Semantic information stored in an extended denormalized database

Piedad Garrido Picazo[a], Jesús Tramullas Saz[b]
[a] *Universidad de Zaragoza, Informática e Ingeniería de Sistemas , Ciudad Escolar s/n, 44003 Teruel, SPAIN*
[b] *Universidad de Zaragoza, Ciencias de la Documentación, Pedro Cerbuna 12, 50009 Zaragoza, SPAIN*

The research project, which we hereby detail, has been utterly dedicated to explain the birth and evolution of an information repository called XTMdb whose basic principles are bound to integrate complementary tagging languages such as: SKOS; MODS, Dublin Core and/or GILS with the paradigm of topic maps. Once the information processing was completed, we could test it by means of designing an efficient information retrieval process which, at the same time, allows the information resources description to be extended in real time. On the one hand, the main objective lies in obtaining greater expressivity, independence in relation to the tagging language, as well as improved searches since the search can be centred on the topic concept. On the other hand, certain aspects remain outstanding which can lead to interesting discussions such as: the topic maps paradigm can provide added value visual information, it helps make the development of a support system decision easy and soft-computing techniques can solve a considerable amount of problems in relation to the information retrieval process.

Keywords: Topic Maps, MODS, Dublin Core, GILS, SKOS, information retrieval, database, XTM.

1 INTRODUCTION

The idea of storing semantic information in an extended denormalized database arose from the research group GRIO (Gestión de Recursos de información en las Organizaciones[1]) which proposed certain improvements to an already designed prototype thanks to the research project (EA 2003 – 52) granted by the Ministry of Education. This first prototype resulted in a tool called Potnia [8], [9], [10] whose information repository was built using a RDBMS (Relational Database Management System), theTopic Maps paradigm [7] and DC (Dublin Core) [6]. The information retrieval process was based on a binary search, it was limited to a static collection of attributes, and restricted to a combination of these attributes with classic boolean operators such as: AND, OR and NOT.

Therefore, we decided to migrate the data to a new design based on a Native XML Database owing to the Topic Maps paradigm. The Topic Maps paradigm is based on the ISO 13250: 2003 standard, and it was thought that it would give good results and solve binary search problems if it was based on an XTM (XML for Topic Maps) structure [11]. At the same time, it would help extend the information resources description to other tagging languages and would not only use Dublin Core. The research team thought the retrieval process would improve, but some storage, data processing and search problems emerged because of the nature of the XML files. While RDBMSs are known as data-centric databases, a native XML database neither has fields nor stores atomic data, but stores documents instead. Thus, data processing does not improve owing to the hierarchical format in which information is stored. Furthermore, information retrieval cannot be carried out with the Structured Query Language (SQL), and a language called XPath is used instead, but this does not allow for ordered searches, cross joins, nor does it make use of set operators because it is only designed to perform searches in a single document. Finally, it was decided that a conceptual structure had to be designed to integrate the Topic Map paradigm in order to complement the relational one. In other words, using Topic Maps to not only enclose the HTML code in a static way by enriching its structure and providing it with semantics, but also to ensure that these tree structures can be generated automatically according to requirements by making use of the superimposed model-based information [2]. Previous results on this matter exist (see Freese 2003). The result was an extended denormalized database that enabled the information resources description to be extended in real time [12] without having to modify the design.

[1] Information Resources Management in Organizations

2 SOFTWARE REQUIREMENT SPECIFICATION (SRS)

All software development processes need to be accompanied by a software requirement specification where the functionalities and necessities of the system to be developed are reflected. This aspect, which must be clear for all disciplines in all fields [29], does not require any in-depth computer knowledge since it does not include technological aspects and can avoid a lot of problems in the software life cycle within a multidisciplinary research group. The Guide IEEE STD 830:1998 has been followed to analyze and design XTMdb along with its software application. The objective is to have at our disposal at all times information about the final product requirements without having to take implementation details into account.

2.1 Potnia versus XTMdb

Table1. Comparison of objectives pursued for both applications

	Potnia	XTMdb
Accesibility	NO	YES
Usability	NO	YES
Whatever discipline	YES	YES
Whatever language	NO	YES

As Table 1 illustrates, the information repository redesign has been taken advantage of by introducing a series of interface improvements, as well as accessibility and usability criteria, for the purpose of reaching a greater number of users and increasing the quality criteria in aspects such as giving users direct access, bandwidth and interaction, simplicity and consistency, design stability, feedback and dialog, and design for the disabled [14]. We are able to deduce from the information shown in Table 1 that the SRS between the first and the second version will obviously have many points in common. The first version will be enriched by the second version. In any case, these documents have proved fundamental since they justify the reason why a native XML database is not the suitable solution to obtain the aim followed, even though the whole underlying technological environment indicates the contrary.

3 EVOLUTION

Potnia was only capable of labeling information in Dublin Core since it was based on a normalized RDBMS, and the addition of any other type of information resources description tagging language meant that the tables became very large, the search process became much slower and semantics would be lost in the representation. Besides, the GUI (User-Guide Interface) type to be used was mainly textual.

.

Fig. 1. Potnia GUI

The migration of this relational information repository to a native XML database [17, 21] was unexpected, and even though XTM is based on XML, the former is not based on an information exchange.

As one of the project requirements was to work with the information resources description in real time rather than having this description available in a static form, as is the case with the majority of tagging languages, we considered to continue using RDBMS technology [18, 19] since referential integrity proved most useful for us for certain types of searches. Therefore, it was very important to maintain it. Finally, we obtained a relational data model that was not based on obtaining a collection of normalized relations, but one that was based on exploiting the possibility of storing and querying a denormalized logic/relational model. In this way, we obtained an extension of it [1]. The stored information integrity is maintained with this structure extension, and with it, the information resources description was extended to other tagging languages, such as SKOS, MODS or GILS. Combinations among these were also accomplished.

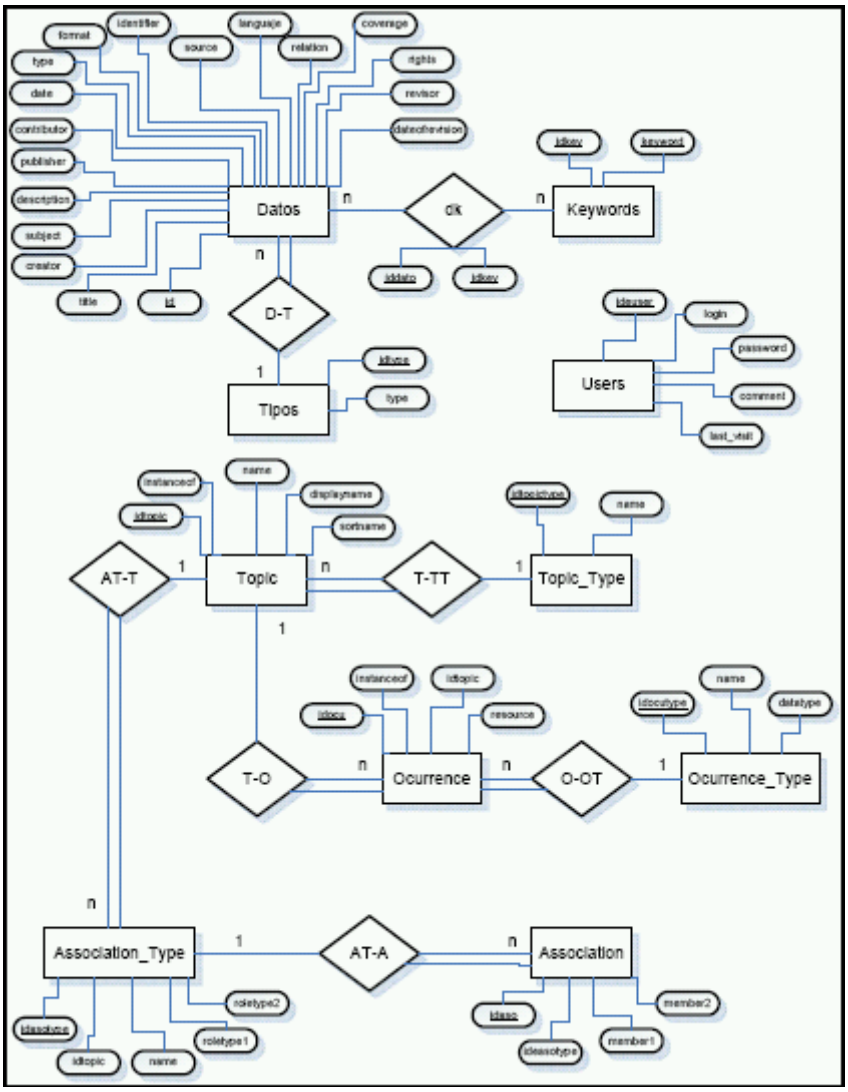

Fig. 2. XTMdb

4 Information Retrieval Improvements

The Topic Maps and the versatility of the topic concept have enabled information to be labeled in any tagging language (see Figure 2). The use of a database that follows the extended denormalized relational model has also allowed search processes to be less restrictive and has enabled the user to see the same resource described in various languages without the problem arising of whether some label has no equivalent label in any other tagging language.

The search process [28] differentiates between a simple search and an advanced search by following

the most common search layout interface presently found in the technology for web-based information resources. Potnia makes use of a RDBMS with textual information so the recovery model needed is boolean, whose main feature is the consideration of relevance that is purely binary. This was achieved by embedding the SQL code within a programming language. With the new search engine version, a Stop words dictionary is used; the search is additive, there are no Boolean operators, and the implicit operator is a logic OR, combined with a weighted indexing.

```
//$consulta="SELECT * FROM datos WHERE title ='" . $titulo . "'";
$consulta="select t.idtopic,tt . name as lenguaje,t.name as etiqueta1,t.sortname as etiqueta2, d . " fror
UNION
select t.idtopic,tt . name as lenguaje,t.name as etiqueta1,t.sortname as etiqueta2, d . " valor from dato:

require('conexion.func');
```

Fig. 3. A brief piece of code

4.2 Advantages & Drawbacks

Since it deals with short contents, the advantages with regard to its performances lie in the fact that indexing is instant. From a statistical point of view, and thanks to the Topic Maps, each content is only indexed once and performance in a multilingual environment is good. Its indexing quality and disk space presents some drawbacks that are worth specifying. Selecting PHP as a programming language could make the word extraction process run slightly slower. This problem could be solved if this process is simplified as much as possible so that indexing times are minimum. As regards disk space, just as with the indexing quality, any emerging problems are not derived from a theoretical idea but from a technological aspect when a specific program or software is used. MySQL was selected as the development platform of XTMdb. This RDBMS is not prepared to store indexing data, therefore this process may prove inconvenient if the site that is available to store the information repository is limited.

As for the stop words file, we should point out that a modification was made later since a problem with acronyms, such as DC, G9, etc., was detected as they were not indexed. Thus, those words with a number of letters >= 2, and which only contained capital letters and numerals, had to be considered as indexed acronyms.

ACKNOWLEDGMENT

The authors wish to thank the IET (Instituto de Estudios Turolenses). This work was supported in part by a grant from this institution thanks to its annual research aids.

CONCLUSIONS & EXPECTANCIES

As conclusions of the present work, we wish to highlight that the experience has been more constructive than expected and certain improvements have been obtained:

- Greater expressivity: not limited to a set of static attributes, no cardinality restrictions and referential integrity was maintained.
- Independence in relation to the tagging language. Information can be shown independently in various formats.
- Improved searches since the search can be centered on the topic concept.

On the other hand, some aspects, which may lead to interesting discussions, have been taken into account for future developments:

- Using paradigms such as Topic Maps, displaying and clustering the results obtained will provide added value visual information on the search results [4].
- Development of a possible support system to perform generic searches in traditional search engines.
- Soft-computing techniques to solve imprecise, uncertain information retrieval processes of textual information [3].

REFERENCES

[1] Balmin, A., Papakonstantinou, P.: Storing and querying XML Data using denormalized relational database. The VLDB Journal — The International Journal on Very Large Data Bases, Volume 14 Issue 1, 30-49 (2005). Springer-Verlag New York, Inc.

[2] Bowers, S., Delcambre, L.: A generic approach for representing model-based superimposed information. In Proceedings of the Workshop on the Semantic Web at ECDL-00; Lisbon, Portugal (2000).

[3] Crestani, F., Passi, G.: Soft computing in information retrieval: techniques and applications. Heidelberg: Physica-Verlag, (2000).

[4] Geroimenko, V. & Chen, C. (Eds.) (2003). Visualizing the semantic web: xml based Internet and information visualization. London: Springer.

[5] ISO 13250:2003. The Dublin Core Metadata Element Set.

[6] ISO 15836:2003. SGML Applications. Topic Maps.

[7] Park, J., (Ed.): XML topic maps. Creating and using topic maps for the web. Boston: Addison-Wesley, (2003).

[8] Tramullas, J., Garrido, P.: Planificación y evaluación de directorios científicos especializados para Internet: su aplicación como instrumentos de docencia en sistemas de enseñanza y aprendizaje virtual. Proyecto EA-2003-52.

[9] Tramullas, J., Garrido, P.: "Constructing Web subject gateways using Dublin Core, RDF and Topic Maps" Information Research, 11(2) paper248 [Available at http://InformationR.net/ir/11-2/paper248.html]

[10] Garrido, P., Tramullas, J.: Potnia: una herramienta para directorios temáticos basada en Dublin Core y Topic Maps. Paper presented at the 7th. Congreso ISKO España, Barcelona, July 2005.

[11] XTM TopicMaps.Org (2001). (XML Topic Maps Specification (XTM) 1.0. XTM TopicMaps.Org. Retrieved 2 March, 2003 from http://www.topicmaps.org/xtm/index.html

[12] IEEE Guide 830-1998. Recommended Practice for Software Requirements Specifications

[13] Garshol, L.M. (2002): What are Topic Maps. XML.com.
Disponible en: http://www.xml.com/pub/a/2002/09/11/topicmaps.html

[14] Lynch, P.; Horton, S.: Web Style Guide (2nd Edition). Yale University Press, 2002.

[15] Hofmann, T. (1999). Probabilistic Topic Maps: navigating through large text collections. In David J. Hand, Joost N. Kok, Michael R. Berthold (Eds.) Advances in intelligent data analysis: third international symposium, IDA-99, Amsterdam, The Netherlands, August 1999. (pp. 161-172). Berlin: Springer.

[16] Hongjun, L., Xu Yu, J., Wang, G., Zheng, S., Jiang, H., Yu, G., Zhou, A.: What makes the differences: benchmarking XML database implementations ACM Transactions on Internet Technology (TOIT), Volume 5 Issue (2005

[17] Nicola, M., Van der Linden, B.: Native XML Support in DB2 Universal Database. Industrial Session: XML Support in Relational System (2005). Proceedings of the 31st international conference on Very large data bases VLDB '05. VLDB Endowment

[18] Florescu, D.; Kossmann, D.: Storing and Querying XML Data using an RDBMS. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 1999.

[19] Jiang, H.; Lu, H.; Wang, W.; Xu, J.: Xparent: an efficient RDBMS-Based XML Database System. ICDE' 02. Proceedings of the 18th International Conference on Data Engineering, 2002, IEEE.

[20] Balmin, A., Papakonstantinou, P.: Storing and querying XML Data using denormalized relational database. The VLDB Journal — The International Journal on Very Large Data Bases, Volume 14 Issue 1.Springer-Verlag New York, Inc.

[21] Thuraisingham, B. (2002). *XML databases and the semantic web*. Boca Raton, FL: CRC Press.

[22] Rath, H.H. (2001). Semantic resource exploitation with Topic Maps. In Henning Lobin, (Ed.) *Sprach- und Texttechnologie in digitalen Medien. Frühjahrstagung der Gesellschaft für linguistische Datenverarbeitung (GLDV), Justus-Liebig-Universität Giessen, 28.-30.03.2001*, (pp. 3 -15). Norderstedt, Germany: Books on Demand.

[23] International Organization for Standardization and International Electrotechnical Commission. Joint Technical Committee 1. (2002) ISO/IEC 13250. Topic Maps. Information technology. Document description and processing languages. Retrieved 22 September, 2004 from http://www.y12.doe.gov/sgml/sc34/document/0322_files/iso13250-2nd-ed-v2.pdf

[24] Hofmann, T. (1999). Probabilistic Topic Maps: navigating through large text collections. In David J. Hand, Joost N. Kok, Michael R. Berthold (Eds.) *Advances in intelligent data analysis: third international symposium, IDA-99, Amsterdam, The Netherlands, August 1999.* (pp. 161-172). Berlin: Springer.

[25] Garshol, L.M. (2004). Metadata? Thesauri? Taxonomies? Topic Maps! *Journal of Information Science*, **30**(4), 378-391.

[26] Freese, E. (2003). Topic maps and RDF. In Park, J., (Ed.) *XML topic maps. Creating and using topic maps for the web*. (pp. 283-325) Boston, MA: Addison-Wesley.

[27] Chakrabarti, S. (2002). *Mining the Web: analysis of hypertext and semi-structured data.* Boston, MA: Morgan Kaufmann

[28] Berry, M.E. & Browne, M. (1999). Understanding search engines: mathematical modeling and text retrieval. Software, environments, tools. Philadelphia, PA: Society for Industrial & Applied Mathematics.

[29] Tramullas, J., Garrido, P. Software libre para servicios de información digital. Prentice Hall, 2006.