Vol. 11 No. 2, January 2006

Constructing Web subject gateways using Dublin Core, the Resource Description Framework and Topic Maps

Jesús Tramullas, Department of Librarianship and Information Sciences and
Piedad Garrido, Department of Computer Science and Systems Engineering
Zaragoza University, Zaragoza, Spain

change font

Abstract

Introduction. Specialised subject gateways have become an essential tool for locating and accessing digital information resources, with the added value of organisation and previous evaluation catering for the needs of the varying communities using these. Within the framework of a research project on the subject, a software tool has been developed that enables subject gateways to be developed and managed.
Method. General guidelines for the work were established which set out the main principles for the technical aspects of the application, on one hand, and on aspects of the treatment and management of information, on the other. All this has been integrated into a prototype model for developing software tools.
Analysis. The needs analysis established the conditions to be fulfilled by the application. A detailed study of the available options for the treatment of information on metadata proved that the best option was to use the Dublin Core, and that the metadata set should be included, in turn, in RDF tags, or in tags based on XML.
Results. The project has resulted in the development of two versions of an application called Potnia (versions 1 and 2), which fulfil the requirements set out in the main principles, and which have been tested by users in real application environments.
Conclusion. The tagging layout found to be the best, and the one used by the writers, is based on integrating the Dublin Core metadata set within the Topic Maps paradigm, formatted in XTM.

Introduction

The tools for retrieving and accessing information available on the Internet to answer users' needs take three main forms, namely search engines, subject gateways and vertical portals. Among these, the subject gateways are gaining in popularity and importance as a source of digital information that has been chosen and assessed, and which is being provided with added value services, as described by Navarro & Tramullas (2005). This type of information service, by which is understood the structures proposed by Colomb (2002), is based on sets of documents which are browsed by the user with the aid of auxiliary information structures. Koch proposed the following definition, which is widely accepted:

> Subject gateways are Internet services which support systematic resource discovery. They provide links to resources (documents, objects, sites or services), predominantly accessible via the Internet. The service is based on resource description. Browsing access to the resources via a subject structure is an important feature. (Koch 2000: 24-25)

Therefore, they are areas of information designed for a user to discover relevant information for a given need (Pitschmann 2001). Although the engines and portals have received a great deal of attention in the bibliography, this is not the case with specialised gateways, in spite of their importance as a tool for searching for and retrieving information (Robinson & Bawden 1999; Bawden & Robinson 2002). Only in Britain, thanks to the excellent development of the Resource Description Network (RDN), can one talk of high quality subject gateways.

The undertaking of a research project on subject gateways in the libraries of Spanish universities by a research group, to which the writers of this article belong, had among its objectives the creation of a software tool which, with minimum technical requirements and applying the basic principles of digital information management, would enable the rapid start-up of a specialised subject gateway A set of guidelines was established for the purpose, which would have to be adhered to both during development and for the final tool:

1. It must be based on free software, and therefore, would also be free.
2. It must use standard data base technology to save and retrieve information.
3. It must comply with XML standards.
4. It must use standard resource description for information.

Bearing in mind the above, and the need for the tool to be able to evolve in line with future developments of information treatment in XML environments, in addition to new techniques for displaying large amounts of information, a free software tool was designed and implemented, called *Potnia* (Garrido & Tramullas 2005), distributed under a Mozilla Public License.

Description and treatment of data on digital information resources

In line with the basic concepts mentioned above, the structure of the resource description of information conforms to standard ISO 15836 Dublin Core. However, this structure for description based on metadata reaches its full potential when it is integrated into XML coding schemes. For the digital information resource description, Dublin Core was embedded within the Resource Description Framework (RDF), with the aim of boosting its description capacity and its (future) use in the semantic web framework. Projects following this structure can be found in the literature , for example Berry & Browne (1999), Chakrabarti (2002), Firestone (2003) and Michalak (2005).

However, one of the objectives set for the *Potnia* tool is to develop interfaces based on visual metaphors, which will complement the usual presentation of a list of replies. The most suitable paradigm for this is to combine DC/RDF with Topic Maps (Lacher & Decker 2001), a paradigm which is contained within standard ISO 13250 (2002; Park 2003). Topic Maps have been formatted as XTM (2001) through the use of XML notation. Once the connection point has been introduced and studied in depth, it would be advisable to integrate the Dublin Core and Topic Maps. As demonstrated by Bowers (2000) in a study of superimposed information based on models, there are many similarities between the structural layers found in XML, RDF and Topic Maps. Therefore, it is perfectly possible, and right, to represent the majority of RDF structures through the use of syntax for topic maps, and vice-versa (Freese 2003). However, in the latter direction (representing topic maps with RDF structures) part of the semantics is lost. Since the main objective of the application is provide a greater degree of precision in the results of searches, such a loss of semantics is detrimental, and therefore it was decided to use the structure and syntax of topic maps, since these are a more modern, flexible and abstract paradigm.

In addition, topic maps have been repeatedly put forward for other projects as an extremely suitable tool for classifying and organising information. Classification tools, such as taxonomies, thesauri, ontology or faceted metadata can be integrated into topic maps, as demonstrated by Garshol (2004). In the same way, they are also more complex and allow for the development of richer, more complex information structures than the widely-used conceptual maps, which can also be integrated into XTM (Garrido & Tramullas 2004). The many examples of proprietary software tools for displaying information through topic maps, and the fact that there are open source development packages, make it possible to say that the paradigm of topic maps is the information tagging environment that offers the highest number of opportunities for developing metadata-rich digital information products, presented through graphic interfaces.

First version: Potnia 1.0

The needs analysis before the development of *Potnia* showed that it would be of most use on a personal or departmental level, answering the needs of special interest groups or communities. The prototype for development had very specific characteristics and was much more manageable than if a structure of specialised subject gateways had been used, such as those used by general gateways like Yahoo! or dmoz. The first objective of the design was to have the simplest possible search system, which would reduce the problem of over-complicated interfaces for novice users, and would reply faster with information in greater detail, and be more agile. The second design objective was for the users themselves from a specific discipline to gradually fill the data base supporting the information resource. The initial architecture for the application planned for this project is shown in the figure below:

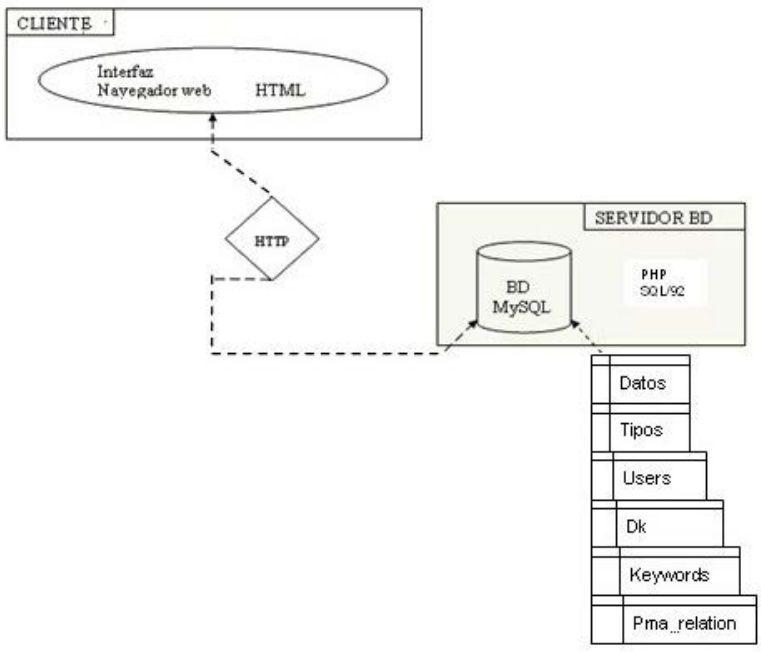

Figure 1: Architecture for the application

The following guidelines were set for working on the development of the first version:

1. Non-declarative programming must be used.
2. It must be supported by a Relational Database Management System.

3. Advanced information display techniques must not be used.
4. The information resource description structure must comply with the ISO 15836 (DCMI 2003) standard.

The technology chosen to implement this was:

1. PHP Script Language (4.3.3RC1) and XHTML 1.0 for Graphical User Interface
2. MySQL Database ServerVersión 4.0.13
3. SQL/92 (Structured Query Language, for the design of the search engine).

Eventually, the final product was given the name *Potnia*. A test application can be found at http://imhotep.unizar.es/potnia. The application is distributed as open source on the following servers:

- http://potnia.sourceforge.net/
- http://freshmeat.net/projects/potnia

Database design

The resulting relational scheme consists of six tables: data, types, users, data-keyword relations (dk), keywords and pma_relation (Figure 1 above). This last relation proves essential when working with this version of MySQL in order to enable the database management system to interpret the relations with many-to-many cardinality. The users table remains apart from the scheme and not related to the other tables, as its only task is to centre on administering the various types of users who access the application. Tables for managing the information from the resources are the types, dk, keywords and data tables, which bear the weight of the application. When these tables are implemented, the Dublín Core Metadata Set (simple) can be used in a database relational structure.

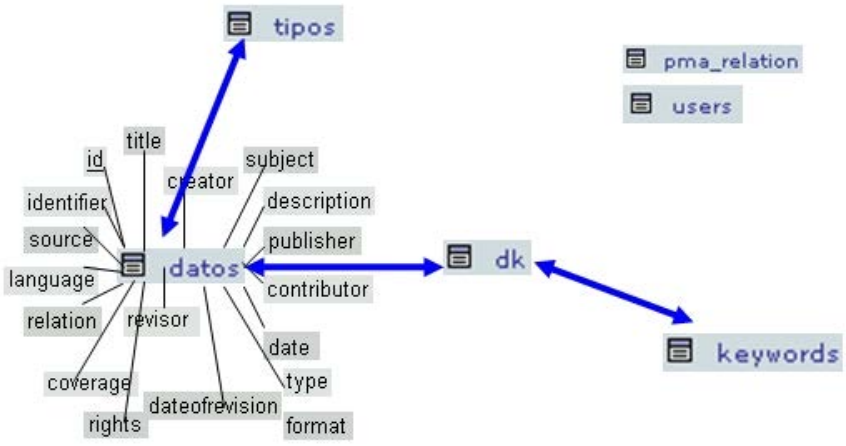

Figure 2: Relations among tables

Search code

The search process differentiates between a simple search and an advanced search, following the most common search layout interface found at present in the technology for web-based information resource gateways. As a relational database management system with textual information is being used, the recovery model needed is boolean, whose main feature is the consideration of relevance that is purely binary. This was achieved by embedding SQL code within PHP programming, which, in turn, included Javascripts and HTML code. The following figures show examples of SQL embedded code needed to execute searches:

```
if ($tema){
$consultatotal="SELECT * FROM datos WHERE subject='".$tema."'";
}
else{
$consultal="SELECT d.title, d.creator, d.description, d.identifier, d.id, dk.idkey, dk.iddato, k.idkey, k.keyword
            FROM datos d, dk, keywords k WHERE (";

 switch ($campol) {
     case "título":
         $consultal2="(title LIKE '%". $elemabuscarl . "%')";
         break;
     case "tema":
          $consultal2="(subject LIKE '%". $elemabuscarl . "%')";
         break;
     case "descripción":
          $consultal2="(description LIKE '%". $elemabuscarl . "%')";
         break;
     case "palabras clave":
          $consultal2="((k.keyword LIKE '%" . $elemabuscarl . "%') AND (d.id=dk.iddato) AND (dk.idkey=k.idkey))";
         break;
```

Figure 3. Treatment of character strings

Second version: Potnia 2.0

As shown below, some decisions taken in implementation, the technology used and suggestions from persons or groups who have used *Potnia*, have motivated a series of decisions to be taken that have gradually affected the architecture of the application, and caused the appearance of several versions of the tool throughout its lifespan. The advantage of this idea lay in the wish to re-design *Potnia*. Once the first version was in operation, checked for any faults and accepted by the various groups using it, it was suggested that a second version should be developed. The first version of *Potnia* had been evaluated partly from suggestions and comments sent in by e-mail from various groups and individuals who had downloaded the application from Sourceforge or Freshmeat; and partly because the team working on the development were well aware that some aspects had not been included in the first version. Important among these lacks was the need to incorporate sessions, to improve the perception and running of searches by the end user, and the extension of the Dublin Core metadata set, in order to be able to use the qualified set.

Once the usefulness of the project had been evaluated, re-designing was approached by incorporating a set of improvements to different design aspects centred on the user, such as information retrieval and security. In order to fulfil these objectives, several markup languages were considered for prior representation of the information. Among these were OWL (Ontology Web Language), XML (eXtensible Markup Language) and XTM (XML Topic Maps), all highly suitable markup languages for the treatment of metadata, and which offered powerful new options to combat the deficiencies and limitations in the usual text access to digital information resources. Integration of OWL into XTM has been proven (Vatant 2004). These languages can be stored in some of the different types of information repository found on the market, such as: SGBDR, BDOO, BD natives in XML or hybrid systems. This approach can provide higher performance treatment to meta-information saved in the database and formatted to ISO 15386 standard requirements, and, obviously, to improve the display interface for the information in the not too distant future. This main objective will centre on replacing the simple interface of a text list of static replies and thus be able to use a language which is capable of representing the semantics inherent in the inter-relations targeted among the objectives (Hofmann 1999). Lastly, from the point of view of security, the application was to be improved by including sessions, a modification that will directly affect the design and programming of the user's graphic interface.

This was achieved by taking a new look at the initial design of the whole project (see work hypothesis), and incorporating XTM Topic Maps (ISO 13250) and a Hybrid Database (SGBR working with an XML manager) into its design:

Relating to the first point, integrating XTM to the resource description for digital information studied should mean that a DTD had had to be used in the first place to enable the topic map paradigms to be read by a computer (see figure 5), followed by the process of describing the information already stored in the MySQL database with the corresponding metadata, in compliance with the ISO 13250 standard. Thus, due to the flexibility of XTM and to the fact that the descriptive information of the resources has been treated previously according to ISO 15386, a simple process is used to enable an information exchange with other tools using mark-up languages such as RDF, making later development possible in the framework of semantic web technology (Beckett 2002).

```
<!DOCTYPE topicMap PUBLIC "-//TopicMaps.Org//DTD XML Topic Map (XTM) 1.0//EN"
                          "http://www.topicmaps.org/xtm/1.0/xtm1.dtd">
<topicMap xmlns="http://www.topicmaps.org/xtm/1.0/"
                  xmlns:xlink="http://www.w3.org/1999/xlink">
<html>
<head>
<title>Documento sin t&iacute;tulo</title>
<meta http-equiv="Content-Type" content="text/html; charset=iso-8859-1">
<script language="JavaScript" type="text/JavaScript">
```

Figure 4: Inclusion of the DTD in the application

```
<!--Código xtm -->
      <record>
      <field>
       <ul><li><a href="mostrar_datos_por_tema.php?titulo=<?php print($result_campo_tema);?>">
       <key><font color="#0000A0"><?php print(" <strong>".$result_campo_tema."</strong><br><br>");?>
       </font></li><ul></key></a>
       <li><strong>Título:</strong><a href="mostrar_datos_por_titulo.php?titulo=<?php print($result_campo_titulo);?>">
       <key><font color="#0000A0"><?php print(" ".$result_campo_titulo."<br><br>");?>
       </font></key></li></a>
       <key><li><strong>Descripción:</strong><?php print(" ".$result_campo_descripcion."<br><br>");?></li></ul></ul></key>
      </field>
      </record>
<!--Fin código xtm -->
<?php
   }//Del for
 }else{
```

Figure 5: Example of a code with integrated XTM

As for the second point, which involves the changeover of a purely relational database manager to a hybrid database manager (Thuraisingham 2002), it must be emphasised that the process of introducing a digital information resource within the application remains the same as in the first version, as far as the end user is concerned. However, in the internal architecture, the user opens with a validation session as the authorised user of the system, and proceeds to include the description of a resource, and, after having passed through an error control filter, passes on the process of instantiation. In this process, part of the information is instantiated in the tablespaces data and keywords, created in MySQL for the purpose, and a parallel instantiation process appears, which stores the metadata in a structure called a description file. This description file saves the data in XML in a secondary memory, thus comprising the XML manager which works with the relational database at all times. As a complement, part of the description file is stored in a hash table, in the main memory, in order to make more rapid and efficient searches. This change in the architecture has, in turn, forced a change in the architecture of the search system, with the search process being structured as shown in Figure 6.

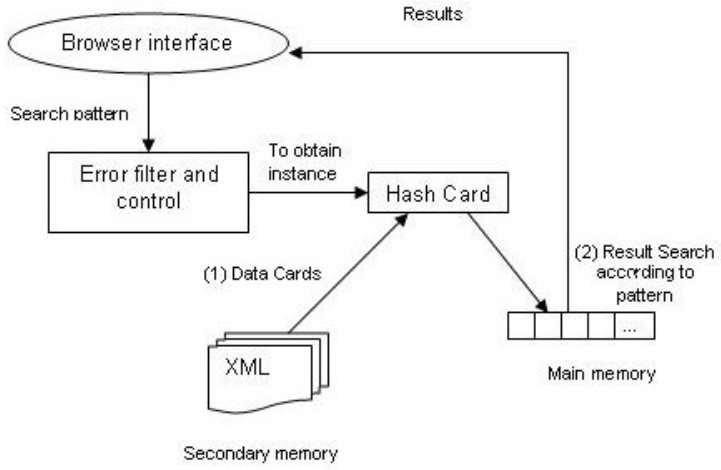

Figure 6: Search process on the hybrid manager

Conclusion and future developments

The development of tools for information management needs to have an inter-disciplinary approach that will ensure the quality of the resulting products. The project described here was aimed at the needs for information management of the end users, and has determined the main features on the *Potnia* application. Technical problems were solved by using the topic maps paradigm to provide the information with semantics (Rath 2001). This enables the search processes to be refined and adjusted, as they establish points of contact between the keywords which were ignored by the traditional process of treatment and retrieval of information used initially. In addition, the greater advantages obtained by incorporating a hybrid manager into the application has improved the speed of reply on the search engine

However, one of the drawbacks that may arise from this new version is the use of an XML manager, since this does not allow for the most suitable treatment for textual information in the case of more complex searches. For this reason, the use of soft-computing techniques is being analysed to solve problems arising from this type of retrieval.

The future development of *Potnia* will take the following aspects into account:

- Display: using paradigms such as self-organizing maps, conceptual maps or topic maps, the display and clustering of the results obtained will provide added value visual information on the search results (Geroimenko & Chen 2003).
- Development of a possible support system to make generic searches in traditional search engines, or other sources of digital information, to feed the Potnia gateway.
- Soft-computing techniques for information retrieval, since the particular features of textual information conform to an imprecise, uncertain information profile that is difficult to categorise; this makes soft-computing techniques into a very useful paradigm to solve the problem (Crestani & Passi 2001).

References

- Bawden, D. & Robinson, L. (2002). Internet subject gateways revisited. *International Journal of Information Management*, **22**(2), 157-162.
- Beckett, D (2002). *Connecting XML, RDF and Web technologies for representing knowledge on the Semantic Web*. Paper presented at XML Europe Conference 2002, Barcelona, May 2002. Retrieved 2 May, 2004 from http://www.idealliance.org/papers/xmle02/dx_xmle02/papers/03-05-07/03-05-07.html
- Berry, M.E. & Browne, M. (1999). Understanding search engines: mathematical modeling and text retrieval. Software, environments, tools. Philadelphia, PA: Society for Industrial & Applied Mathematics.
- Bowers, S. (2000). *A generic approach for representing model-based superimposed information.* Beaverton, OR: Oregon Health and Science University. Retrieved 17 December, 2005 from http://www.cse.ogi.edu/tech-reports/2000/00-008.pdf&ei=MlWkQ4rAJKKQiALG56ydBw&sig2=qycTahRJnfO2HwaZ5mGX3g
- Caussanel J., Cahier J-P., Zacklad M. & Charlet, J. (2002). *Les Topic Maps sont-ils un bon candidat pour l'ingénierie du Web Sémantique?* Paper presented at Actes de la conférence Ingénierie des Connaissances, Rouen, 2002. Retrieved 1 April, 2005 from http://tech-web-n2.utt.fr/ssw/cahier/docs/IC2002.pdf
- Chakrabarti, S. (2002). *Mining the Web: analysis of hypertext and semi-structured data.* Boston, MA: Morgan Kaufmann.
- Colomb, R.M. (2002). *Information spaces: the architecture of cyberspace*. London: Springer.
- Crestani, F. & Passi, G. (Eds.) (2000). *Soft computing in information retrieval: techniques and applications*. Heidelberg: Physica-Verlag
- Firestone, J.M. (2003). *Enterprise information portals and knowledge management.* Oxford: Butterworth-Heinemann.
- 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.
- Garrido, P. & Tramullas, J. (2004). Convergence of topic maps and concept maps: prototype foundations. In *Proceedings of the IADIS International Conference. WWW/Internet 2004, Madrid*, vol. 2, (pp. 1105-1108). Lisbon: International Association for the Development of the Information Society.
- Garrido, P. & Tramullas, J. (2004). Topic maps: an alternative or a complement to concept maps? *Concept Maps: Theory, Methodology, Technology. First International Conference on Concept Mapping*, Pamplona, vol. 2, (pp. 185-188). Pamplona, Spain: Universidad Pública de Navarra.
- Garrido, P. & Tramullas, J. (2005). *Potnia: una herramienta para directorios temáticos basada en Dublin Core y Topic Maps.* Paper presented at the 7th. Congreso ISKO España, Barcelona, July 2005.
- Garshol, L.M. (2004). Metadata? Thesauri? Taxonomies? Topic Maps! *Journal of Information Science*, **30**(4), 378-391.
- Geroimenko, V. & Chen, C. (Eds.) (2003). *Visualizing the semantic web: xml based internet and information visualization*. London: Springer.
- 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.
- 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
- Koch, T. (2000). Quality-controlled subject gateways: definitions, typologies, empirical overview. *Online Information Review*, 24(1), 24-34.
- Lacher, M.S. & Decker, S (2001). On the integration of Topic Maps and RDF data. In *Aggregated Proceedings for the Extreme Markup Languages® Conferences (2001-2005)*. Rockville, MD: Mulberry Technologies Inc. Retrieved 17 December, 2005 from http://www.mulberrytech.com/Extreme/Proceedings/html/2001/Lacher01/EML2001Lacher01.html
- Michalak, S.C. (Ed.). (2005). *Portals and libraries*. Binghamton, NY: Haworth Press.
- Navarro, D. & Tramullas, J. (2005). Directorios temáticos especializados: definición, características y perspectivas de desarrollo. *Revista Española de Documentación Científica*, **28**(1), 49-61.
- Park, J., (Ed.) (2003). *XML topic maps. Creating and using topic maps for the web*. Boston: Addison-Wesley.
- Pitschmann, L. A. (2001). *Building sustainable collections of free Web third-party resources*. Washington, DC: Digital Library Federation.
- 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.
- Robinson, L. & Bawden, D. (1999). Internet subject gateways. *International Journal of Information Management*, **19**(6), 511-522.
- Thuraisingham, B. (2002). *XML databases and the semantic web*. Boca Raton, FL: CRC Press.
- Vatant, B. (2004). *Ontology-driven topic maps*. Paper presented at XML Europe 2004, Amsterdam, May 2004. Retrieved 2 May, 2005 http://www.idealliance.org/papers/dx_xmle04/papers/03-03-03/03-03-03.pdf
- 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