003 [7], known as Topic Maps, is a semantic model with a subject-centric vision used to represent the conceptual structure of the information contained in an online information resource. They are used to richly describe relationships between 'things' rather than between documents and pages, they improve the findability of information, and they are in a more high-level because they are human-oriented. The biggest contribution of such structures to knowledge representation lies in the fact that, when combined, they are much more effective than when considered separately [8]. In 2000, Topic Maps were defined using an XML syntax [2] known as XTM (XML for Topic Maps), which was updated in 2006.

DITA (Darwin Information Typing Architecture) is a XML specification, which IBM created in 1999, and it was transferred to the OASIS association in 2004. This content model is a topic-centered approach generated to design, produce and distribute technical information. The philosophy of this content model is based on the fragmentation of the content in short themes to facilitate their reuse in different statements. Since this is an extendable specification, different organisations possibly define the specific informative structures without relinquishing the use of generic tools [9].

We believe that the combination of both models could prove profitable to develop automatic systems that process large amounts of textual information as the former adds contextual semantics to the text by greatly minimising the problem since it only provides the information contained in it. In parallel, DITA covers the traditional part since it is based on its content and deducing the relevant information through content.

The most interesting characteristic to undertake the automated analysis of textual documents containing historic information is based on the fulfilment of two conditions: (i) the presentation of a series of occurrences, taking place at a given time and in a specific place, associated with a group of entities, and (ii) most of an entity's intrinsic information is performed by virtue of relationships with other entities. The semantics of these relationships lies in the roles they play. In this sense, all the smaller fragments into which we are interested in dividing each entity (the internal structure) have to be identified beforehand to subsequently interrelate the entities (the external structure) either in the encyclopedia database computer system or with other external online information sources. Later a content analysis has to be done which, in the terminology used in this research work, is the equivalent of the semantic description of these documental contents. The information retrieved after the analysis could have a network structure in which each node is an entity related with other nodes. Effective semantic information for the human user would be stored in the tags of the arcs containing the role of the internode association, along with the date and place if dealing with a dated event. The response to all these challenges is based on a hybrid scheme employed between the XTM specification and the DITA architecture. Both are models that define the way all the associations among the entities are stored and interpreted, and independently describe each entity from the rest.

3 Related Work

We now discuss the different application types and developments related to our proposal: (i) applications that carry out the analyses of event (see Table 1), and (ii) the developments which use the XTM markup languages (see Table 2) and the DITA ones.

Table 1. List of tools that analyse events

Name	Description
ESA	**Event Structure Analysis** It is programmed in JAVA, works with XML and is based on a qualitative methodology to understand the sequence of events and how they link people and things through narrative prose. Only examples of its use with textual documents and historical information are available in the article of Richardson [10] on the Everett massacre in 1916, and in Griffin's historic sociology matter [11].
TABARI	**Textual Analysis By Augmented Replacement Instructions** This open-source system is based on pattern recognition. It is designed to work with short summaries based on three forms of information; actors (proper names), verbs that determine the actions among actors, and phrases to distinguish among the various verbal meanings and to supply syntactic information related to the position the verb takes in the phrase.

To the best of our knowledge, only two contributions have been made [10, 11] which use these textual document analysis tools with historic information. Unlike our proposal, neither works with superimposed information nor uses a language analyser including events analyses, and they do not include an indices base capable of searching among various document types because the document to be imported in nearly all the applications has to be supplied in flat text. Table 2 presents the most outstanding Topic Maps-related developments.

Table 2. Specific developments which use Topic Maps and their XTM specification

Name	Description
Merlino	Merlino is a prototype of the semi-automatic events generation. It takes a Topic Map as input, creates search queries, and uses many search engines to automatically identify relevant information resources and to identify them as events of the topic supplied for the search. It is set up in Perl and its strong point is its ability to express semantic relationships in Topic Maps using a powerful search engine.
Siren	**Semantic Information Retrieval Environment for Digital Libraries** It appears in digital libraries as a Topic Maps-based semantic information retrieval environment. It is part of the DMG-Lib (Digital Mechanism and Gear Library), a set of components that make up a collaborative work environment which uses the TMwiki and Merlino tools in some of its development phases [12].
Houston	The work of Ann Houston and Grammarsmith, presented in TMRA 07, entitled 'Automatic Topic Map Generation from Free Text using Linguistic Templates', shows how using the textual information available online and Ontopia's Omnigator software automatically constructs a Topic Map by comparing passages from a free text with linguistic templates [13].

We now highlight the works done using DITA. First the work of Hennum [14] that uses information superimposed with DITA and SKOS to manage the formal subjects of the document content. Then Gelb's contributions to international congresses which consider the theoretic use of Topic Maps and DITA [15] which integrated a methodology known as SOTA (Solution Oriented Topic Architecture) into the development of projects involving content management. Next there is Garrido's practical approach [16] to demonstrate that work with superimposed information done with both markup text languages improves human-machine interaction subjects when working with textual documents.

The nearest development to the proposal presented herein, that is, the automatic processing of textual documents with historic information, is Wittenbrik's work [17] that considers the incorporation of encyclopaedic information online by using the international Topic Maps norm, which could only be consulted previously in a printed form. The genuine contribution to this predefined structure is made only at the coding level. Topic Maps enabled us to map legible data online to an available semantic structure. Thus its tagging, among other things, was not carried out automatically.

Briefly, if we wish to correctly digitalise, store, process and diffuse textual historic documents in Spanish, we conclude that there is no tool other than our proposal available in the market to do this.

4 Designed Algorithms

The aim of analysing the available textual documents with historic information is to extract a series of relationships among the entities and a series of events that, collectively, describe the relevant information. So the algorithm used for automatic processing purposes must be capable of: (i) reading and interpreting the text, (ii) detecting relevant information, (iii) extracting it, (iv) shaping the association among entities or for an entity-related event, and (v) storing it in one or several ways.

The model designed assumes the same information resource to be the object of several representations, all of which are interdependent within the process but have different results. However, the crux of the problem lies in: (i) the detection method used to determine which specific part of the processed textual document is converted into an association or event, and (ii) the method to infer the tagging and to indicate the role, date, place and entities participating. For this purpose, our system is able to use two different approaches (see Table 3).

Table 3. Developed algorithms

FIRST-LEVEL ALGORITHM	**Detection method:** Based on the word search in the text **Inference method:** To check against a predefined knowledge-based system
SECOND-LEVEL ALGORITHM	**Detection method:** To process natural language at the morphological, syntactic and semantic levels **Inference method:** A typical rule engine which incorporates a series of morphological combinations.

4.1 The first approach (first-level algorithm)

The first of the two approaches we implemented was a first-level algorithm where the analysis was done directly on the text. The text was first fragmented into phrases to identify associations and entities, and then into words. Each word was analysed to determine if it was the role of an association (by inspecting a small database of relevant roles: successoral lines, hierarchies, titles, etc.), or whether it was a word which mentioned an entity (detecting if it was a proper name). When an entity was detected, the presence of a role in the rest of the phrase was investigated, and if successful, an association among the topics was created. To detect events, a basic exploration based on finding parentheses was followed to indicate a date of an event in most cases.

This exploration algorithm has been classified as a first-level algorithm because detection is based directly on the words in the text, and its associations and events are produced from a search of keywords which provide clues as to the presence of relevant information. Therefore it deals with the subject of indexing by establishing relationships between natural language and documental languages. However, the basic service has not been developed further since the analysis essentially depended on a predefined knowledge-based system, and it failed proportionally to the complex expression of natural language regarding its variation and linguistic ambiguity characteristics. So we had to go one step further by distinguishing between the information analysis and the projection phase.

4.2 The second approach (second-level algorithm)

The next step is a second-level approach in which the information analysis phase would be independent of grammars, codes, predetermined deductive languages and predefined knowledge-based systems because it would be impossible to cover the peculiarities of processing the natural language from Spanish with such complex structures. The intention was to firstly eliminate complex expressions and use others to be subsequently analysed by a simple processor. In our particular case, the projection phase involved having to regenerate the documental base with a different structure, so it was necessary to develop a second-level processing algorithm. This uses natural language processing techniques at the morphological, syntactic and semantic levels such as a typical rule engine as a novelty to incorporate a set of morphological combinations which, unlike what is normally done, conducts an in-depth analysis to neither identify only certain significant structures, nor to lead to the loss of relevant information, a mistaken extraction or redundant information.

At the morphological and syntactic levels, the Freeling software package [18] was integrated and adapted to correct certain errors made (cataloguing common names as proper names as they are at the beginning of sentences, assigning verbs to common names, etc.) and to solve some widely used acronyms (e.g., Z. or id.). Moreover, a twin-type system was included to process linguistic ambiguity in such a way that a word could be simultaneously catalogued in several ways to avoid errors in the following analysis phases. An example of a fictitious entity:

'Al comienzo de su reinado, se casó con Magdalena de Folcaquier' (At the beginning of his reign, he married Magdalena de Folcaquier). Following the syntactic and morphological level: 'A (SP, preposition) comienzos (NC, common noun) de (SP, preposition) su (DP, possessive determiner) reinado (NC, common noun) casó (VM, Spanish conjugation of the verb 'casar') con (SP, preposition) Magdalena de Folcaquier (NP, proper noun)'.

At the semantic level, this morphological structure would be verified with the series of predefined algorithm patterns for the purpose of finding relevant information. If one of the predefined patterns was 'VM SP NP', then this sentence would fulfil this pattern as from the word 'casó'. If we look closely, we can observe how a wide spectrum of sentences describing an action on another entity is covered with this pattern.

Table 4 presents the different steps of our second-level algorithm. This algorithm receives the original information split into XML voices with a simple semantic description (i.e. $voz, vozid, nombre, descripción$). After processing the original voices applying our knowledge-based strategy which combines subject-centric computing, topic-oriented approach, and superimposed-information, the obtained output is semantically richer (see in Figure 1).

Table 4. Steps of the second-level algorithm

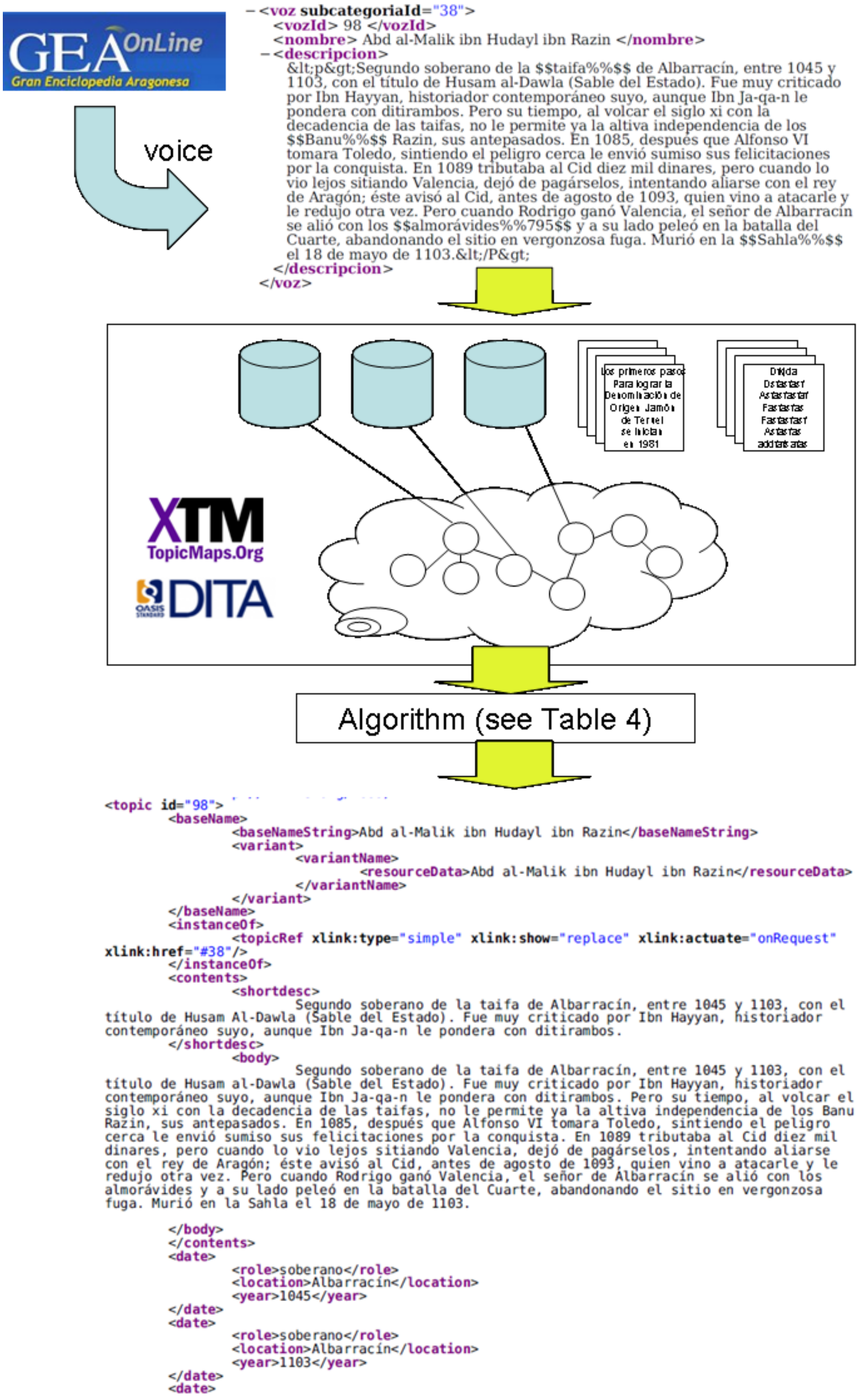
Step	Task	Step	Task
1	Filtering out the stop words	7	Obtaining a series of lists of the associations and events
2	Word-by-word analysis	8	Verifying the lists and search by references.
3	Phrase-by-phrase analysis	9	Having generated the dependences, the lists of associations will be calculated by a cross-search
4	Structure check against a series of ad-hoc patterns	10	The relevance of the roles of both the associations and events will be calculated
5	Eliminating redundant information with summary patterns	11	A rescue method will be called to detect orphan associations
6	Detecting events based on three pattern types		

5 Performance evaluation

This section presents the analysis of the results obtained from using the two algorithms proposed for the automatic text analysis. The relevance of the roles of both the associations and events that the two algorithms detect will be calculated by two different approaches.

The first approach consists in assigning the preset relevances associated with the predefined roles by rendering the remaining roles null relevance. In this way, the algorithm will assign relevances in relation to this predefined table. However, this method is not flexible.

The second approach, which is used in the final algorithm, consists in counting the roles. The more frequent the roles, the greater their relevance, and vice versa. This is a very flexible method and depends on neither predefined tables nor information domains. Having assigned relevance, a filter is run to eliminate the most irrelevant data (that is, the data with the least frequent roles).



```
– <voz subcategoriaId="38">
    <vozId> 98 </vozId>
    <nombre> Abd al-Malik ibn Hudayl ibn Razin </nombre>
  – <descripcion>
      &lt;p&gt;Segundo soberano de la $$taifa%%$$ de Albarracín, entre 1045 y
      1103, con el título de Husam al-Dawla (Sable del Estado). Fue muy criticado
      por Ibn Hayyan, historiador contemporáneo suyo, aunque Ibn Ja-qa-n le
      pondera con ditirambos. Pero su tiempo, al volcar el siglo xi con la
      decadencia de las taifas, no le permite ya la altiva independencia de los
      $$Banu%%$$ Razin, sus antepasados. En 1085, después que Alfonso VI
      tomara Toledo, sintiendo el peligro cerca le envió sumiso sus felicitaciones
      por la conquista. En 1089 tributaba al Cid diez mil dinares, pero cuando lo
      vio lejos sitiando Valencia, dejó de pagárselos, intentando aliarse con el rey
      de Aragón; éste avisó al Cid, antes de agosto de 1093, quien vino a atacarle y
      le redujo otra vez. Pero cuando Rodrigo ganó Valencia, el señor de Albarracín
      se alió con los $$almorávides%%795$$ y a su lado peleó en la batalla del
      Cuarte, abandonando el sitio en vergonzosa fuga. Murió en la $$Sahla%%$$
      el 18 de mayo de 1103.&lt;/P&gt;
    </descripcion>
  </voz>
```

```
<topic id="98">
        <baseName>
                <baseNameString>Abd al-Malik ibn Hudayl ibn Razin</baseNameString>
                <variant>
                        <variantName>
                                <resourceData>Abd al-Malik ibn Hudayl ibn Razin</resourceData>
                        </variantName>
                </variant>
        </baseName>
        <instanceOf>
                <topicRef xlink:type="simple" xlink:show="replace" xlink:actuate="onRequest"
xlink:href="#38"/>
        </instanceOf>
        <contents>
                <shortdesc>
                        Segundo soberano de la taifa de Albarracín, entre 1045 y 1103, con el
título de Husam Al-Dawla (Sable del Estado). Fue muy criticado por Ibn Hayyan, historiador
contemporáneo suyo, aunque Ibn Ja-qa-n le pondera con ditirambos.
        </shortdesc>
                <body>
                        Segundo soberano de la taifa de Albarracín, entre 1045 y 1103, con el
título de Husam al-Dawla (Sable del Estado). Fue muy criticado por Ibn Hayyan, historiador
contemporáneo suyo, aunque Ibn Ja-qa-n le pondera con ditirambos. Pero su tiempo, al volcar el
siglo xi con la decadencia de las taifas, no le permite ya la altiva independencia de los Banu
Razin, sus antepasados. En 1085, después que Alfonso VI tomara Toledo, sintiendo el peligro
cerca le envió sumiso sus felicitaciones por la conquista. En 1089 tributaba al Cid diez mil
dinares, pero cuando lo vio lejos sitiando Valencia, dejó de pagárselos, intentando aliarse
con el rey de Aragón; éste avisó al Cid, antes de agosto de 1093, quien vino a atacarle y le
redujo otra vez. Pero cuando Rodrigo ganó Valencia, el señor de Albarracín se alió con los
almorávides y a su lado peleó en la batalla del Cuarte, abandonando el sitio en vergonzosa
fuga. Murió en la Sahla el 18 de mayo de 1103.

        </body>
        </contents>
        <date>
                <role>soberano</role>
                <location>Albarracín</location>
                <year>1045</year>
        </date>
        <date>
                <role>soberano</role>
                <location>Albarracín</location>
                <year>1103</year>
        </date>
        <date>
```

Fig. 1. The original GEA voice in XML (input), and the XTM-DITA (output) obtained after applying our algorithm

The tests done have shown that the second-level algorithm produced around three thousand associations and approximately one thousand events as from two hundred words, unlike the first-level algorithm that generated hundreds of associations and dozens of events. This reflects an increase in the order of magnitude of the results. Moreover, the reliability (in other words, the degree of success of both the associations and events related to the text) for the second-level algorithm is 90%, while it is 70% for the first-level algorithm[1].

6 Conclusions

In digital means, integrating the 'context' into the developments performed with thesauri and ontologies, and the solution being a versatile one, is an important aspect to investigate. In this research work, a contextualised, reusable and interoperable solution was firstly planned and then constructed thanks to the use of standards and free software. Since no traditional classification tools were included in the prototype, with the proposed information architecture we managed to: (i) enhance user-friendliness, especially with non-specialised users, (ii) capture the search that this work contemplated in natural language without having to simulate or imitate a performance, and (iii) merge it with other types of external information sources thanks to Topic Maps. Our proposal goes beyond the traditional solutions in the sense that it provides a framework within 'things' can be represented as they are, and it significantly extends and improves the information retrieval process.

Secondly, not only the work done with superimposed information for the online development in this field of application, but also the combination of both subject-centric vision (semantic model) and topic-centered vision (content model), are novel proposals that make the capture, development and reuse of semantics and content strongly relevant, enabling the difficult process of restructuring the volume of information to work properly.

Thirdly, incorporating AI techniques into the algorithm provides coverage of the peculiarities of the Spanish language, such as semantic ambiguity and the wide spectrum of available linguistic formulae to express the same thing.

References

1. "GEA, Gran Enciclopedia Aragonesa," 2009, available at http://www.enciclopedia-aragonesa.com/.
2. ISO/IEC JTC1/SC34, "ISO/IEC 13250-3:2007: Information technology – Topic Maps – Part 3: XML syntax," ISO Intl. Organization for Standardization, 2007.
3. "DITA, Oasis Darwin Information Typing Architecture," 2009, available at http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=dita.

[1] An example trace is available in http://e-archivo.uc3m.es/bitstream/10016/4945/1/Tesis.pdf (pages 324-328)

4. J. Tramullas and P. Garrido, "Constructing web subjects gateways using dublin core (DC), resource description framework (RDF) and topic maps (TM)," *Information Research: an international electronic journal*, vol. 11, no. 2, January 2006, available at http://informationr.net/ir/11-2/paper248.html.
5. L. Maicher and L. M. Garshol, *Subject-centric Computing*, 2008.
6. S. Nakamura, S. Chiba, H. Kaminaga, S. Yokoyama, and Y. Miyadera, "Development of a topic-centered adaptive document management system," in *Computer Sciences and Convergence Information Technology, 2009. ICCIT '09. Fourth International Conference on*, 24-26 2009, pp. 109 –115.
7. ISO/IEC JTC1/SC34, "ISO/IEC 13250:2003 Information technology–SGML Applications–Topic Maps," ISO Intl. Organization for Standardization, 2003.
8. P. Garrido, "El procesamiento automático de documentación textual con información histórica: una aplicación XTM y DITA," Ph.D. dissertation, University of Carlos III de Madrid, 2008, available at http://e-archivo.uc3m.es/dspace/handle/10016/4945.
9. J. Linton and K. Bruski, *Introduction to DITA: a User Guide to the Darwin Information Typing Architecture*. Comtech Services Inc., 2006.
10. L. Griffin, "Millowners and wobblies: an event structure analysis of the everett massacre of 1916," *Annual meeting of the American Sociological Association*, August 2004, available at http://www.allacademic.com/meta/p109966_index.html.
11. L. Griffin and R. Korstad, "Historical inference and event-structure analysis," *International Review of Social History*, vol. 43, 1998.
12. H. Thomas, R. Brecht, B. Markscheffel, S. Bode, and K. Spekowius, "TMchartis — A Tool Set for Designing Multiple Problem-Oriented Visualizations for Topic Maps," in *Scaling Topic Maps: Third International Conference on Topic Maps Research and Applications, TMRA 2007 Leipzig, Germany, October 11-12, 2007 Revised Selected Papers*. Berlin, Heidelberg: Springer-Verlag, October 2008, pp. 36–40.
13. A. Houston and S. Grammar, "Automatic Topic Map Generation from Free Text using Linguistic Templates," in *Scaling Topic Maps: Third International Conference on Topic Maps Research and Applications, TMRA 2007 Leipzig, Germany, October 11-12, 2007 Revised Selected Papers*. Berlin, Heidelberg: Springer-Verlag, October 2008, pp. 237–253.
14. E. Hennum, D. Day, J. Hunt, and D. Schell, "Design patterns for information architecture with dita map domains: defining a type for collections of topic," IBM DeveloperWorks," Technical Library, September 2005, http://www.ibm.com/developerworks/xml/library/x-dita7/.
15. J. Gelb, "DITA and Topic Maps: bringing pieces together," in *Proceedings of the Topic Maps International Conference*, April 2008, http://www.suite-sol.com/downloads/DITA-and-TopicMaps-Bringing-the-Pieces-Together.pdf.
16. P. Garrido, J. Tramullas, F. Martínez, M. Coll, and I. Plaza, "XTM-DITA structure at human-computer interaction service," in *Actas del XI Congreso Internacional Interacción Persona-Ordenador*, June 2008, pp. 407–411.
17. H. Wittenbrik, "The GPS of the information universe: Topic map in an encyclopaedic online information platform," in *Proceedings of XML Europe Conference*, June 2000.
18. "Freeling: an open source suite of language analyzers," 2009, available at http://www.lsi.upc.edu/ nlp/freeling/.