# The Galician Virtual Library[1]

Ángeles S. Places, Nieves R. Brisaboa, Antonio Fariña, Miguel R. Luaces, José R. Paramá, Miguel R. Penabad

asplaces@udc.es, brisaboa@udc.es, fari@udc.es, luaces@udc.es, parama@udc.es, penabad@udc.es

Department of Computer Science
University of A Coruña
Facultade de Informática
Campus de Elviña, s/n
15071 A Coruña, Spain
Tlf: +3498116700
Fax: +34981167160

## ABSTRACT

**Purpose –** To present the digital library *Galician Virtual Library* (BVG, for "Biblioteca Virtual Galega" in Galician, "Galician Virtual Library" in English).

**Design/methodology/approach-** Shows the objectives pursued by the BVG, its development, making special emphasis in the main technological challenges, and presents some data about its usage.

**Findings-** A digital library can be used to stimulate a lesser used language and to promote the culture and tourism of a region.

**Original/Value-** Shows how a Digital Library can be used to strengthen the Galician language, which is currently categorized as a "Lesser Used Language" in the European Community and to contribute to the preservation and spreading of Galician culture and literary works, either from current authors or antique documents. It also provides a digital publishing house for new authors and opens a communication channel between current authors and their readers. Finally it helps to connect a scattered community like the Galician, offering a centralized access point to any information about Galicia. This work also presents some technological innovations included in the BVG, especially from the viewpoint of user interface design and search by content.

**Keywords-** Digital libraries, user interfaces, text retrieval, geographic information systems.

**Paper type-** Case study.

## 1. INTRODUCTION

Digital libraries have attracted much attention during the last decades (Baeza-Yates and Ribeiro-Neto, 1999; Borgman, 1999; Gonçalves, *et al.,* 2001; Gonçalves, *et al.,* 2002; Lesk, 1997). However, there is still no consensus on the definition of a digital library. From a very general point of view, Lesk (1997) defines it as a "collection of information that is both digitized and organized". Borgman (1999) offers a thorough revision of the definitions of a digital library. Among them, the following is one of the most interesting, since it includes the technical point of view, but it also considers the people that create it and reflects the information they use. According to her, a digital library is a collection of electronic resources —its content— and the infrastructure to manage search and use the information. The content of a digital library includes the information itself, that can be as simple as plain text or more complex data (audio, video, pictures, etc.), but it also includes metadata, such as bibliographic information, the format of the document, copyright, or links to related documents. Furthermore, it is usually built for a community of users to fulfil their information needs. A digital library can be seen as

[1] This work is partially supported by Grants TIC2003-06593 and TIN2006-15071-C03-03 of Ministerio de Educación y Ciencia and PGIDIT05SIN10502PR of Xunta de Galicia.

an enhancement and complement of physical places and institutions for the interaction among a community of users that takes advantage of the information collected and organized in the digital library.
According to the previous definition, it seems appropriate to consider two different types of issues that affect the development of digital libraries: technological issues that affect its architecture and functionalities, and sociological and cultural aspects related with the community that will use it.

Among the technological aspects, besides the design of the data model and the global architecture of the digital library, we pay special attention to the detailed analysis of the design of the user interface and the search by content (text retrieval) service. We consider the user interface a fundamental part of a digital library (especially if it is a Web site), because if users do not find it intuitive and friendly they will not use it. With respect to searches, although BVG has metadata to perform the typical searches, we believe that an advanced digital library should provide the capability of seeking literary works also by their content, taking advantage of the digital nature of the literary works.

We shall describe in this paper the ambitious project that led to the development of the Galician Virtual Library (BVG, for "Biblioteca Virtual Galega" in Galician, available at http://bvg.udc.es/). Such a description includes the main goals and requirements, as well as the special characteristics of the architecture, content, services, and user interface of the BVG. An early version of the BVG project was published in Spanish in (Brisaboa, *et al.,* 2002).

Regarding technological aspects, we shall describe in this paper the most interesting challenges, such as the Geographical Information System included in BVG, or the connection of BVG, as a *data provider* of the standard architecture of Open Archives Initiative, with the EMILL (European Minority Languages Library) project. We also considered BVG's user interface a fundamental issue, and we have followed some general guidelines and approaches for its design, like the systematic use of cognitive metaphors (Zetie, 1995) and the *browsing approach* (which has already been successfully used in (Brisaboa, *et al.,* 2000; Paramá, *et al.,* 2006)) to help users navigate in our library. Finally, we show the search by content service built for the BVG. This service uses two technologies. First, an inverted index was built, adapting open source libraries from the Lucene project (Lucene, 2006). Second, we modified and implemented several pattern matching algorithms which are suitable for the different types of searches included in the BVG.  As we will show, both technologies are combined to obtain optimum response times.

Considering sociological and cultural aspects, a digital library like BVG can be considered an excellent medium to connect scattered communities, as well as to promote heritage, cultural or linguistic values. The special characteristics of the Galician culture and language influenced the design of the digital library. The Galician community is dotted among many different countries, mainly due to the emigration that happened during the 19th and 20th centuries, but even second or third generation Galicians feel a strong tie with their home country. Galician language, on the other hand, is widely understood in Galicia and by the Galician Diaspora, but it is not as frequently used as it should be.

The rest of this paper is organized as follows. First, we provide a brief description of the history of Galicia and Galician culture. The next section details the goals of our project; this is followed by discussion of the contents, architecture and services offered by BVG. BVG's user interface is treated in a separate, following section due to its importance. We then move on to a presentation of the search by content (or text retrieval) service; this is followed by discussion of information about the impact and usage of BVG.

## 2. A BRIEF HISTORY OF GALICIA AND THE GALICIAN LANGUAGE

Galicia is a region of Spain that occupies an area of 29.432 km$^2$ in the Northwest of the Iberian Peninsula, located just to the north of Portugal. Nowadays, about 2.800.000 people are living in Galicia. Most of them, around 84%, claim to speak Galician and about 91% claim they understand it perfectly.

Romans conquered Galicia and transformed it into a province of the Empire, calling it Gallaecia. Galician, the current official language in Galicia in conjunction with Spanish, comes from the Latin brought by the Romans, but it also includes many words coming from previous cultures like Celts or Swabians, a German culture.

The variation of Latin in the West of the Iberian Peninsula produced a new language called Galician-Portuguese, the language of the western Peninsula between the 12th and 15th centuries. In this stage, the "golden age" of Galician literature, Galician-Portuguese lyrical poetry emerges strongly boosted by the pilgrimage to Santiago. At this time, Galician was an "international" language used not only by writers but also in royal and feudal courts in the West of the Iberian Peninsula.

When, in 1640, Portugal and Spain definitely became two independent countries, Portugal took Portuguese as its official language. However, Galician was excluded in Spain from official texts for the sake of Spanish, and survived only orally. This era is known as the "Séculos Escuros" or The Dark Centuries, where Galician suffered a high level of dialectalism. The "Rexurdimento" or Revival period occurred in the 19th Century. The recovery of Galicia was literary but also cultural, political and historical. Writers of this period had to "invent" orthography because Galician language had not been written for three centuries. At the end of the Spanish Civil War (1936-1939) and the start of Franco's regime, Galician language was forbidden also in official forums, until the last years of the 20th century. With the democracy, Galician became alive again, being protected by law and spoken by the Government and by a quite large number of people. In the last years of 20th century, Galician literature was rapidly gaining importance.

At the same time, during the 19th and 20th centuries, due to their poor economic status, many Galician people (almost 25% of the population) emigrated to different countries. These people and their descendants still maintain a close contact with Galicia, although they are a very scattered community. It is important to note that, for a region with less than 3 million inhabitants, the Galician Diaspora is formed by about 500.000 people.

## 3. THE GALICIAN VIRTUAL LIBRARY PROJECT

The project was developed by two teams of researchers, on Galician Philology and on Computer Science, from the University of A Coruña (more information about these researchers can be found on the BVG web site, http://bvg.udc.es).

Taking into account the special characteristics of Galician language, we started a project to build a digital library devoted to the Galician culture and language with three main goals:

- *To promote the use of, and knowledge about, Galician language*. In general terms, Galician is understood by a large percentage of Galician people, but it is not so widely used. According to the European Bureau of Lesser Used Languages, 91% of Galician population understands Galician, and 84% can also speak it, but only 48% use it on a daily basis (see http://www.eurolang.net/State/spain.htm#Galician for a full report and statistics). The Government of Galicia is trying to increase this rate, and in fact it is used, jointly with Spanish, in all official dealings with the administration.

  Teachers of any level of education are also encouraged to use Galician. In this sense, and trying to help using a Galician of better quality according to the rules of the Royal Academy of Galicia (which is the official institution in charge of ruling the use of Galician, trying to remove the high level of dialectalism suffered by it), our objective was to offer a section entitled "recursos de lingua" (language resources), which leads to a number of dictionaries, grammars, and other documents useful to learn or improve Galician.

- *To spread and promote Galician literature*. Nowadays, probably any language that is not present in Internet quickly looses importance and sociological "prestige", which can even lead to its eventual disappearance. To avoid this happening to Galician, our second goal was to spread Galician literature by including in BVG works from classical and current writers. For the latter, we also offer a direct communication channel between authors and their public. We have created a Web page for each author, where he or she can announce news, such as encounters or the appearance of a new book written by them. These Web pages can be used by their readers to send messages to the author, or to make comments on his/her works.

Finally, we want also to encourage new authors, by offering them a digital publishing house (called e-Dixital) to publish their works on BVG.

- *To spread Galician culture*. Galicia has many cultural elements apart from literature. Among them, there are many archaeological places, spas, or tourist places that deserve a visit. We included in BVG a Virtual Tour through Galicia (see (Penabad *et al.,* 2003) for details), a Geographical Information System with information about several interesting places, and where it is also possible to plan routes or to obtain information about routes travelled by some prominent Galician personages in the past. Additionally, since the Galician community is scattered around the world, we try to offer a centralized point to have access to any type of information about Galicia, including its culture, and its literature and other important aspects such as tourist information, newspapers, universities, etc.

Once BVG was developed we have verified that it is a very useful tool for many different areas. It is a multimedia encyclopaedia of Galician works and writers from all times. As such, BVG is now the main Web reference for schools and high schools in Galicia. As a research tool, it is also useful on several areas, such as user interface design, natural language processing (having a Galician tagged corpus) or text retrieval (we developed an inverted index and several ad hoc pattern matching algorithms). Other apparently distant disciplines, such as sociology or genre studies, also benefit from BVG, allowing research like *women in BVG* (Fernández P-Sanjulián, 2005).

# 4. CONTENTS, SERVICES AND ARCHITECTURE OF BVG

## 4.1 Contents of BVG

BVG is a multimedia digital library that can be seen under three perspectives: an encyclopaedia, a library, or a digital publishing service. Under the first perspective, BVG contains a large volume of (metadata) information about authors and their works, including biographies of classical and contemporary authors. In the case of living authors, BVG is continuously evolving since such authors provide their autobiographies, and they can write news, opinions and comments at any moment. BVG can also be seen as a multimedia library, since it contains complete texts, digitized pages, video or audio clips of Galician literary works. Finally, BVG is also a digital publishing house. We have included a section (“New authors” section or “Sección de Novos”, linked to the digital publishing house *e-Dixital*) which tries to encourage new writers. It contains works by new, unknown authors, mainly in text or HTML format.

BVG currently stores information about 344 authors and 3849 works, some of them having several editions. More than 800 of these works are available for reading, either as texts or HTML pages. We have 62 works digitized (a total of 5912 digitized pages). There are also audio files with excerpts from 34 works (authors reading their works, or prominent writers reading classical books) and 17 video files of the same nature.

We asked authors to write a new small literary work to offer it as an exclusive of our digital library. Many of them agreed to write it, so we are proud of having in our library many literary works of great quality, which are not published elsewhere. Other authors preferred to transfer their copyright on some works to BVG, so these works were digitized completely and are available at BVG as well as in paper.

As a summary, the contents available at BVG can be classified on the following categories:

- *Language resources*, containing information about paper resources about Galician language, including dictionaries and grammars.
- *Digitized pages from original works*, either classical or contemporary writings. We must note that, due to copyright limitations of some works, only 10% of the complete work could be digitized and included in BVG.
- *Transcribed texts*, that can be the result of an optical character recognition (OCR) process over digitized pages, or typed (either by us, or directly sent by the author). These texts can be either plain text files or HTML formatted documents.
- *Video and audio clips* of Galician writers reading their own works or classical ones.

- *New authors writings*: This category contains works from new authors, included through e-Dixital. It is a section that is continuously growing.

We want to emphasize the double role of BVG as a library and as a publishing house. We believe that, for a language with few speakers, it is very important to provide a digital publishing company, because it permits the publication of works that due to publication expenses, would never be published on paper.

## 4.2 Services offered by BVG

Any community of users of a virtual environment, especially of a virtual library, would benefit from value-added services that allow them to share their knowledge, cooperate in their work, or simply provide an open environment where discussions can take place. BVG offers several additional services. Among them, we can highlight the following:

*Services for authors*

- *Personal web pages for authors:* every author on our library has a personal Web page that they can manage autonomously. It includes his/her autobiography, which most of them wrote in a very creative way. Another section includes a list of the writings by the author, with a link to access them, if such works are available at BVG. Finally, it also has a *news* section where they can write announcements or even use it as a discussion forum. This news section is under his/her sole responsibility, meaning that there is no censorship by the library maintainers. It is commonly used to announce a new book by the author or forthcoming events of interest to Galician writers or readers, but it is also sometimes used to publish personal opinions about news, political happenings, etc. An additional advantage of the use of this section by any author is that all authors having news appear on the BVG main page, so it serves as a "publicity service" for the author. We have verified that authors having news have a higher number of accesses than those not offering any news. Figure 1 shows an example of an author Web page.

  This service uses the Virtual Library database, but files uploaded by authors are stored in a separate repository. These files are processed by the BVG management team (converting them to the required format and dimensions) before making them available to the public.

- *Messages to authors:* As it would seem obvious, authors value very positively the comments on their works by their readers. Similarly, readers are very interested in getting "closer" to the authors they like, so they can send them either personal messages or comments on their works. In order to provide a virtual communication channel between authors and their public, we have included on every author Web page the possibility to send a message to the author, which is directly sent to him or her (again, there is no censorship on the message contents). If the public is interested in getting an answer by the author, he or she has only to give his/her e-mail address in the message. We have received very positive feedback from both authors and public after including this service in BVG.

- *New authors:* One of the goals of our project was the promotion of the Galician language. Thus, encouraging new authors to write in Galician seems a good idea. Unfortunately, the publication (in paper) of a book for a new, unknown author is difficult and expensive. Trying to overcome this problem, BVG offers a virtual editorial, where unpublished works by either known or unknown authors can be published. Using the New Authors section ("Sección de Nov@s"), any person can submit a work, together with some personal information. The work can be in any of the electronic formats accepted by the library, and will be accessed, and commented online, by the public. These comments on the works convert this service into a debate forum, very valuable for both authors and the public, because they can refer to the style of the writing but also about its contents. There are also periodical contests where the public "grades" the new works. The prize for the winner is to be included, as a regular author, in BVG.

*Services for the public*

- *Search by content:* Most of the searches performed in BVG are made using the browsing approach described in the next section, because people usually search for a specific author or books of a given genre and epoch. However, we wanted to offer a more powerful query engine, using a natural language interface to query the library documents by their contents. In this way, it is possible to find all works containing a specific word or sentence, but more complex queries are also available, such as fuzzy queries (containing words with a number of typos or mistakes), Boolean searches, etc.

This service is implemented as an independent module. It needs the texts to be indexed, so we decided to use the Lucene Java library (Lucene, 2006) and our own implementations of some pattern-matching algorithms to manage the search module. The description of this service is presented in later in the paper.

- *Direct information sharing with lg3 official page:* lg3 (http://www.culturagalega.org/lg3) is a Web portal from the "Consello da Cultura Galega" (Galician Culture Secretary), depending on the State Government of Galicia, which is mainly dedicated to Galician Literature. With this service, we share information both ways: we have on the main page of BVG the most important news in lg3 Web, and when a search for a specific author is performed in lg3 Web page, the answer includes results found in BVG, allowing users to directly access the information that BVG has about such an author, or even his/her works, if they are available for reading, listening to, or watching as videos.
- *Virtual Tour through Galicia:* We decided to include this service to provide our users with a useful tool to know the great archaeological, artistic and monumental heritage, as well as our tourist resources (beaches, country houses, spas, etc.). This service is a Geographic Information System (GIS) that includes information about a number of aspects of Galician culture and tourist resorts. Museums, monuments, archaeological sites, country houses, natural areas and a long list of other interesting sites can be located in a map of Galicia. Figure 2 shows this service, where 2 different areas can be found: the map itself, and a menu area that allows users to interact with the map. Initially, the map showing the 4 Galician provinces is displayed, showing their councils. From this base, users can choose which information is added to the map, in different layers. The map in Figure 2 shows only beaches, country houses and rivers. The map itself is also interactive: when the mouse is moved over an element, it is highlighted, and its name is shown. By clicking on the element, a pop up window appears with more information. For example, if the chosen element is a museum, this window shows a picture of the museum, its address, opening hours, and its phone number, if available.

  There is also the possibility to perform searches: for cities or councils, just selecting them on a drop-down list. For a specific element (such as a river, museum, or country house), the user chooses the category of the element and types part of its name. A list of elements matching these search conditions is shown and, once one element is selected, it is highlighted on the map. Users can also zoom in or out, or select only one province, to do the searching.

  There is also another service implemented on the virtual tour, which might be very interesting for those people planning a trip to Galicia. It is the possibility to display one of the many "programmed" tourist itineraries or literary routes, travelled in the past by a classical author in our library. All interesting elements along the routes are automatically selected. Finally, and also useful for people planning a travel through Galicia, the different network roads of Galicia can also be displayed on the map. More information on the Virtual Tour can be found on (Penabad *et al.,* 2003).

## 4.3 An overview of the architecture of BVG

Figure 3 shows the general architecture of BVG. It is fully modularized, having a subsystem for each of the services BVG offers. As the figure shows, there are two main modules, the *Virtual Library System* and the *Virtual Tour System*. The Virtual Library system is the core of BVG, and it is responsible of most of the services described under *services offered by BVG* and the digital library itself. However, due to the special nature and complexity of the *Virtual Tour through Galicia* service, it has its own module (the Virtual Tour System module).

The *Virtual library system,* which is a Web application, was mainly implemented using Java, and it accesses the underlying database (we are using MS SQL Server 7 as database server) and a set of file repositories. Most of the modules of this system are in charge of one of the services previously described. These include LG3, author management, and text retrieval. The library catalogue, which was implemented using JSP pages with embedded Java code, had its most interesting challenges in the design of its user interface, so we devote the next section to describe it. Let us, then, focus on the remaining modules.

Since BVG participates in EMILL (European Minority Languages Library) project (http://www.emill.org) with some other digital libraries (by now, only a representation of Frisian language is available), we developed a software module to accept OAI PMH requests (Van de Sompel and Lagoze, 2000), which can be sent to http://bvg.udc.es/oai/. As a provider, BVG publishes data about editions of works, using a Dublin Core Metadata Element Set with an identifier of each edition, its creator (author of the work), date of publication, type (narrative, poetry, essay, etc.), title, publisher, coverage, language (Galician) and source (URI of the work at BVG).

The *BVG Management subsystem* implements the modules to carry out the daily management tasks of BVG, and it is being used by a multidisciplinary team, formed by specialists in Computer Science and Humanities: a philologist, 3 specialists in librarianship, and 2 analysts/programmers. This team is directed by two well known researchers on the areas of Computer Science and Galician-Portuguese Philology. This subsystem includes a module to obtain statistics and generate reports about the usage of BVG, a module to perform database and repositories backups, a metadata management module, and a mail generator module. This module is used to send mails to our authors (either electronic and/or postal mails, as they choose) with the comments sent to them by the public.

This management subsystem also includes the applications needed to incorporate all the literary works in BVG. The process followed to include these writings varied depending on two factors: the type of source documents (either paper or electronic format) and the desired target format (image, plain text, or HTML). Documents written by classical authors were available on paper only, therefore we developed specific applications to help our management team process them: first, the pages were digitized using a scanner or digital camera, and processed with an OCR application to obtain text format. Later, these texts were corrected and formatted, obtaining either text files or HTML pages.

The architecture of the *Virtual Tour System* is shown in Figure 4. The data is stored using a PostgreSQL database management system, extended with the PostGIS module to manage spatial information. The data is extracted from the database by a Web Map Service (WMS). This web service receives map requests following the OpenGIS standard operations (see http://www.opengeospatial.org/standards/wms) and serves Scalable Vector Graphics (SVG) maps. There are already free implementations of the OpenGIS WMS, in our case we used the Deegree WMS (http://www.deegree.org/). The SVG map provided by Deegree is enriched with more information by a Post Process Module implemented in JSP by us. This added information is needed, for example, to display the name of a geographic object (say a city) when the user moves the mouse over such a geographic object. Then, the enriched SVG map is transmitted through the net until it reaches the client (a web browser). The web browser receives JSP and a Java Applet, which is in charge of displaying the final map using the information of the received SVG map.

## 5. BVG USER INTERFACE

As mentioned in the introduction, many interesting technological challenges on the design and development of BVG are related to its user interface. In this section we shall describe the different sections of BVG, and how their usability was improved by using special techniques, such as the use of cognitive metaphors or a browsing approach.

Cognitive metaphors or analogies are a well known and widely used technique on the design of user interfaces (Zetie, 1995). They are based on a simple, but effective principle: the use of something known by the user, translated into another domain. An example of analogy is the Windows calculator, which uses the metaphor of a calculator, a well known tool. Since the aspect and functionalities of a real calculator and the "virtual" one are the same, the user knows how to use it. In the case of the BVG, for example, the main page shows a room representing a real library, that is, it shows several desks with drawers (which, in real libraries, use to contain the index cards). Among those desks, for example, there is one with an array of hi-fi to denote that by clicking on it, the user can access literary works in audio format.

The *browsing approach*, which we have already used successfully in other digital libraries (Brisaboa, *et al.,* 2000; Paramá, *et al.,* 2006)) is also based on a simple principle: when offering a service to search for information, try to avoid complex query systems (usually query languages) and, instead, provide simple interfaces that lead the user to obtain the desired information by following several easy steps. The reason we found to start using this technique was that several studies demonstrate that users prefer to build single queries, retrieving a large set of results and navigating through them, than building complex queries that retrieve fewer and more precise results, even a unique response. Lucas (2001) shows a study done by popular search engines like Excite or Altavista, where they found that up to 72% of the queries only contain one or two words, and nearly 80% of all queries did not include any operator (either Boolean, the symbols "+" that forces a word to be included in the document, or "–" that forces a word not to appear, etc.).

The main page of the Galician Virtual Library, which gives access to all its services, is shown in Figure 5. Notice that the digital library itself uses a *library metaphor*, that is, it has similar aspect and functionalities as a real library. Each category of information is accessed through a different section of the digital library. Thus, by clicking on the TV set or camcorder, the user accesses the video library, and clicking on the door labelled "Sección de Nov@s", the user enters a "room" where he or she can read and comment works by new authors, or even submit his or her own writing.

One of the most used services offered by BVG is the catalogue of authors and works, shown on the left of Figure 6. Notice again the use of a cognitive metaphor, with the catalogue being represented by a set of drawers. By "opening" (clicking on) some of these drawers, namely those referring to some specific kind of literary works, the user obtains a list of works. Then, he or she can access the contents of these works, reading or listening to them, or simply watching a video of the authors reading them, depending on the available formats for the work on our library.

If the user opens a drawer on the second line, which is devoted to the authors, he or she accesses a new page that uses yet another cognitive metaphor: a set of cards containing the authors' names, as shown on the right of Figure 6. It resembles an actual drawer full of cards, arranged in alphabetical order. By clicking on a letter, the user obtains the list of authors whose name begins with this letter (note that several authors can appear on one or more cards, as we have indexed them by all the names or pseudonyms they use). Clicking on an author name, the user accesses his or her web page, which has been already described.

## 6. SEARCH BY CONTENT

Galician language has had several orthographic regulations. Even in the current regulation of the Galician language (approved by the Royal Academy of Galicia), several alternatives for the same word are allowed. Thus, some typical information retrieval approaches (like, for example, the vector space model (Salton, 1968)) are not suitable for our case. We chose the information retrieval technique called Extended Boolean (Baeza-Yates and Ribeiro-Neto, 1999), which allows exact searches for words and phrases, search for prefixes and fuzzy or approximate searches. In addition, the results can be sorted by relevance, which in our case was done using the inverted frequency (Baeza-Yates and Ribeiro-Neto, 1999). An early version of this module was published in Spanish in (Vazquez *et al.,* 2005).

### 6.1 Types of searches

BVG allows two basic types of searches: *simple* and *advanced* searches. In turn, simple searches can be divided in two types of searches: searches specifying a *fragment of the title* of the work and searches specifying an *exact word or phrase,* which is in the text of the literary work.

Advanced searches allow three types of searches: *exact* (including several words), *pattern*, and *approximate* searches. In the first type, it is possible to enforce the presence of *all* words (AND connector) or the presence of *some* of them (OR connector). The second type allows the search for works which include words *matching a pattern*. For example, if the user writes "sanct", the system will return works containing words like *sanctification, sanctify, sanctimonious* or *sanctimoniously.* Finally, approximate searches allow seeking works that contain words similar to a given word. For example, if a user writes "new", the system will return works containing words like *news*, *mew*, *mews* or *mewl.*

## 6.2 Text retrieval subsystem

As we already noted, the text retrieval subsystem is based on two technologies: an inverted index and several pattern matching algorithms. The global search process can be seen in Figure 7. From the index, we obtain the identifiers of the literary works containing the searched pattern/s. With those identifiers the system accesses the metadata of the literary works to build the list of retrieved works. The user selects one of these literary works, and then the system shows the list of pages where the pattern is present. To do that, a pattern matching algorithm seeks the first occurrence of the pattern in each page, if it exists. Once the first occurrence is found, then that page is reported as one of the pages including the pattern, and the search skips the rest of the page to continue from the beginning of the next page.

After selecting one page, it is traversed by the pattern matching algorithm to highlight the positions of the occurrences of the pattern. Thus, in this case the whole page is always processed.

### *6.2.1 Inverted index*

The inverted index was developed using libraries of the Lucene project (Lucene, 2006). Lucene represents documents like objects. Observe in Figure 8 that a document is an instance of a class that aggregates a list of *fields*. Each field contains a *name* and a *string* of characters. Examples of names could be *title, author,* etc. The text of the work is always one of these fields. The exact list of fields is chosen by the developer for each case.

In our case, each edition of a literary work is represented by an object of the class *Document*. The used fields are the *content* of the work and the *identifier* of the work in the database of the digital library. In order to obtain the *content,* the source files containing the text of the literary work have to be pre-processed to remove the html tags, since our text source files are the html files that are used to present the literary work in the web. After this, the resulting text is converted to lower case.

Once we have the object *document* of all literary works, it is possible to build the inverted index. The index stores for each word in the vocabulary (list of all different words present in the indexed collection of documents) the list of documents where it appears, plus the relative positions of that word inside each document.

The relative positions do not represent the exact physical position of the word, but the order of that word inside the text. The first word in the text is numbered with 1, the second one with 2 and so on. Relative positions are used to seek phrases, where the searched words should be present in a certain order in the text, but they are not useful to know the physical position of the words.

### *6.2.2 Pattern matching algorithms*

Following the search process, once the inverted index provided the literary works (documents) where a word (or phrase) appears, the system presents to the user the sections (pages of the literary work) where the word appears. When the user selects a page to display, the system highlights the searched word/s. To do that, the inverted index is not useful since it only stores which documents contain the searched word/s and the relative position/s of the words. Thus we cannot know the section (page) and the exact position of the occurrences of a word. In order to implement these tasks, we used several pattern matching algorithms.

Pattern matching algorithms search across the text for the occurrences of a pattern. We performed a study including the most important pattern matching algorithms to choose the most suitable ones for our system. We implemented the *Brute Force, Knuth-Morris-Pratt (KMP), Shift-Or Backward Nondeterministic Dawg Matching Algorithm (BNDM)* and an *Approximate String Matching Algorithm* (Navarro and Raffinot, 2002).

Brute force is the simplest one. It follows the text from left to right, character by character seeking the searched pattern. It is valid for patterns of any length, but it is obviously the less efficient.

KMP belongs to the family of *Morris-Pratt.* KMP improves the brute force by avoiding the execution of comparisons that were already done in previous iterations. As brute force, it can search patterns of any length.

BNDM belongs to the Booyer-Moore family. It builds the inverse of the pattern in a nondeterministic automaton. During the search, the algorithm introduces the text from right to left in the automaton until it finds the pattern or a suffix. The complexity of this algorithm is sub lineal (*$O((n \log_{|\Sigma|} m)/ m)$*, being ∑ the alphabet of our vocabulary, *n* the size of the text where we are seeking and *m* the length of the searched pattern). Therefore this technique is very suitable to search for patterns of 32 characters or less; this limitation is due to the use of the automaton, which only has room for 32 characters.

Shift-or algorithm also uses an automaton, but introduces the text from left to right making easier the process. As in the case of BNDM, it is suitable for patterns of 32 characters or less.

The *Approximate String Matching Algorithm* is a modification of a normal pattern matching algorithm that uses the Levenshtein length (Levenshtein, 1965) as coincidence criteria. Levenshtein length measures the number of insertions, deletions and substitutions which are necessary to convert one word into another.

To compare the algorithms, we used a collection of texts from the TREC conference. We searched for patterns of different length, a non existent word (housa) and a very frequent word (any). Using a 2 GHz Pentium IV with 512 MB, Table 1 shows the results.

| | Short word | Big word | Large phrase | Wrong word |
|---|---|---|---|---|
| **Brute force** | *28.63* | *25.47* | *24.93* | *23.83* |
| **KMP** | *20.73* | *21.80* | *21.43* | *20.23* |
| **Shift – Or** | *19.76* | *19.23* | - | *21.40* |
| **BNDM** | *11.97* | *3.67* | - | *8.80* |
| **Approximate** | *455.20* | *645.83* | *1581.23* | *510.40* |

**Table 1. Average search time in milliseconds**

The most efficient is BNDM due to its sub lineal complexity, while all the rest has a complexity *O(n)*. The problem is that BNDM is only valid for patterns of 32 characters or less. Therefore we used BNDM for seeking words, prefixes and phrases of up to 32 characters and KMP for bigger patterns. Brute force and shift-or are not used.

## 6.3 Empirical results

To validate the efficiency of the text retrieval subsystem we performed a study using a corpus of 26.9 MB. The index was built in 2 minutes and 25 seconds and occupied 6.78 MB (25.20% of the corpus size). The experiment was done in a 2 GHz Pentium IV with 512 MB. Table 2 shows the results. Searches of simple words (the most common) are very efficient, less than one millisecond. The search for phrases takes time proportional to the size of the phrase, but in a sub lineal progression. The approximate search is the slowest due to the necessity of computing the distance.

| **Search type** | **Simple words** | **2 words phrases** | **4 words phrases** | **Wild card searches** | **Approximate searches** |
|---|---|---|---|---|---|
| **Average time (ms)** | *< 1* | *44.75* | *62.63* | *81.2* | *542.95* |

**Table 2. Average time for 20 random searches**

# 7. USAGE AND IMPACT

We present here the data of usage and impact of BVG dated June 2006. The BVG received from its launch in 2002 more than 1,300,000 visits. We extracted these data from the web server logs, which include for each request, the URL and time. With respect to countries, we used the domain to identify the country. The *.com, .edu* and *.org* domains were not considered to discriminate the country of the visit, since such domains do not represent visits from USA in many cases. To analyze the time of connection, our scripts observe the time of consecutive requests from the same URL.

## 7.1 Visits per section

The activity in the *catalogue of works and authors* reflects the success of the digital library. The author's web pages were visited 1,318,027 times and the index cards of literary works 2,395,931 times. With respect to the literary works which are available for reading, listening or viewing, we have the following information: the videos were accessed 14,774 times (we have to take into account that currently the BVG has only 17 videos, although we are about to add new ones); literary works in audio format were accessed 9,106 times (currently with 35 works) and the text version, more than 225,000 accesses.

Another important section of the BVG is the *new authors section*. It received 228,928 visits and more than 2,787 reader's comments about the literary works introduced by new authors.

### 7.2 Source of the visits

One of the aims of the BVG is to spread Galician culture through the world. BVG visits are not constrained to Galicia and Spain, having visits from many countries. Table shows the number of visits per country.

| *Country* | *Percentage* |
|---|---|
| Spain | 76.18 |
| Brazil | 5.2 |
| Mexico | 4.82 |
| Peru | 2.85 |
| Argentina | 1.96 |
| Portugal | 1.61 |
| France | 0.89 |
| Colombia | 0.85 |
| Chile | 0.61 |
| Italy | 0.53 |
| Great Britain | 0.37 |
| USA | 0.36 |
| Rest of the world | 3.75 |

**Table 3. Visits per country**

It is possible to classify the visits in three types depending on the source country (excepting Spain). First, Latin American countries like Argentina, Mexico, Peru, Colombia or Chile; these countries received many Galician emigration in the last half of the 19th century and the first half of the 20th century, thus many descendants of those emigrants are still interested in Galician culture. Second, countries where the first language is Portuguese, this is because Galician and Portuguese are very similar. As we already noted, they were the same language in the past. In a third group, there is a mixture of countries; France and Italy are big European countries, maybe housing Galician emigrants, but never in an amount bigger than Germany or Switzerland. Great Britain is now receiving a new type of Galician emigration, which is more qualified. A new type of immigrant with university degree tries to improve his/her curriculum and to learn English. USA can be the same case as Great Britain; good salaries for qualified workers and the language attract this new emigration.

### 7.3 Other useful data

Regarding organizations which accessed BVG, we would like to highlight the high percentage (around 11.21%) coming from secondary schools in Galicia (these visits come from the *edu.xunta.es domain*). This fact, combined with comments from teachers of Galician in different schools, leads us to believe that BVG is a very useful tool to help students improve their Galician and get to know better our authors and their works. It is also important to note that accesses to BVG were not casual visits, because each user visiting BVG accesses at least 4 pages, and stays connected an average of 12 minutes, and that many of them where during work days, on school hours.

## 8. CONCLUSIONS AND FUTURE WORK

Following the information shown in the previous section, as well as direct feedback from the authors who have collaborated with us and the public, we can conclude that the development of BVG was a success, and all our goals were met.

Our idea of building the digital library to help connecting a disperse community like ours, and promote cultural and linguistic values, proved to be correct, as shown by the accesses from throughout the entire world.

BVG also helped the prestige of Galician language and writers, and we are especially proud to build a digital publisher house to help new writers to publish (electronically) their works and to have built a communication channel between writers and their public. Specifically concerning the Galician language, BVG also showed that it was a very useful tool very highly valued by school and high school teachers in Galicia.

As future work, we are trying to add more resources to BVG, including didactic guides and WebQuests (Dodge, 1995), so it becomes even a better tool for Galician language and literature teachers and students. Also in the interest of this community, and to try to overcome the problems of Lesser Used Languages in general, we are trying to help the EMILL project include more researchers and languages in their Web page.

## 9. REFERENCES

Baeza-Yates, R. and Ribeiro-Neto, B. (1999), *Modern Information Retrieval*, Addison-Wesley, New York, NY.

Borgman, C. (1999), “What are digital libraries? Competing visions”, *Information Processing and Management*, Vol 35 No 3, pp. 227-243.

Brisaboa N, Duran M, Penabad M and Places A. (2000), “A collaborative framework for a digital library”, *Proceedings of the VI International Workshop on Groupware (CRIWG’2000), Madeira, Portugal*, IEEE Computer Society Press, Los Alamitos, CA, pp. 104-111.

Brisaboa, N. R., Paramá, J. R., Penabad, M. R., Places, A. S., Rodríguez, F. J. (2002), “BVG. La Biblioteca Virtual Gallega”. José Hilario Canós, Purificación García (Ed.). *Actas de las III Jornadas de Bibliotecas Digitales (JBIDI'2002)*, pp. 163-172. El Escorial, Madrid (Spain).

Delos, (2007), “A Reference Model for Digital Library Management Systems”. Available http://www.delos.info/index.php?option=com_content&task=view&id=345&Itemid=.

Dodge, B. (1995), "Some Thoughts About WebQuests", *TheDistance Educator*, Vol 1 No 3, pp. 12-15.

Fernández Pérez-Sanjulián, C. (2003), “As mulleres na Biblioteca Virtual Galega”, *Actas do VII Congreso da Asociación Internacional de Estudios Galegos: Mulleres en Galicia. Galicia e os outros pobos da Península,* Barcelona (Spain) (In press).

Gonçalves, M., Fox, E., Watsom, L. and Kipp, N. (2001), *Streams, structures, spaces, scenarios, societies (5S): A formal model for digital libraries*. Technical Report TR-01-12, Virginia Tech, Blacksburg, VA.

Gonçalves, M., Mather, P., Wang, J., Zou, Y., Luo, M., Richardson, R., Shen, R., Xu, L. and Fox, E. (2002), “Java MARIAN: From an OPAC to a modern digital library system”, *Proceedings of the 9th International Symposium on String Processing and Information Retrieval (SPIRE 2002)*, *Lecture Notes in Computer Science,* Vol. 2476, Springer-Verlag, Berlin, pp. 194-209.

Lesk, M. (1997), *Practical Digital Libraries: Books, Bytes, and Bucks*, Morgan Kaufmann Publishers, San Mateo, CA.

Levenshtein, V. I. (1965), “Binary codes capable of correcting spurious insertions and deletions of ones”, *Problems of Information Transmission,* Vol 1, pp. 8-17.

Lucas W. (2001), “Search engines, relevancy, and the World wide Web”, in Goyal A. (Ed.) *Text Databases & Document Management: Theory & Practice,* Idea Group Publishing, Hershey, PA.

Lucene (2006), Available at: http://lucene.apache.org/. Accessed 31 May 2006.

Navarro, G. and Raffinot, M. (2002). *Flexible Pattern Matching in Strings,* Cambridge University Press, Cambridge.

Paramá, J. R., Places, A. S., Brisaboa, N. R., and Penabad, M. R. (2006), “The Desing of a Virtual Library of Emblem Books”, *Software: Practice and Experience*, Vol 36 No 5, pp. 473-494.

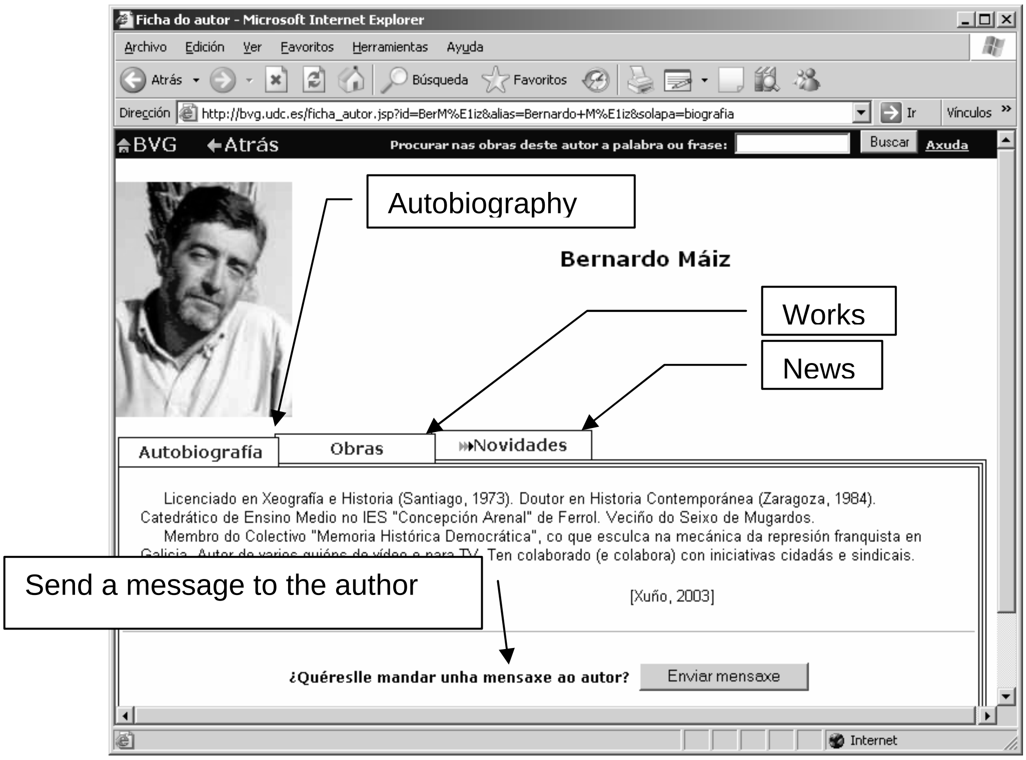
Penabad M., Brisaboa N., Fariña A., Luaces M., and Paramá J. R. (2003), Using Geographical Information Systems to browse touristic information. *Information Technology & Tourism*, 6, 31-46.

Salton, G. (1968), *Automatic information Organization and Retrieval*, McGraw-Hill, New York, NY.

Van de Sompel, H. and Lagoze, C. (2000), "The Santa Fe Convention of the Open Archives Initiative", *Dlib Magazine*, Vol 6 No 2. Available http://www.dlib.org/dlib/february00/vandesompeloai/02vandesompel-oai.html.

Vázquez, E., Places, A. S., Fariña, A., Brisaboa, N. R., Paramá, J. R. (2005), "Recuperación de Textos en la Biblioteca Virtual Galega". *Revista IEEE América Latina*, Vol 3 No 1. IEEE Press.

Zetie C. (1995), *Practical user interface design: Making GUIs work*, McGraw Hill, New York, NY.


Ficha do autor - Microsoft Internet Explorer
Archivo Edición Ver Favoritos Herramientas Ayuda
Atrás Búsqueda Favoritos
Dirección http://bvg.udc.es/ficha_autor.jsp?id=BerM%E1iz&alias=Bernardo+M%E1iz&solapa=biografia
Ir Vínculos
BVG Atrás
Procurar nas obras deste autor a palabra ou frase: Buscar Axuda
Autobiography
Bernardo Máiz
Works
News
Autobiografía
Obras
Novidades
Send a message to the author
Internet

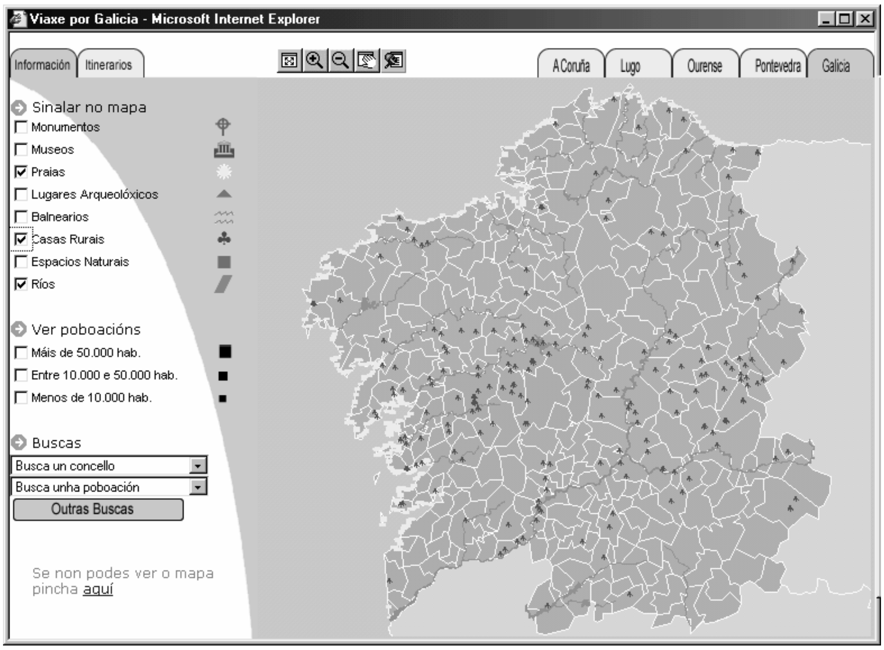
Licenciado en Xeografía e Historia (Santiago, 1973). Doutor en Historia Contemporánea (Zaragoza, 1984). Catedrático de Ensino Medio no IES "Concepción Arenal" de Ferrol. Veciño do Seixo de Mugardos.

Membro do Colectivo "Memoria Histórica Democrática", co que esculca na mecánica da represión franquista en Galicia. Autor de varios guións de vídeo e para TV. Ten colaborado (e colabora) con iniciativas cidadás e sindicais.

[Xuño, 2003]

¿Quéreslle mandar unha mensaxe ao autor? Enviar mensaxe

**Figure 1. Web page for a living author**



Sinalar no mapa
☐ Monumentos
☐ Museos
☑ Praias
☐ Lugares Arqueolóxicos
☐ Balnearios
☑ Casas Rurais
☐ Espacios Naturais
☑ Ríos

Ver poboacións
☐ Máis de 50.000 hab.
☐ Entre 10.000 e 50.000 hab.
☐ Menos de 10.000 hab.

Se non podes ver o mapa pincha aquí

**Figure 2. Virtual tour through Galicia**

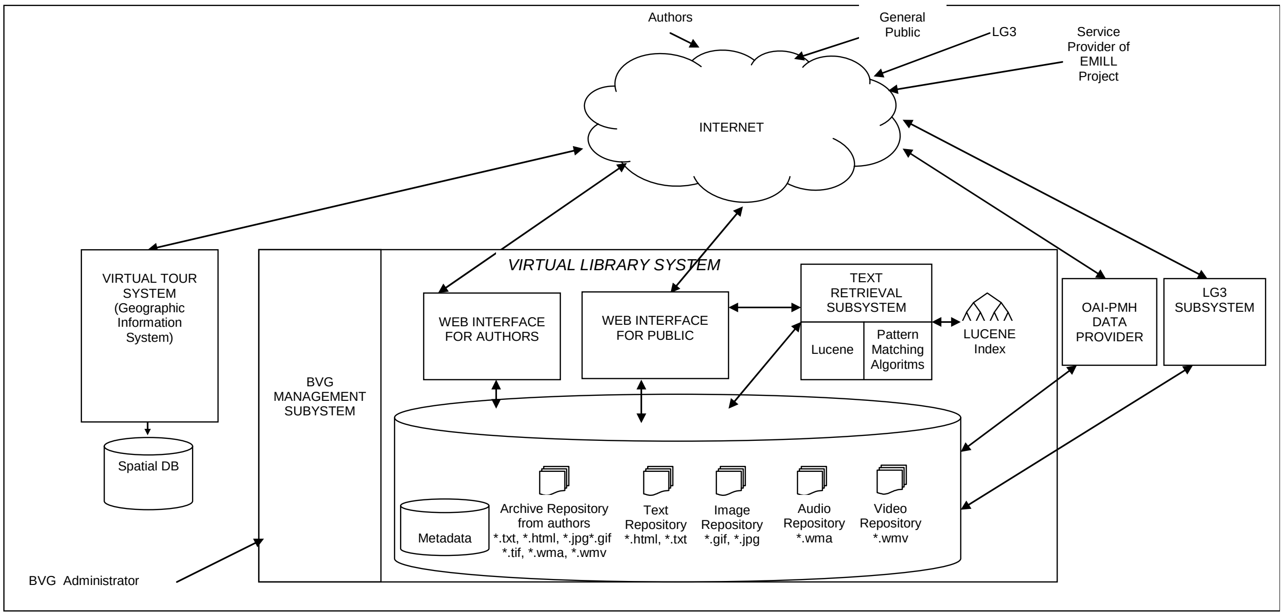

Authors
General Public
LG3
Service Provider of EMILL Project
INTERNET
VIRTUAL TOUR SYSTEM (Geographic Information System)
Spatial DB
BVG MANAGEMENT SUBSYSTEM
VIRTUAL LIBRARY SYSTEM
WEB INTERFACE FOR AUTHORS
WEB INTERFACE FOR PUBLIC
TEXT RETRIEVAL SUBSYSTEM
Lucene
Pattern Matching Algoritms
LUCENE Index
OAI-PMH DATA PROVIDER
LG3 SUBSYSTEM
Metadata
Archive Repository from authors *.txt, *.html, *.jpg*.gif *.tif, *.wma, *.wmv
Text Repository *.html, *.txt
Image Repository *.gif, *.jpg
Audio Repository *.wma
Video Repository *.wmv
BVG Administrator

**Figure 3. Architecture of BVG**

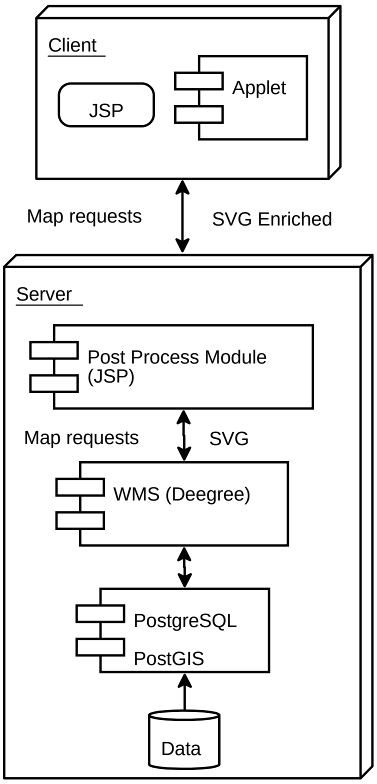
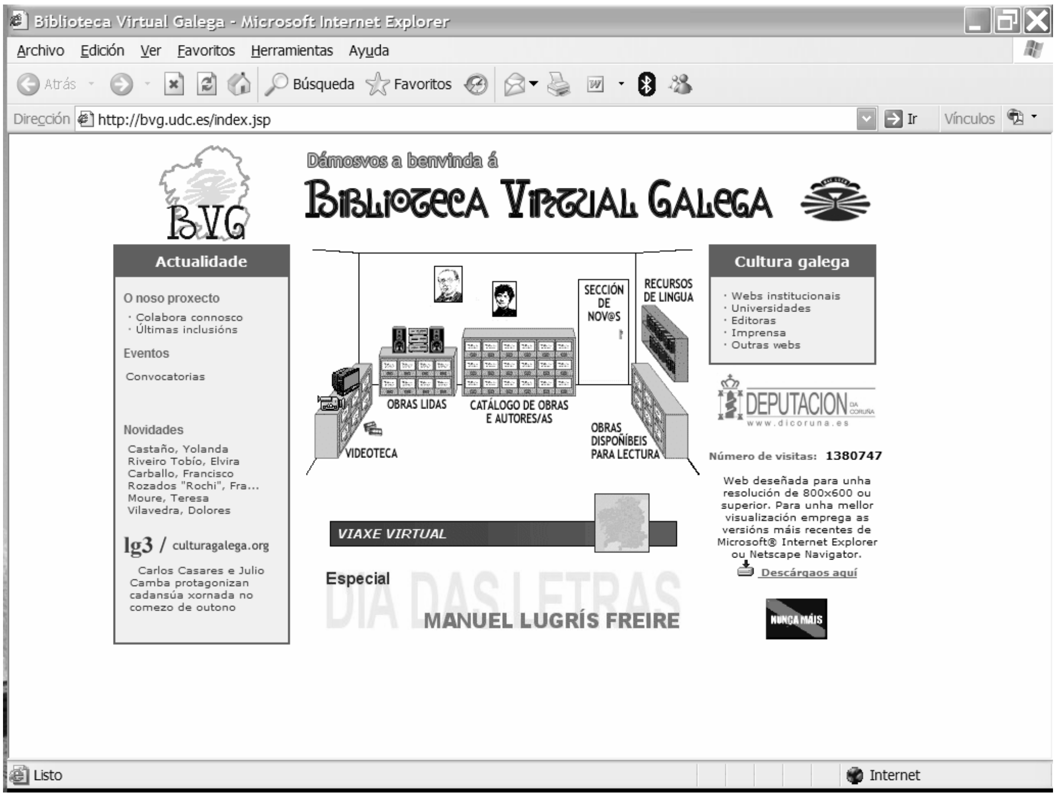


**Figure 4. Architecture of the virtual tour**



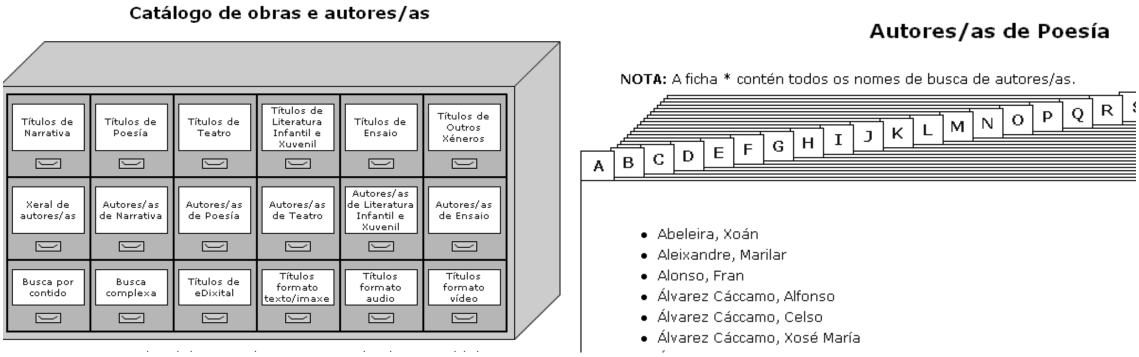
Castaño, Yolanda
Riveiro Tobío, Elvira
Carballo, Francisco
Rozados "Rochi", Fra...
Moure, Teresa
Vilavedra, Dolores

Carlos Casares e Julio Camba protagonizan cadansúa xornada no comezo de outono

Número de visitas: 1380747

Web deseñada para unha resolución de 800x600 ou superior. Para unha mellor visualización empregra as versións máis recentes de Microsoft® Internet Explorer ou Netscape Navigator.

**Figure 5. The Galician Virtual library**

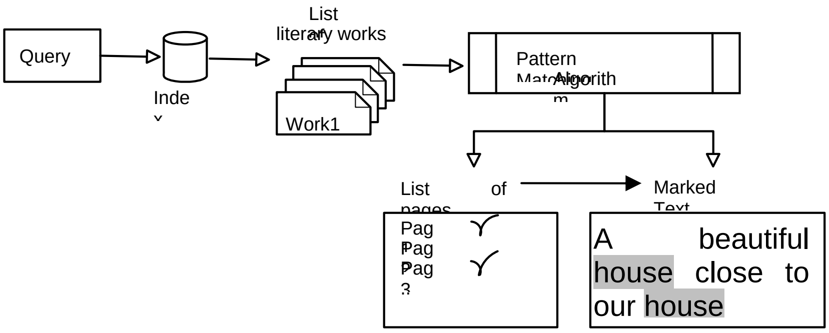

Catálogo de obras e autores/as
Autores/as de Poesía
A B C D E F G H I J K L M N O P Q R S

| | | | | | |
|---|---|---|---|---|---|
| Títulos de Narrativa | Títulos de Poesía | Títulos de Teatro | Títulos de Literatura Infantil e Xuvenil | Títulos de Ensaio | Títulos de Outros Xéneros |
| Xeral de autores/as | Autores/as de Narrativa | Autores/as de Poesía | Autores/as de Teatro | Autores/as de Literatura Infantil e Xuvenil | Autores/as de Ensaio |
| Busca por contido | Busca complexa | Títulos de eDixital | Títulos formato texto/imaxe | Títulos formato audio | Títulos formato vídeo |

NOTA: A ficha * contén todos os nomes de busca de autores/as.

- Abeleira, Xoán
- Aleixandre, Marilar
- Alonso, Fran
- Álvarez Cáccamo, Alfonso
- Álvarez Cáccamo, Celso
- Álvarez Cáccamo, Xosé María

**Figure 6. Catalogues of works and authors**



A beautiful house close to our house

**Figure 7. Search system process.**

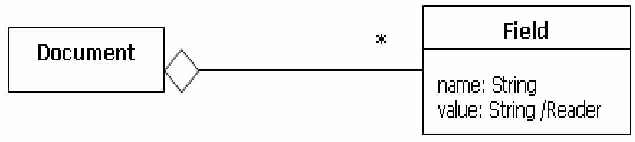


**Figure 8. Representation of a document (literary work in our case) in Lucene.**