Developing a simple production and dissemination model for a complex DL-news service using XML, XSLT and TEI♣

© Alejandro Bia and Jaime Gómez

Departamento de Lenguajes y Sistemas Informáticos
and Miguel de Cervantes Digital Library
University of Alicante
{abia, jgomez}@dlsi.ua.es

Abstract

Electronic publishing makes it possible to reach every corner of the world and opens up new research and communication paths.

In this article we describe the production model and implementation of an electronic news service for a DL, that manages altogether five different DL-newsletters plus a monthly journal, each one of them delivered in several optional output formats, and all managed and produced based on XML-TEI and XSLT technology. Articles and news may appear in different newsletters and/or the journal, and they have different periodicity (some are published quarterly and some monthly). The news and articles come from different sources, which generally coincide with departments or units of the DL. A general editor reviews the articles or news, decides where must they appear, and also manages the distribution lists.

The newsletters are not preserved, but the most relevant news and articles are published in the monthly journal called Dulcinea, which is published through Internet and preserved for historical purposes. Its contents are indexed by means of an XML indexer that allows advanced XML searches to be performed.

In addition to describing the production model we will discuss our current efforts to improve the dissemination model used for these newsletters. We will also talk about the spirit and objectives of the MCDL's interactive services in general.

1 Introduction

The Miguel de Cervantes Digital Library is the biggest electronic publishing project in Spain, and perhaps the biggest digital library of Spanish texts on the Internet, currently with more than 10000 entries in its catalogue [1]. It produces an average of 150 XML-TEI digital texts per month, most of which are Spanish classics from the 12th century up to these days, comprising a wide variety of subjects and styles such as poetry, narrative, drama, history, geography, law, etc. These texts are used both by the casual reader and by specialized researchers that take advantage of the power of complex-structural-markup.

Before going any further, it is essential to explain the four roles of the MCDL:

1. ***digital library:*** The MCDL is a systematically organized collection of all kinds of digital materials which are freely accessible through our Web site .
2. ***digital publisher:*** The MCDL produces its own digital resources using the latest technologies.
3. ***research centre:*** The MCDL is also focused on research and creation of advanced tools and methods to be applied to humanities research and to the development of digital libraries.
4. ***communication channel:*** The MCDL promotes the diffusion of the Hispanic culture and academic work.

The MCDL synthesizes these four concepts: It is a collection of digital materials; a digital publisher; a place where research takes place; and a communication instrument. Its main purpose, according to its foundational objectives, consists of promoting and spreading the Hispanic cultural values by the application of the most modern technologies to relevant literature, science and Hispano-American works.

2 Objectives

"It is interesting to think on the revolution that the Internet has meant for the publishing industry, especially with the academic world urging for the latest ideas and innovations. This was a giant achievement because information is now readily available in a very short time (Ph.D. theses are available on the net as soon as dissertations take place at universities) promoting as a result a rich interaction among scholars [2]."

Proceedings of the 5th Russian Conference on Digital Libraries RCDL2003, St.-Petersburg, Russia, 2003

One of the goals of the MCDL is to act as a communication channel for the academic community. In this sense we have implemented a number of communication services, and we try to maintain a permanent communication with our readers. These communication services can be classified into two categories: interactive and non-interactive. Amongst the interactive communication services there is a "digital librarian" service[1], which is an online help-desk to answer questions sent by users, trying to compensate for the lack of personalized human attention that traditional libraries provide.

We would also like to make a special reference to another of our interactive services, our discussion forums[2], which have resulted in an increasing number of visits to our Website. Forums provide researchers with a continuous information exchange flow. Section forums gather opinions, comments, questions and any other contribution from readers regarding authors and their work. Forum comments sometimes give way to replies from the authors, opening then an asynchronous debate with fruitful results for all.

A complementary service to discussion forums is the online chats service[3], where discussions are also held about specific topics requested by researchers. But online chats, unlike discussion forums, are held in real time, so they must be organized according to a prefixed time schedule. An online chat is a virtual meeting point for researchers and users where discussions are held about topics like literature, history, education, philosophy, literary creation, etc. To keep the scientific accuracy, once we identify a popular topic, we contact academic specialists who will develop their views online during the chat, with the help of a moderator. The text lines of these online chats are permanently published later on.

Finally, to complete our wide variety of communication services, we have our non-interactive news services, which serve to inform of the release of new digital books, and of events of interest for certain groups of subscribers, like cultural or academic events, scheduled online chats, etc. These news services are comprised of online daily briefings[4] of the library's activities, quarterly or monthly electronic newsletters from several thematic sections of our library, the Dulcinea monthly journal (which is a compendium of the newsletters plus some articles), and digital versions of press releases related to our DL[5]. In addition to these news services directly related to our DL, there is a press catalog with links to all Iberoamerican Web-accessible newspapers[6] and an electronic search engine that is updated daily from these news sources[7].

All these different communication services complement each other in several ways, like online chats that need to be announced by news bulletins and daily briefings.

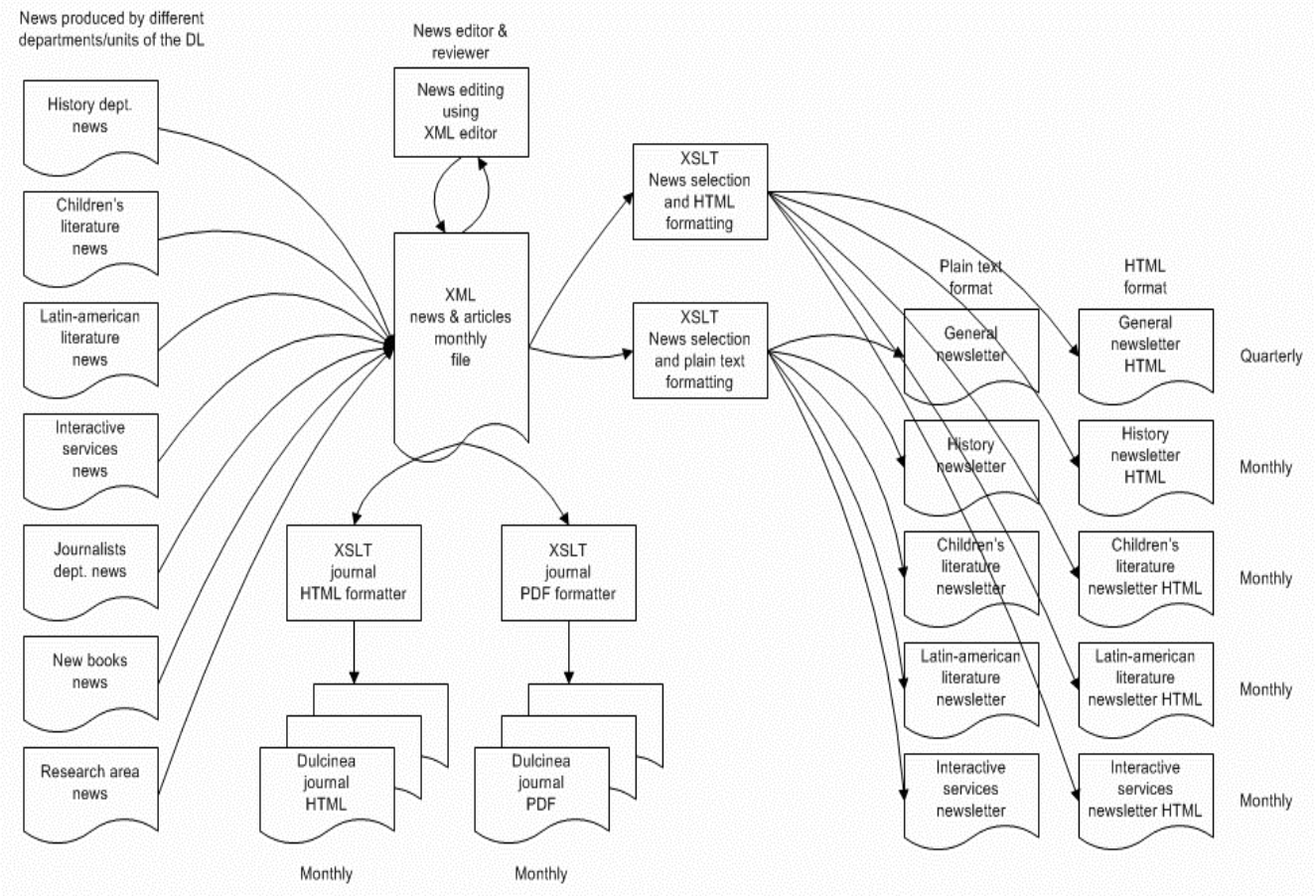

Fig. 1: Newsletter and journal production workflow.

Five newsletters plus a monthly journal

We will now describe the production model and implementation of the electronic news service of the MCDL, that manages altogether five different DL-newsletters plus a monthly journal, each of which is delivered in several optional output formats, and all managed and produced based on XML-TEI and XSLT technology [3].

Requirements

From the requirements analysis the following needs were detected: Articles and news may appear in different newsletters and/or the monthly journal at the same time according to editorial decisions, but this overlapping has a monthly periodicity (no news of one month appear on another bulletin or journal of another month). Newsletters and the journal have different periodicities (some are published quarterly and some monthly and all in different dates). The monthly journal is issued at the end of the month as a compilation of the monthly news plus some added articles.
The news and articles come from different sources, which generally coincide with different departments or units of the DL.

The editor coordinates the production of the different news sources. He reviews the articles or news, corrects grammatical or stile errors and decides in which publication should each note appear. He also manages the distribution lists and sends the emails to subscribers.

The newsletters are not preserved, but the most relevant news and articles are published in a monthly journal called Dulcinea, which is published through Internet and preserved for historical purposes. Its contents are indexed by means of an XML indexer that allows advanced XML searches to be performed. Just like any other of our digital books, the Dulcinea journal is stored permanently, and it can be accessed by means of the MCDL general catalogue.

A simple solution

First, we considered the possibility of developing a management system based on a database. Then, while reviewing the requirements, we realized that a much simpler solution using an XML editor, XML-TEI encoding and XSLT transformations was possible (see figure 1).

A database based system would be better for editing and maintenance of news, but news have no maintenance. Once the are published, there is no more editing of the news. However, they can be searched, but this service can be provided by an XML searcher without the need of an DBMS (database management system).

With this simple solution, no system programming was needed, only the XSLT transforms needed to be built. In addition, it takes advantage of the same TEI XML markup scheme and the same processing technology we use to produce our digital books. The differences are mainly the overlapping of news and the interrelated times of publication.

How it works

News are submitted by the contributing departments of our library, first by email, and now by means of an intranet Web form. They are submitted and temporarily stored in XML format for editorial processing. The editor picks up the news and inserts them into a monthly file, joining them with journal articles. At different deadlines during the month, an XSLT script is applied to extract each partial newsletter simultaneously in different formats (plain text or HTML). At the end of the month, the Dulcinea journal is issued by the same XSLT processing system. This journal includes almost all the news of the different newsletters (except those that became outdated, like online-chat announcements that have expired), and also some articles that only appear in the journal. The selection of where to include each piece of news is done by different XML attributes used as flags for this purpose.

The markup scheme

We used XSL-TEI to markup each piece of news. TEI is also the markup scheme we apply to all of the digital books we produce . In the case of the newsletters and the journal we use a subset of the tagset we currently use for books, so we developed a small DTD for this purpose. This DTD was obtained by automatic simplification from our cervantes.dtd using a DTD simplification tool of our own, called DTDprune [4].

Processing environment tools

For the development and testing of the XSL transformations we used CookTop and XMLSpy, both editors with interesting tools and capabilities for XML design. However, for news editing we use XMetaL, an editor that provides the nicest and more user friendly interface available for text editing. For XSLT processing we currently use the Saxon XSLT processor.

XSLT transformations

We had to develop several XSL [5] transformation scripts to produce the different output formats required. Newsletters are generated both in plain text (for traditional mail readers) and in HTML for those who prefer a richer format.

The journal, on the other hand, is given a format similar to that of our digital books: several HTML pages, one for each section, and a hypertext table of contents at the beginning. Navigation buttons are added to allow the user to navigate the journal also in a cyclic way, jumping from one chapter to the next, or to the previous (see figure 2).

How it works

Once the monthly news file is complete and supervised, we enter the automatic phase of generation

of different output formats, as described by Arms in chapter 9, pg. 163-166 of his book on Digital Libraries [6].

The final output in HTML format is obtained from a double transformation of the XML-TEI file [7] (see figure 3). First an XSL transformation processes the XML monthly file to generate a single HTML file (with special formatting marks embedded). Then a parsing program of our own design called MakeBook transforms that file into a digital book, generating a file for the table of contents, and a file for each section of the journal. Notes are extracted from the main file and placed in small external files, leaving hyperlinks to these files in their place. MakeBook was also thought to add page headers and footers for each section, that include buttons to implement a navigation pattern called "indexed guided tour" [10]. This navigation pattern allows both the presence of a central index or table of contents and also buttons to move back and forth the different sections of the journal (hence the metaphor "guided tour"). Buttons to jump up and down article headings with a mouse click are also provided. The result is an electronic journal that is nothing but a set of web pages interconnected with a ring-star topology: a bidirectional ring of connections to navigate the sections, plus a central table-of-contents page with bidirectional connections to each section (see figure 2).

A set of templates is used for formatting and providing navigation functionality to the generated HTML files. The purpose of using templates is to give a uniform appearance to all the monthly journals, as well as to facilitate the maintenance, so that format changes can be applied easily and evenly to the whole set (e.g., section headers and footers, background colors and textures and navigation buttons can be changed through these templates).

To avoid having to reprocess the whole collection of journals every time a template is changed, this templates are applied on-the-fly by the Web-server by means of Java Servlets. In this way, rendering changes don't force the regeneration of the whole lot but they instantly take effect on the next journal to be served.

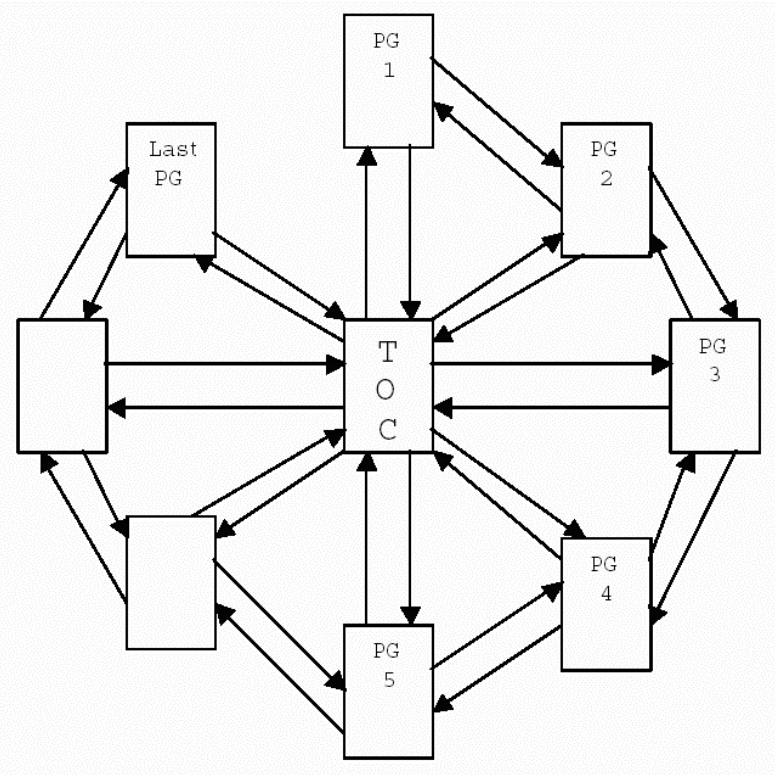

Fig. 2: Navigation topology of the journal

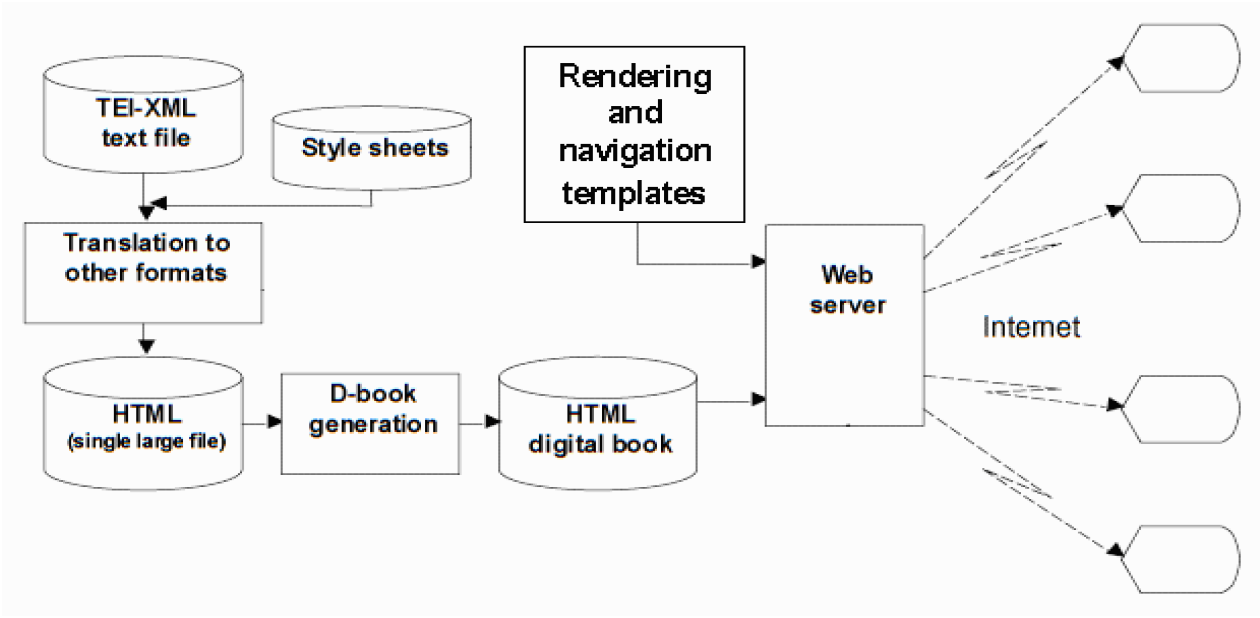

Fig. 3: Generation of the HTML publication format for the journal.

Dissemination approach

According to Ceri et al. [11], personalization is: *to give the user the impression of interacting with the application by means of a dedicated interface, specifically tailored to the user's needs and preferences*[8]. Content personalization systems select, adapt and generate contents according to user models that define information needs [9]. To achieve this we first followed the dissemination model, as described by Yan and García-Molina [8]. In this model, users subscribe to an information dissemination service by submitting a set of personalization settings that describe their interests, which are usually called *profiles*. Then they passively receive information filtered according to such profiles. However, we now believe the process doesn't need to be so passive. What's more, it is desirable that the sender of such information (DL) get a feedback of the users interests, which may evolve with time, or of the users opinions on the information received (as described by Díaz et al. [9]). This would result in interesting metrics.

In the first stage, our profile was very simple and allowed only the choice of one or more possible thematic newsletters out of five (see fig. 4).

In a second stage, we are now remodeling this personalization approach to both offer a finer granularity (more detailed personalization choices) and also obtain feedback from users which will allow dynamic changes in personalization settings and will give us useful information on user preferences (user metrics).

Koch [12] states that user profiles and preferences can be acquired in three ways:

1. Can be set by the application designer by means of *stereotyped* techniques.
2. Can be given by the users themselves by means of *interview* techniques (this includes input forms).
3. Can be automatically inferred by the application based on user activity: *observation* techniques.

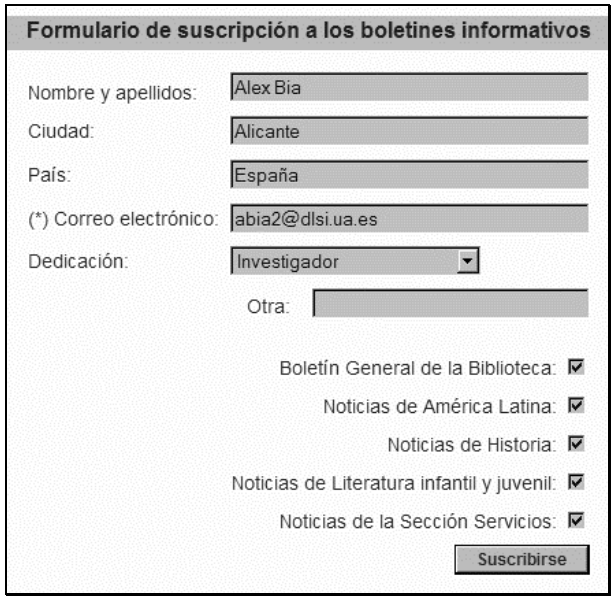

Fig. 4: Original subscription form.

Mi Bilioteca: a successful experience

In another personalization service we offer, called *Mi Biblioteca* (My Library), we have successfully applied the third method –*observation*– in a user-supervised fashion (see figure 5). Through this service, users can create a personalized view of the library by clicking on marker buttons as they navigate the DL. In this way, users can mark:

- entire books of their choice
- selected pages of a book and attach a personal reminder to it
- selected authors

- discussion forums
- selected subjects according to the CDU classification
- selected subjects according to the UNESCO classification
- apart from these markers, *Mi Biblioteca* automatically stores the last five pages of the last five books accessed

As a result, a personal view of the library is created where users can start and continue working with materials of their choice without having to search for them every time they connect to the DL. So in this way, *Mi Biblioteca* acts like a persistent personal work space. Additionally, an optional email-reminder service tell users of new additions to the library that match the items or categories selected: e.g. new works of a selected author, or new works under a selected CDU or UNESCO category.

Personalization and filtering of newsletters.

The user model we are implementing now for newsletters (second stage) falls under the second category of Koch's classification: the interview model, although we may expand it in the future (third stage) to incorporate information gathered from user navigation, like we do in *Mi Biblioteca*. The final model will be based both on expressed interests on certain DL sections (information gathered during navigation) and on certain types of news that compose the user profile. Apart from the five types of newsletter already mentioned above on fig. 1 (general DL news, Latin-American literature, children's literature, history and interactive services), we will provide one further level of personalization by allowing the users to select the types of news they are interested in with finer detail (more granularity). There are three main categories (DL news, suggestions and new books) and several subcategories:

NEWS
- new sections added
- history
- Catalan/Valencian language
- literature
- humanities and computing
- humanities research
- conference announcements
- future publications
- new publications
- visits of important people
- chat announcements
- contests

SUGGESTIONS ON
- Latinamerican literature
- humanities research
- history
- children's literature
- theatre and movies
- interactive services
- computer tools

NEW BOOKS
- Latinamerican literature
- children's literature
- critical studies
- humanities research
- history
- interactive services
- Tribute to Hispanists
- Exile Library
- Argentine Academy of Letters
- Thesis
- Movies
- Magazines
- Recently printed works

Finally, in a third stage of incremental development, we plan to enhance the personalization of newsletters by filtering news by areas of preference expressed by users as they navigate, using the same approach of *Mi Biblioteca*.

Conclusions

This solution to the production of newsletters and a journal for a DL saves an important amount of time: now we can produce the newsletters and the journal in less time and with less effort than what was previously required to produce only the newsletters by hand.

The output formats are uniform and regular, and less error prone. Previous newsletters showed les uniformity and some rendering errors.

Compared to the other production models of our DL (for digital text books and for digital facsimiles books), this model is different in three ways:

- The production does not begin with scanning. This model deals with the production of digitally born material.
- This model includes many more different output formats.
- This model is tied to a given fixed periodicity proper of this kind of publication.

The general newsletter, which is the most demanded one, currently has 14,000 subscribers worldwide.

Concerning the dissemination model for these newsletters, we are now enhancing the granularity by offering more detailed personalization options and in a near future we plan to apply to newsletters the same approach we have successfully applied to the *Mi Biblioteca* personalization service, which will allow us to filter news based on preferences gathered from user navigation (observation model).

The MCDL, with this effort, struggles to fulfill its objective of spreading research knowledge to the global academic community through the Web. Our aim is to be

able to offer a better service by optimizing searches of digital resources, by reducing waiting times for digital publication, by promoting dynamic scientific research communication and by developing efficient preservation strategies.

Daily experience results in the continuous integration of new ideas. Our goal is not only the mere publication of research work, but to build a rich and open communication channel for the global scientific community, with the aim of making true the Enlightenment dream of the "Republic of Letters".

Appendix: file formats

News in XML format:

The following is a piece of news in XML format. Attributes indicate in which newsletter/journal should be included:

```
<div1 general="yes" AmLat="no" children="no"
history="no">
<head type="main">2.4- Ediciones Multimedia:
</head>
<p>-Hemos publicado la edici&oacute;n
multimedia del gui&oacute;n de "La Corte de
Fara&oacute;n" (1985), de Rafael Azcona y
Jos&eacute; Luis Garc&iacute;a S&aacute;nchez.
Se trata de una pel&iacute;cula clave para
acercarse desde una perspectiva
humor&iacute;stica al fen&oacute;meno de la
censura teatral durante el franquismo. Puede
encontrarse, adem&aacute;s, una serie de
materiales diversos sobre este divertido film y
la opereta hom&oacute;nima, con los que es
posible conocer algunos de sus detalles
m&aacute;s relevantes.</p>
<p>          <xref          doc          =
"http://cervantesvirtual.com/086.pdf?incr=1">
</xref></p>
</div1>
```

Newsletter in plain text format:

```
*******************************************
Boletín de la Biblioteca Virtual M.de Cervantes
Número 25
http://cervantesvirtual.com
*******************************************

CONTENIDO DE LA REVISTA:

[1] NOTICIAS:

-->1.1- Mañana, tertulia virtual sobre novela e
Historia con Jorge Edwards, Premio Cervantes de
las Letras de 1999.

-->1.2-      La      Biblioteca      Virtual
digitalizaráfondos del Ministerio de Educación,
Cultura y Deporte, nuevo patrono de la
Fundación Biblioteca Virtual Miguel de
Cervantes.

-->1.3- Inauguramos la Biblioteca de Autor de
Larra cuando se cumplen 166 años de su muerte.

-->1.4- La Biblioteca Virtual participará en la
"7.° Conferencia Europea de Investigación y
Tecnología Avanzada para Bibliotecas
Virtuales".

-->1.5- Ya está disponible la edición facsímil
de la "Fábula de Piramo y Tisbe", de Luis de
Góngora.
```

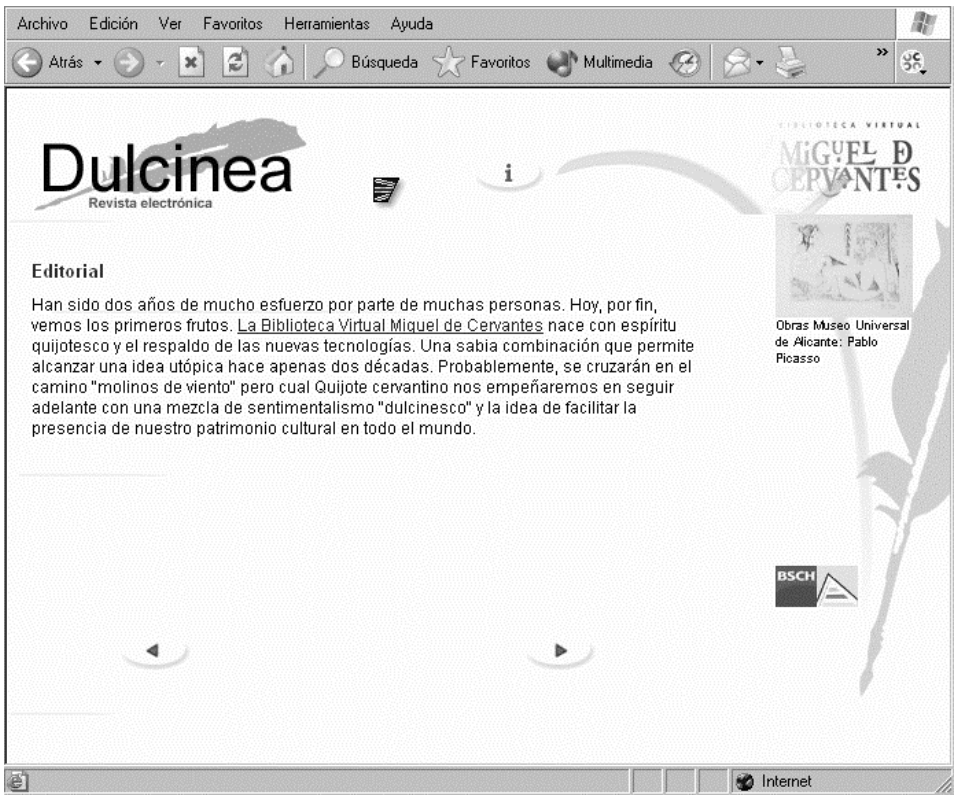

Fig. 3: The Dulcinea monthly journal

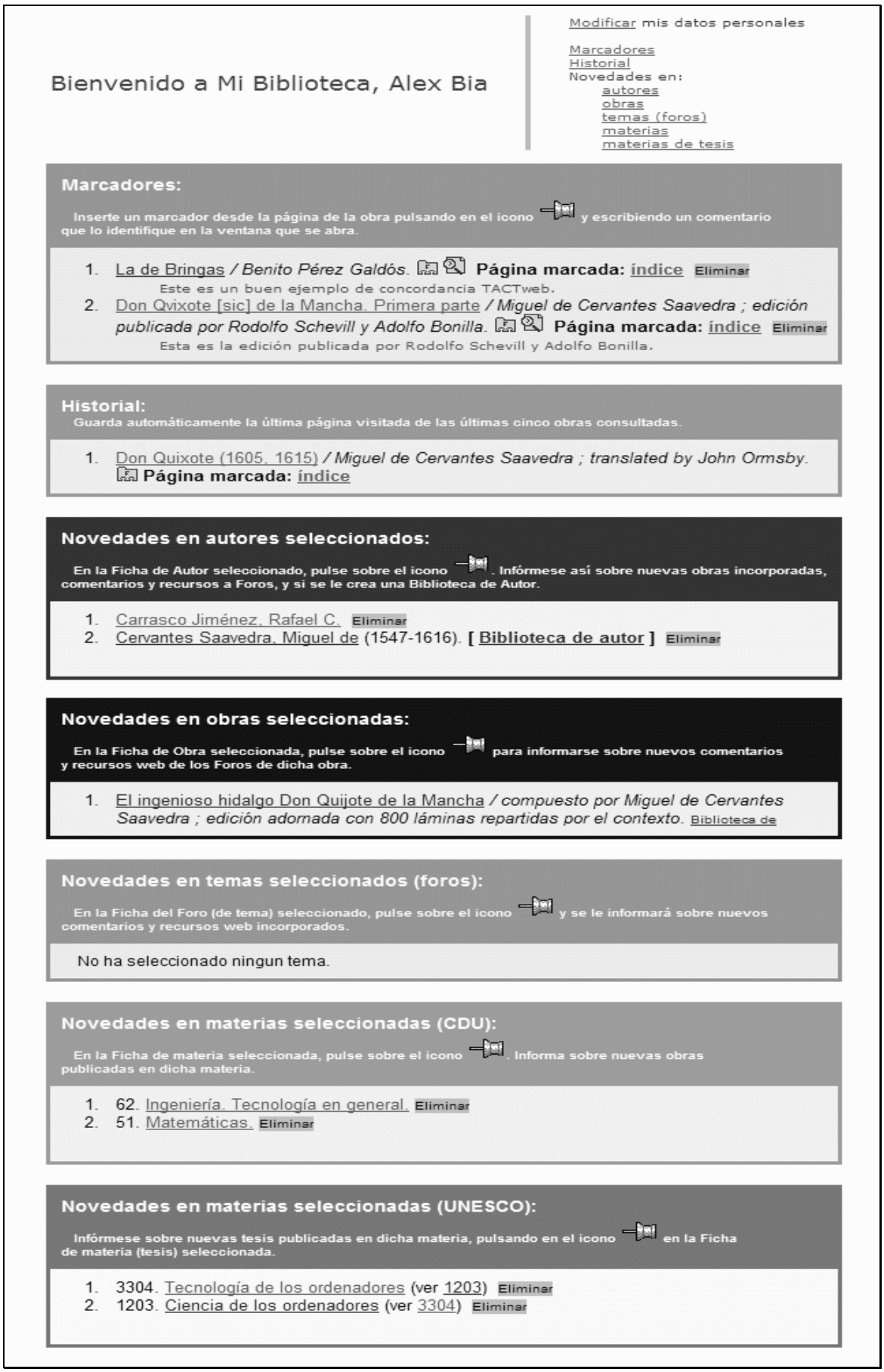

Fig. 5: Mi Biblioteca: a personalized view of the DL (image captured from browser).

References

[1] Alejandro Bia and Andrés Pedreño. The Miguel de Cervantes Digital Library: the Hispanic voice on the Web. *LLC (Literary and Linguistic Computing) journal*, Oxford University Press, 16(2):161–177, 2001.
ISSN: 0268-1145
(First presented at *ALLC/ACH 2000, The Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities*, 21/25 July 2000, University of Glasgow.)
http://www3.oup.co.uk/litlin/hdb/Volume_16/Issue_02/160161.sgm.abs.html

[2] Julia Puig and Alejandro Bia. The Miguel de Cervantes DL's Digital Theses Program. In *ETD 2003 (Sixth International Symposium On Electronic Theses and Dissertations)*, Humboldt-University, Berlin, 21-24 May 2003. Networked Digital Library of Theses and Dissertations (NDLTD).
http://www.hu-berlin.de/etd2003/

[3] Neil Bradley. The XML Companion. Addison-Wesley Longman Limited, 2000.
ISBN 0-201-41999-8

[4] Alejandro Bia and Rafael C. Carrasco. Automatic DTD simplification by examples. In *ACH/ALLC 2001. The Association for Computers and the Humanities, The Association for Literary and Linguistic Computing, The 2001 Joint International Conference*, pages 7–9, New York University, New York City, 13-17 June 2001.
http://www.nyu.edu/its/humanities/ach_allc2001/papers/bia1/index.html

[5] Michael Kay. XSLT Programmer's Reference. Wrox Press, 1102 Warwick Road, Acocks Green, Birmingham, B27 6BH, UK, 1st edition, 2000.
ISBN 1-861003-12-9

[6] William Arms. Digital Libraries. MIT Press, Cambridge, Massachusetts, 2000.

[7] Francesca Marí-Domènec and Alejandro Bia-Platas, Técnicas de producción masiva de contenidos para bibliotecas digitales, in *III Jornadas de Bibliotecas Digitales - JBIDI 2002*, José H. Canós, and Purificación García eds., pgs. 123-132, Universidad Politécnica de Madrid, 18-23 November 2002.
ISBN 84-688-0205-0

[8] Tak W. Yan and Hector García Molina, SIFT - A Tool for Wide-Area Information Dissemination, in *1995 USENIX Technical Conference*, pgs. 177-186, 16th February 1995.

[9] Alberto Díaz-Esteban, Pablo Gervás Gómez-Navarro and Antonio García-Jiménez, Desarrollo de una colección de evaluación para personalización de periódicos digitales (development of an evaluation collection for personalization of digital newspapers). In *JOTRI 2003 - II Jornadas de Tratamiento y Recuperación de la Información*, pgs: 10-17, ISBN 84-89315-33-7, organized by the "Departamento de Biblioteconomía y Documentación" and the "Departamento de Informática" of the Carlos III University, 8-9th September 2003, Universidad Carlos III, Leganés (Madrid).

[10] Cristina Cachero. *OO-H: Una Extensión a los Métodos OO para el Modelado y Generación Automática de Interfaces Hipermediales*. Tesis doctoral, Depto. de Lenguajes y Sistemas Informáticos, Universidad de Alicante, 2 de Diciembre de 2002. Directores: Dr. Jaime Gómez Ortega y Dr. Oscar Pastor López.

[11] Stefano Ceri, Piero Fraternali, Andrea Maurino, and Stefano Paraboschi. One-to-One Personalization of Data-Intensive Web Sites. In *WebDB (Informal Proceedings)*, pages 1–6, 1999.

[12] Nora Koch. *Software Engineering for Adaptive Hypermedia Systems: Reference Model, Modeling Techniques and Development Process*. PhD thesis, Institut für Informatik, Ludwig-Maximilians-Universität München, December 2000.
http://www.pst.informatik.uni-muenchen.de/kochn/abstract.html.

♣ This paper has been partially supported by the Spanish Ministry of Science and Technology (Ministerio de Ciencia y Tecnología de España), project TIC2001-3530-C02-02.

[1] http://www.cervantesvirtual.com/bibliotecario/
[2] http://www.cervantesvirtual.com/foros/foros.shtml
[3] http://www.cervantesvirtual.com/tertulia/
[4] http://cervantesvirtual.com/noticias/noticias.htm
[5] http://cervantesvirtual.com/noticias/dossier/index.shtml
[6] http://cervantesvirtual.com/noticias/catalogo/index.shtml
[7] http://cervantesvirtual.com/noticias/catalogo/buscador.shtml
[8] This ability to dynamically alter an application to give the impression that it was specially tailored to oneself is called one-to-one delivery.