Internet Search Engines and OPACs: Getting the best of two worlds

JOSÉ BORBINHA[1]; NUNO FREIRE[2]; MÁRIO SILVA[3]; BRUNO MARTINS[4]

[1] National Library of Portugal
Campo Grande, 83, 1749-081 Lisbon, Portugal
jose.borbinha@bn.pt
[2] INESC-ID
Rua Alves Redol, 9, Apartado 13069, 1000-029 Lisbon, Portugal
nuno.freire@bn.pt
[3,4] LASIGE - Faculty of Sciences of the University of Lisbon
Campo Grande, 1749-016 Lisbon, Portugal
mjs@di.fc.ul.pt; bmartins@xldb.di.fc.ul.pt

This paper reports a pragmatic experiment, consisting on the integration of a Web search engine with a traditional library's catalogue. PORBASE is the Portuguese bibliographic union catalogue. It holds more than one million of bibliographic and nearly half a million of authoritative structured records about authors, families, organizations and subjects. tumba! is a search engine specialized in indexing web pages in Portuguese language. It indexes actually more than one million of pages, found potentially everywhere in the web. Since the early days of the Web that we have been assisting to a discussion about the pros and cons of web indexes built automatically by search engines as an alternative to traditional human-built databases, as the library's catalogues are. The main argument against search engines has been that they might be very effective in indexing the surface skins of the web, but they miss the richest contents hidden behind the more complex web sites and databases, the so called "deep web". On the other side their defenders point the relatively low cost of those solutions, which make it possible to provide good services without the costs of manual cataloguing, classification, indexing of the resources. This paper reports about an experiment to explore the best of both worlds, by using PORBASE in conjunction with tumba!. We extracted from PORBASE the major index of the authorities (names of persons), along with the frequency of each entry, and integrated it within the tumba! search engine. We also developed a very simple HTTP-based interface, ESPONJA, which tumba! can use to launch queries in PORBASE using the authoritative descriptions as arguments. In the scenario of this experiment, a user performing a search session using tumba! will be able to search also in PORBASE using the authoritative form of the search terms without knowing them in advance. The process starts when tumba! presents its own results complemented with suggestions (tips) to search in PORBASE. If the user chooses to launch a search there, a new window is opened showing the results of a search for the term. After that it is up to the user to continue interacting with PORBASE, to explore new results and take advantage of the specific functions of its specialized catalogue. This work aims to be only a proof of concept for this approach. Next steps will comprise the tuning of the algorithms to process the authority information in tumba!, and the formalization of the ESPONJA interface toward a stable, open, generic and reliable interface for interoperability with BN's specialized bibliographic systems.

Keywords: Digital Libraries; Information Retrieval; Search Engines; World Wide Wed; Bibliographic Databases; OPAC; Metadata; Interoperability; Authority Files; UNIMARC

INTRODUCTION

With the increasing importance of the World Wide Web (Web) in our society, we assisted to the emerging of a new kind of valuable tools to search for information in that space. Several Internet search engines got special notoriety, such as Altavista and Google, making them valuable resources for all the Internet users. Since the early days of those tools that we have been assisting also to a generalized discussion about their pros and cons as alternatives to traditional human-built databases, as they are the library's OPAC (On-line Public Access Catalogues). The main argument against the use of search engines has been that they might be very effective in indexing the surface skins of the web, but they miss the richest contents hidden behind the more complex web sites and databases, the so called "deep web". On the other side the defenders of those services point the relatively low cost of their solutions, which make it possible to provide good search services without the need and the expensive costs of cataloguing, classification, indexing of the resources, especially if performed manually (as it happens in the libraries catalogues).

Libraries are organizations with relatively stable processes, standards and procedures, especially in what concerns the creation and use of their information resources. Cataloguing rules and OPACs are practices and concepts nowadays very well established. The introduction of the computer in the library started with the digital catalogue and the definition of the first standards for bibliographic description. That was followed by the first data communication and interaction services (X.25, TELNET, BBS - Bulletin Board Systems, etc.), providing remote access to the catalogues and library's services. In the late 80's we had the emerging of the personal computer and the CD-ROM, which brought the digitized library, providing now access to also the contents. Finally, we had the Internet and the World Wide Web, with which we are working today. This evolution brought

us to the problem of the definition of the “virtual library”, or in a more common term, of the “digital library” [1]. This became a recent hot topic of discussion, with some demagogy but also with lots of real serious work, both conceptual and technical. It attracted also professionals and communities from outside the traditional library's world, especially from Computer Science and Engineering.

The digital library is now part of the World Wide Web, and a mandatory actor in the Semantic Web [2]. In this scenario, the users expect not only to reach the library from anywhere, but also to reach anything. For examples, users would appreciate very well if they could use interoperable services when searching about web pages and books of and about "José Saramago", the Portuguese writer Nobel Prize in 1998, and especially if that could be done in ubiquitous ways.

In order to be able to offer services of this kind, the library needs to be designed as an interoperable service, available to be used in aggregation with other heterogeneous services [3]. This requires cooperation from generic and specialized libraries and archives, museums, and other classes of services. The ability to automate this interoperability is crucial for its effectiveness, bringing requirements for new classes of interfaces and metadata, defined or simply adopted by those actors. That has been done traditionally by means like Z39.50 [4], complemented recently by new models and solutions involving bibliographic records coded in XML [5], taking advantage of simple structures such as Dublin Core [6], or to provide bulk of records for harvesting by OAI-PMH [7]. This is technology that has been especially conceived by digital libraries communities, sometimes for specific contexts, but for the future we must start thinking in scenarios reusing more generic solutions, such as for example those based on Web Services [8].

The work reported in this paper is a small step in that direction! It aims to demonstrate, as a proof of concept, that it is feasible to conceive simple interoperability solutions to integrate a normal web search engine with a formal bibliographic database, providing immediate valuable added value for the users at a very low cost.

PORBASE

BN, the National Library of Portugal, is a legal deposit library. It is also responsible for the national bibliographic database, PORBASE [9]. This database holds more than one million of bibliographic records, representing information from 150 libraries from all over the country and of various areas and magnitude. Those records are created and maintained by specialized professionals, according with the Portuguese Cataloguing Rules, which are based on the Anglo-American Cataloguing Rules [10], and the coded according the UNIMARC format. BN runs also the program CIP (Cataloguing in Publishing), which receives in advance from the publishers the information about the works expected to be put in the market soon. That information is also registered in the same system supporting PORBASE.

Almost all the records in PORBASE have classification information according with the Universal Decimal Classification system, and an important part of them have also indexing information according with SIPORBASE, a subject heading language. Others specialized indexing languages are also used in some cases, by specialized libraries members of PORBASE, according with the rules of the initiative CLIP (Harmonizing Indexing Languages in Portuguese).

PORBASE holds yet authoritative information about authors, organizations, and subjects, in a total of about half a million of records. The term "authority" means that this information is especially validated. For the authors, for example, when possible it is recorded his/her full name, along with his/her birth and death year, all the known pseudonyms, and all the known alternative forms. Those authority records are linked to the bibliographic information, making it possible to perform retrieving actions with good precision and recall. TABLE 1 presents two examples of authority records for two writers (the main entries for these authorities are in the fields 200, while the alternatives are in the fields 400).

TABLE 1 - AUTHORITY RECORDS FROM PORBASE (SHOWN IN UNIMARC "FORMAT"), OF "JOSÉ SARAMAGO", A PORTUGUESE WRITER, AND "ÉMILE ZOLA", A FRENCH WRITER.

Etiqueta de registo: 00538cx 2200169 45 001 10526 095 ## $aPTBN00007456 100 ## $a19900619apory0103 ba 152 ## $aRPC$bSIPOR 200 #1 $aSaramago,$bJosé,$f1922- 400 #1 $aSaramago,$bJosé 550 ## $aRomancistas portugueses$zSéc. 20 675 ## $a821.134.3 Saramago, José .09vBNzpor 675 ## $a869.0 Saramago, José .09vBNzpor 675 ## $a929 Saramago, José$vBN$zpor 801 #0 aPTbBN$c19900619 830 ## $9PT$aJornalista, ficcionista, poeta e tradutor. Prémio Nobel da Literatura 1998.	Etiqueta de registo: 00492cx 2200193n 45 001 18786 095 ## $aPTBN00013338 100 ## $a19901121apory0103 ba 152 ## $aRPC$bSIPOR 200 #1 $aZola,$bÉmile,$f1840-1902 400 #1 $aZola,$bÉmile 400 #0 $aZola 400 #1 $aZola,$bEmílio 550 ## $aRomancistas franceses$zSéc.19 675 ## $a840 Zola, Emile .09$vBN$zpor 675 ## $a929 Zola, Emile$vBN$zpor 801 #0 aPTbBN$c19901109 810 ## $aDic. de Literatura 830 ## $9FR$aRomancista

TUMBA!

A recent project in Portugal, promoted by the University of Lisbon, has developed tumba!, a specialized search engine. The service tumba! is a Web search engine specially crafted to provide better results to those searching information on the Portuguese Web [11]. This space is defined as the collection of pages from the global web that satisfy one the following conditions: it is hosted on a site under a ".PT" domain; it is hosted in a site under other domain (except ".BR"), written in Portuguese language and with at least one incoming link originating in a web page hosted under a ".PT" domain.

The service tumba! has been running has a free service since November 2002. It has a similar architecture and adopts many of the algorithms of global search engines [12] [13]. Like most of the state-of-the-art search engines, it ranks web pages based on concepts taken from bibliometrics, measuring document's relevance according to the number of documents that reference it, which is the number of documents having a hypertext link to the given document [14]. However, its configuration data is much richer in its domain of specialization. It has a better knowledge of the location and organization of Portuguese web sites (both in qualitative and quantitative terms).

As no other Portuguese organization is systematically crawling and archiving the contents of this Web, the importance of the development of tumba! has more than a simple cultural or commercial interest: it may become strategic both for government entities and national industries as the resource for locating information and services for web users communicating primarily in the Portuguese language or interested in locating resources related to Portugal.

The technology of tumba! is now the result of about two years of development by the members of the XLDB research group of LASIGE _ Large-Scale Information Systems Laboratory of Faculdade de Ciências da Universidade de Lisboa. It implements some the best know algorithms for web search and is pushing the envelope in some domains. A substantial part of the software, for example, was initially developed for other projects. A selective harvesting system for the digital deposit of Web publications for the National Library of Portugal has provided the environment for building the first prototype of tumba! [15].

The architecture follows the pipelined (tiered) model of high performance information systems, as illustrated in Figure 1. Information flows from web publishing sources to end users though successive stages. At each stage a different transformation is performed on the data.

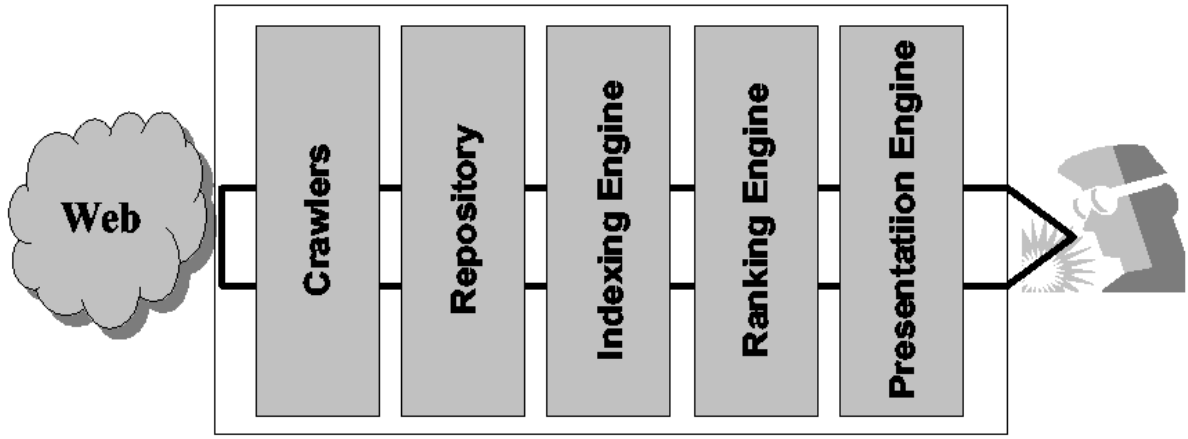

FIGURE 1: TUMBA! PROCESSES THE INFORMATION CRAWLED FROM THE WEB IN A PIPELINE OF SPECIALIZED STEPS.

In the first stage, a distributed crawler, Viúva Negra [16] ("Black Widow") harvests web data referenced from an initial set of domain names and/or web site addresses, extracts references to new URLs contained in that data, and hand the contents of each located URL to the web repository. Viúva Negra detects contents that are replicated in the repository by computing and maintaining MD5 hashes for each stored element, making it suitable to process successive crawls of web spaces without the need to maintain complete replicas of all the crawls. It also features a language detection mechanism based on N-Gram analysis, in order do crawl web pages written in Portuguese from other domains other than ".pt". Note also that tumba! can crawl and index not only HTML files, but also the most popular data types containing text information on the Web, such as Adobe PDF, PostScript, Microsoft Office and Macromedia Flash.

The Web repository is a specialized database management system that provides mechanisms for maximizing concurrency in parallel data processing applications [17]. It can partition the tasks required to crawl a web space into quasi-independent working units and then resolve conflicts if two URLs are retrieved

concurrently. This mechanism is implemented upon a versioning model, derived from the conceptual models of engineering databases.

The indexing engine is a set of services that return a list of references to pages in the repository, matching a given list of user supplied keywords.

The ranking engine sorts the list of references produced by the indexing engine by the perceived relevance of the list of pages. tumba! uses an implementation of the PageRank algorithm, complemented with dynamic scores that weight factors such as the presence of query terms in the title or description meta-tags of web pages. Indexing and ranking are tightly coupled and in tumba!. They are currently performed in Oracle Intermedia7, with ranking computation using custom algorithms [18]. Since Intermedia was designed to index corporate webs, it proved hard to adapt to index large-scale networks as required by tumba!. Therefore the latest prototype version is now using SIDRA [19], a new custom indexing engine.

Finally, the presentation engine receives search results in a device-independent format and formats them to suit multiple output alternatives, including web browsers, mobile phones, PDAs and Web Services. More functionality is being added to this block. For instance a version of tumba!, currently under development, is offering a new interface that will provide a hierarchical organization of results using a clustering algorithm.

THE EXPERIMENT

For the experiment we extracted from PORBASE a report of the major authorities, along with their frequency in the system. For the sake of simplicity, we choose only the authors for this proof of concept. Those authorities were integrated within the tumba! search engine as a new index. We also defined a new HTTP-based interface, ESPONJA, which tumba! can use to launch queries in PORBASE using the headers of the authorities as arguments. The complete architecture is sketched in Figure 2.

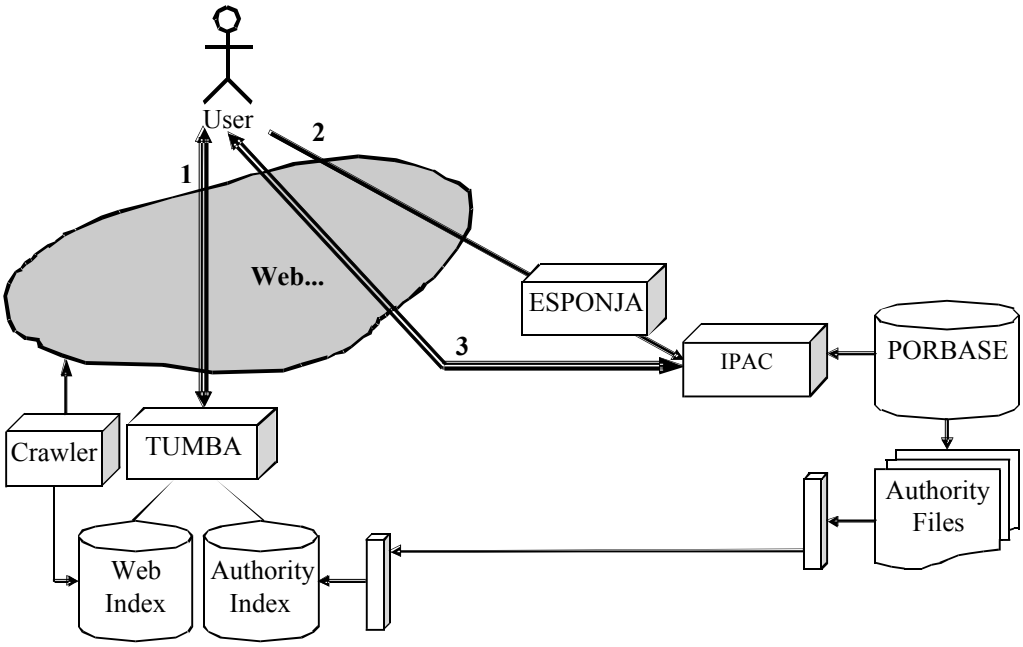

FIGURE 2 – ARCHITECTURE OF THE EXPERIMENT.

The scenario conceived for this experiment works like this:

1. A user starts a session using tumba! with the searching terms. The tumba! engine performs a normal search with those terms in its Web index, and in parallel a search also in the Authority index built from the authorities received from PORBASE. If the term (or terms) matches in this index, an extra side box is presented in tumba!'s page, as a hint, proposing a search in PORBASE.
2. If the user chooses to search in PORBASE following the suggestion proposed by tumba!, he/she just has to click in a link. That send a request to ESPONJA, which processes it and transforms it in a new request to IPAC, the PORBASE's OPAC (PORBASE is accessible by several interfaces; the main OPAC in the moment is the IPAC product, from Dynix, a module of the HORIZON library management system -http://ipac.bn.pt).
3. Finally, the result of the request is presented to the user in a new window, side by side with the original window used to interact with tumba!. This is a very important detail, since now the

user is able, if wanted, to continue the search task in two separated sessions, interacting with two specialized tools, each one tuned for its specific context!

A practical example is shown in the next sequence of images, for the case of the 16th century Portuguese author "Gil Vicente" ("the father of the Portuguese theatre"). In Figure 3 we can observe the page presented by tumba!.

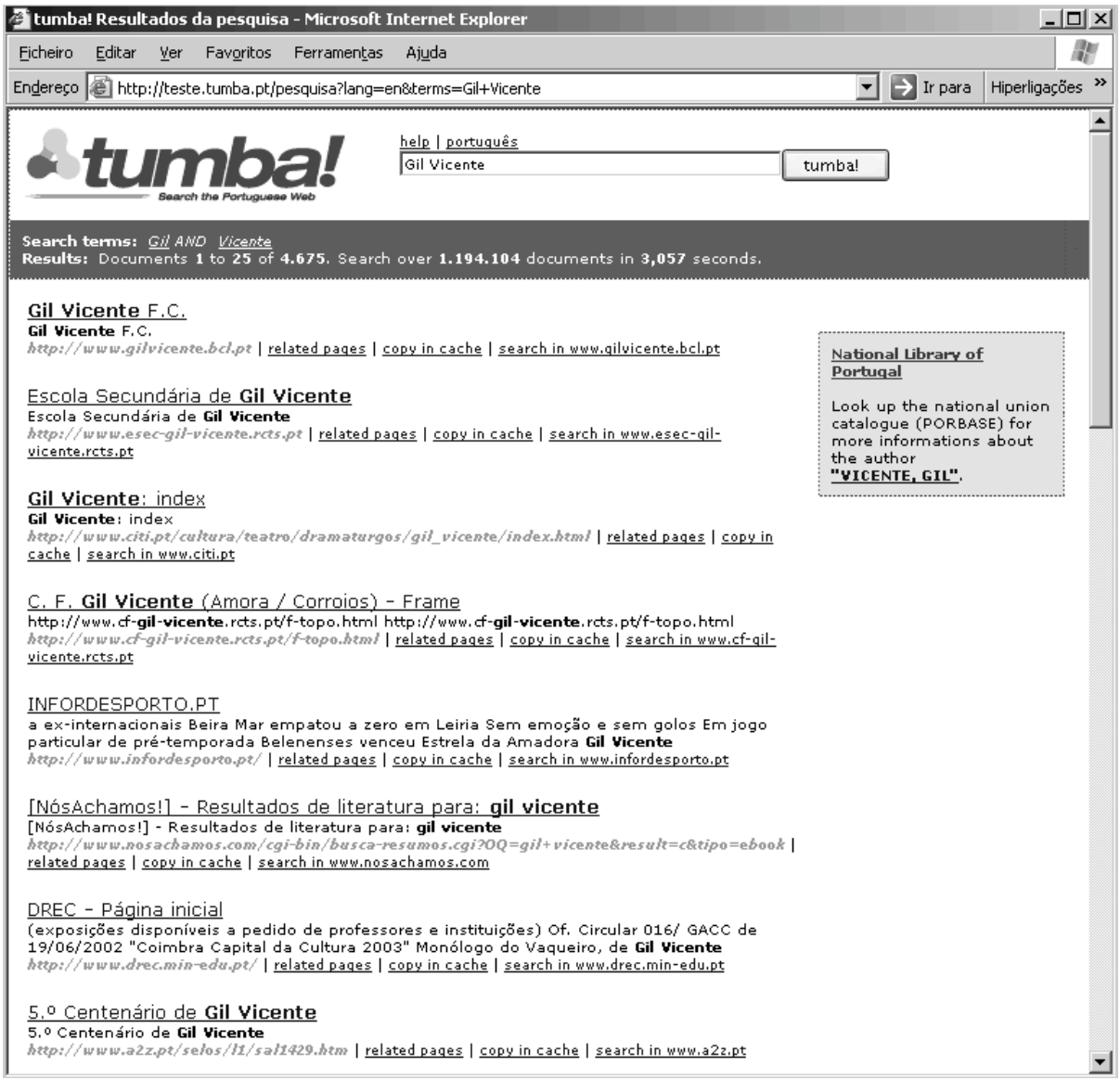

FIGURE 3: TUMBA!'S RESULTS FOR THE REQUEST "GIL VICENTE".

From this set of results we can learn about a football team named "Gil Vicente F.C.", a school named "Gil Vicente", and few pages related with the author "Gil Vicente". The reference to PORBASE is presented in the box in the right side, where the author is shown as "Vicente, Gil". This is because BN asked tumba! to show it this way (see ahead for a more detailed discussion about this issue).

If the user follows this link, he/she will get a reply from IPAC, as shown in the left side of Figure 4 (the window in the back). Because the request was about a specific author, IPAC presents the set of results from a search about works of that author. Now the user can select a specific record, from the total of 402 records that IPAC reports, and see it in detail, as shown in the window in the top, in the right side of the same figure. In this moment the user is already "inside the deep web", interacting with a structured bibliographic database, which is part of a complex information system.

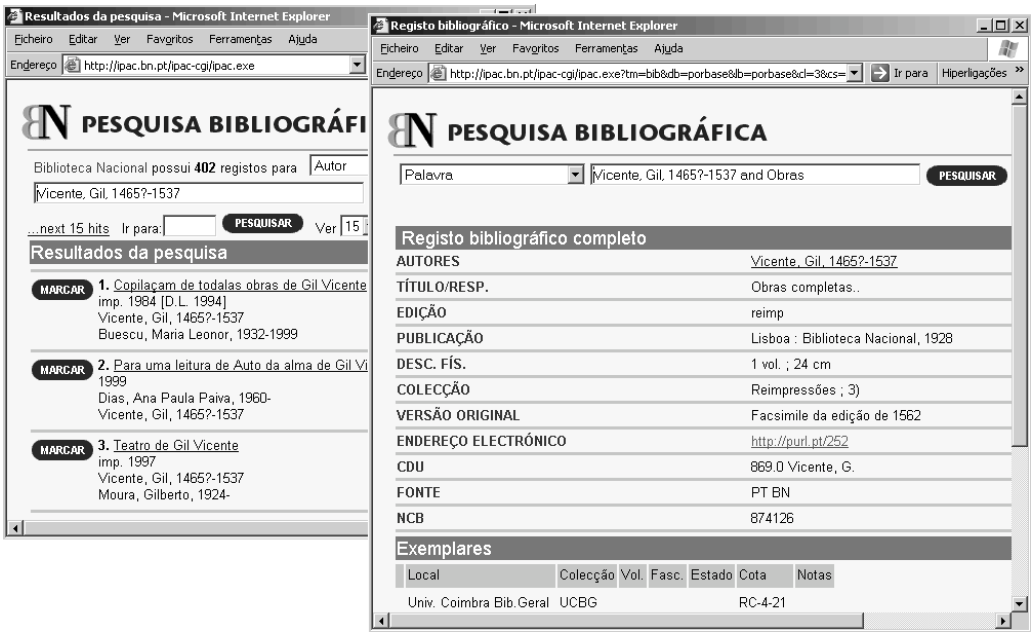

FIGURE 4: LIST OF RECORDS IN PORBASE FOR "GIL VICENTE", AND ONE OF ITS RECORDS IN DETAIL.

Looking carefully to the windows in Figure 4, we can find two interesting details. The first is that the author is now referenced as "Vicente, Gil, 1465?-1537". This means that, according to PORBASE, Gil Vicente might have been born in 1465, and died for sure in 1537. It intends to assure to the user that it is really about the author he/she is looking for.

The second detail is in the top window, in the right side. Here we can see a record of a work published in 1928 with the complete works of the author, which is a facsimile of an original from 1562. In this record we can notice also a field named "ENDEREÇO ELECTRÓNICO", which means "electronic address". This is telling us that such work exists in electronic format, in the address "http://purl.pt/252". In fact, following the link, we get access to a digitized version of the work, available from the National Digital Library (BND - Biblioteca Nacional Digital), as shown in Figure 5. Yet in Figure 4, looking in the bottom of the right window, bellow "EXEMPLARES", we learn too that this work exists also in the collection of the University of Coimbra, as also in others collections of members of PORBASE (which would be visible by scrolling down the window...).

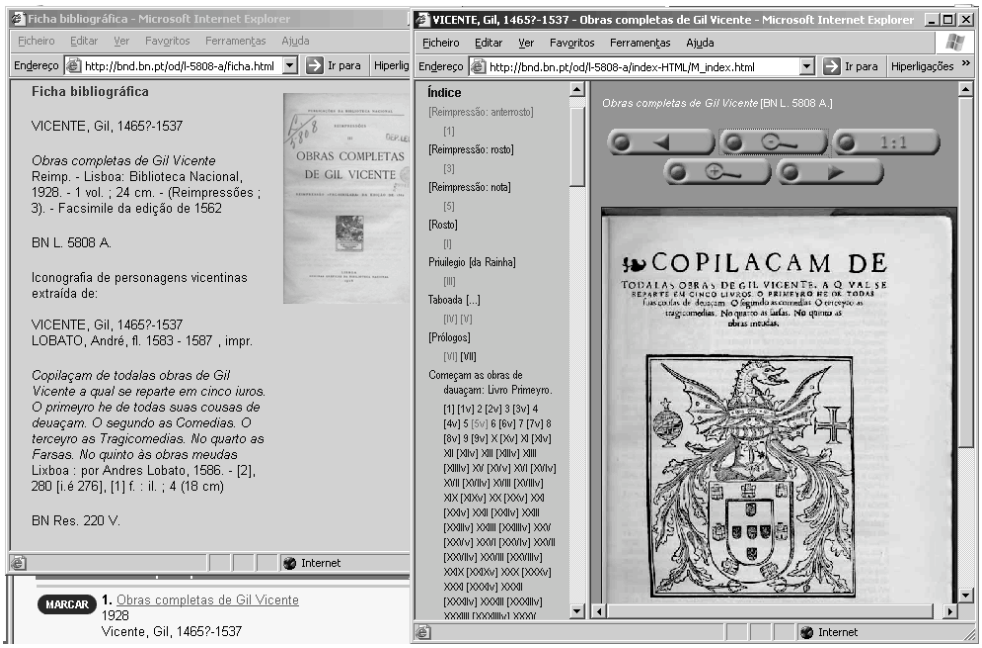

FIGURE 5: THE COMPLETE WORKS OF "GIL VICENTE", AVAILABLE FROM THE NATIONAL DIGITAL LIBRARY.

Resuming the case, we have here an example of how a user was able to get "deep" information from a specialized system, at a cost of a very few "clicks", and mainly browsing after having started a simple interaction with a normal search engine. More than that, it was even possible to close the circle by going back to the "superficial web" and get access to a work from Gil Vicente that is available online! In fact, all of this was possible not only at a low cost for the user, but also at a very low cost for the engineering teams of tumba! and PORBASE!

IMPLEMENTATION DETAILS

The authorities sent to tumba! for this first experiment were only the authors. The information was formatted in triplets in the form **(String-match, Code, Occurrences)**, where:

- String-match: Is the string to be matched by tumba!.
- Code: Is the code of the string in PORBASE's authors index. It is the Authority Control Number in the system.
- Occurrences: Is the number of occurrences for this entity in PORBASE (important for ranking, if there is more than one string-match with the same value.

For our example, the triplet that made it possible had therefore the form **('Vicente, Gil', 10663, 402)**. This made it possible for tumba! to reasoning that:

- The string "gil vicente" matches in the Authority Index, so exists one author in PORBASE with this name...
- ...which code is 10663...
- ...and PORBASE as 402 occurrences for it!

The "Authority Index" in tumba! was implemented as a database that ties a specific query to the text and the hyperlink that is going to be shown to the user when he types that query. For this case it was necessary to implement trivial string processing operations to filter the text in the queries and the information provided by BN, to remove double quotes from queries and converting bibliographic author's entries in the form "Vicente, Gil" to the terms "gil vicente". The total cost of all of this work was a meeting of couple of hours between the two engineering teams, one working day to get the authors from PORBASE and send them to tumba! (very simple programming, but a lot of time for the server to export the data), and less than one working day to create the new index in tumba!. The development of ESPONJA was made at in parallel with this last task.

The conceptual integration of tumba! with ESPONJA was made through a mechanism similar to the advertisement service used in search engines like Google, integrated in the Presentation Layer of the pipeline presented above for the tumba! architecture. A query in the search engine containing the name of an author triggers the appearance of the hint box, with a small description and a hyperlink to the ESPONJA system.

A problem that we had to face was when we had to deal with many records for authors with the same name. As the space available to show the hints is limited, we used the number of works published by the author as a discriminating factor to decide the ranking in this situation. The current implementation presents only one single hint, for the most relevant match, but it can be extended to more. However, probably the best solution will be to have ESPONJA to deal with that...

THINKING AHEAD

Until now, it was not clear in this paper why should we need ESPONJA. In fact, it'd be possible to have tumba! building the final link to IPAC, without the need of a mediator. But we were already thinking ahead, since the main purposes of our work in this project are two: explore ways to expose and disseminate PORBASE; and collect experience to design simple interfaces for interoperability with PORBASE. As a result of that we are now preparing a second step, with the architecture described in Figure 6.

This first experiment was interesting to help us to understand what it might be the best structure to export the authority information from PORBASE, in order to be reused in systems like tumba!. As a result we decided to change that for a new format: **(String-match, String-show, Code, Occurrences),** where:

- String-match: Is the string to be matched by tumba!.
- String-show: Is the string to be shown by tumba! with the link in the hint box when String-match matches. This is important to provide richer semantic information to the user for terms in special scopes, such as transcriptions of codes from the Universal Decimal Classification system, subjects from SIPORBASE, etc. Another important case is when we want to show a specific string or name for more than one String-match (for example, for the possible multiple pseudonyms of one author).
- Code: the code of the string, according the scope of the authority. For the first experiment we used only PORBASE's authors, but now we'll extract more indexes, such as subjects (transcriptions of codes from the Universal Decimal Classification system, subjects from

SIPORBASE, etc. Each index that might match in tumba! will have its own hint box. Also, if there is only one String-match, ESPONJA will redirect the search directly to IPAC. However, if there is more than one String-match, than the value for Code will be zero. ESPONJA will filter that, and instead of redirect the search to IPAC, it will ask it to present its index file, for browsing or search refinement by the user (the index file will show the multiple values for String-match).

- Occurrences: the number of occurrences for the entity in the scope. If there is more than one String-match the value for Occurrences will be zero, so tumba! will be able to take that in account, as an exception for its ranking processing.

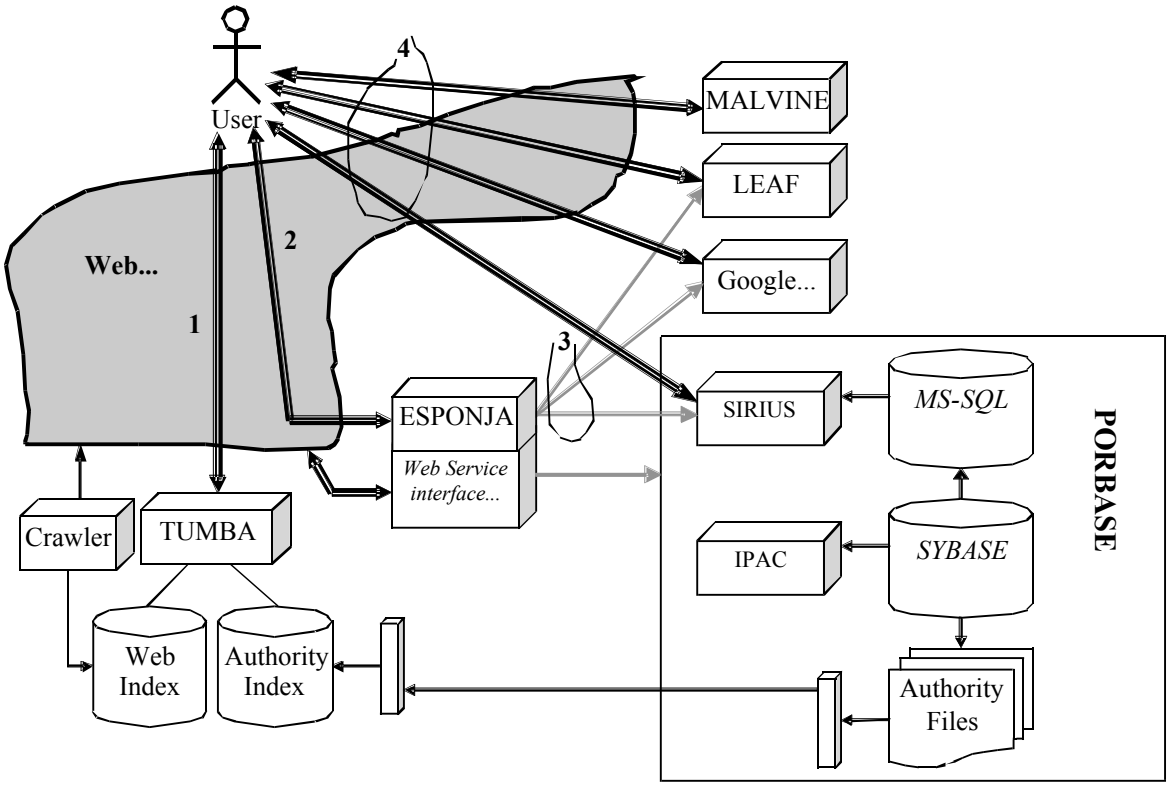

FIGURE 6– A MORE GENERIC ARCHITECTURE FOR THE EXPERIMENT.

This new architecture will make it possible for ESPONJA to redirect the searches to alternatives to IPAC, such as SIRIUS (an alternative replica of PORBASE - http://sirius.bn.pt), or to suggest more sources, such Google or... tumba! (very relevant if the request comes from other system than tumba!). An example of how a case like this can be presented is illustrated in Figure 7.

Another example where ESPONJA will be used is in the LEAF project [20]. LEAF is a European project where BN and a group of European libraries and archives are trying to build together an authority file of authors of manuscripts that they hold. The bibliographic descriptions of those manuscripts are also available trough another interesting system, MALVINE, the result of another previous European project [21]. With ESPONJA, the users will be able to find information about authors LEAF and MALVINE starting from tumba!, without the need to pass through any of the PORBASE systems, if desired.

CONCLUSIONS

The objective of this work was not to define any kind of final interoperability technology. This is just an initial experiment and further work remains to be done. In this trial we simply wanted to explore the relevance of putting together two different concepts. With this experiment we were able to show that both the approaches are not only relevant for information retrieval but also that they can be complementary. Systems and tools like tumba! might not work well as "one-stop shop" solutions to find "deep web" resources. But they can be important gateways to provide simple entry points to reach those resources in logically perceived sequences of steps, which might be very relevant for users that are not specialized in those formal systems. This experiment made it possible to demonstrate the potential of these solutions.

Based on the lessons learned, we intend to define now more formal structures in order to make it possible to expose and export the indexes of the BN's databases to external third-party systems, such as the

Internet search engines and portal, with richer semantics. This will make it possible for systems like tumba! to improve the relevance of their suggestions concerning the BN databases. Finally, another purpose will be the identification of additional databases to be used in this approach, especially from archives and other cultural heritage institutions. The purpose is to show that services like tumba! can be very important to bring those systems and their contents to a new level of awareness.

We can register already examples of the usage of structured information in Internet search engines. Yahoo has been doing that since the beginning, and more recently there is the AdWords Select interface from Google, providing extra search information to sponsored links. Some of these services or experiments are already based on TopicMaps [22] or RDF [23]. Other interesting clues to be explored will be the recent proposal for the new generation of the Z39.50 protocol [24], or the general concept of web services. Our next steps will be to explore those techniques to formalize ESPONJA, making it easier to search in PORBASE from tumba! and other systems (portals, etc.).

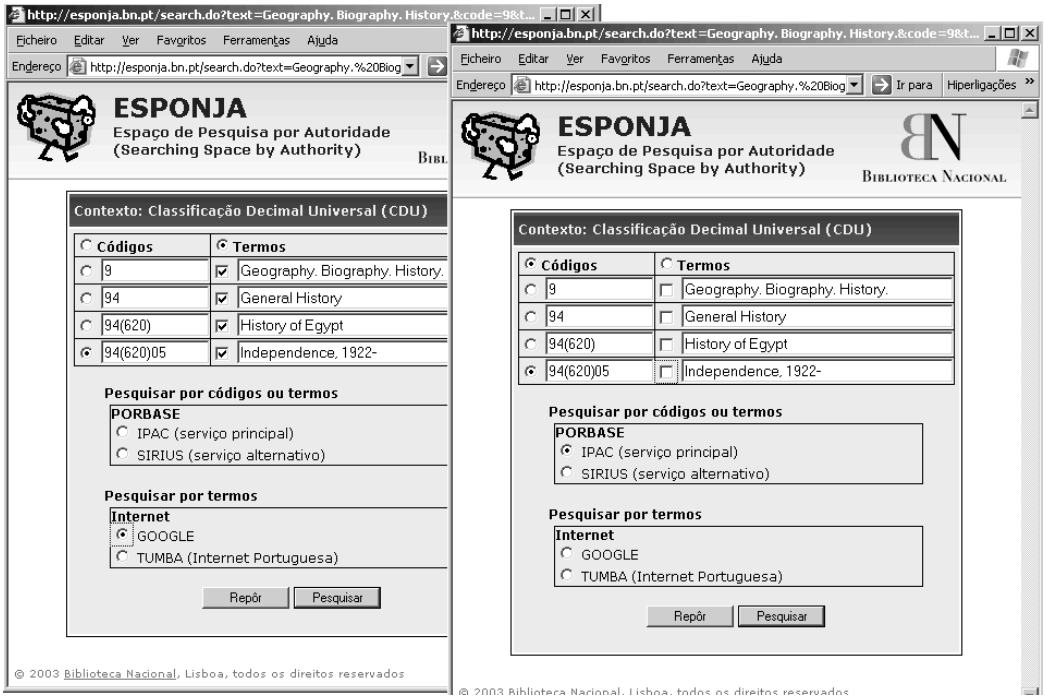

FIGURE 7– ESPONJA AS AN ACTIVE MEDIATOR, FOR AN EXAMPLE OF A SEARCH ABOUT THE INDEPENDENCE OF EGIPT, IN 1922, USING TERMS OR THE UNIVERSAL DECIMAL CLASSIFICATION.

NOTES AND REFERENCES

1. Borbinha, José. The Digital Library - Taking in Account Also the Traditional Library. Elpub2002 Proceedings, VWF Berlin, p.p. 70-80. 2002.
2. W3C. Semantic Web [online]. <http://www.w3.org/2001/sw/>
3. Borbinha, José. Authority control in the world of metadata. Authority Control: Reflections and Experiences. 10-12 February 2003, Florence, Italy [online]. <http://www.unifi.it/universita/biblioteche/ac/relazioni/borbinha_eng.pdf>
4. LOC. Z39.50 Maintenance Agency [online]. <http://www.loc.gov/z3950/agency/>
5. W3C. Extensible Markup Language (XML) [online]. < http://www.w3.org/XML/>
6. DCMI. Dublin Core Metadata Initiative [online]. <http://www.dublincore.org>
7. OAI. Open Archives Initiative [online]. <http://www.openarchives.org/>
8. W3C. Web Services Activities [online]. <http://www.w3c.org/2002/ws/>
9. PORBASE. Base Nacional de Dados Bibliográficos [online]. <http://www.porbase.org>
10. Joint Steering Committee for Revision of Anglo-American Cataloguing Rules. Anglo-American Cataloguing Rules [online]. <http://www.nlc-bnc.ca/jsc/>
11. LASIGE. tumba! - Temos Um Motor de Busca Alternativo [online]. <http://www.tumba.pt>
12. Arasu, Arvind and Cho, Junghoo and Garcia-Molina, Hector and Paepcke, Andreas and Raghavan, Sriram. Searching the Web. ACM Transactions on Internet Technology. 2001
13. Kobayashi, M. and Takeda, K. Information Retrieval on the Web. ACM Computing Surveys, vol. 32, no. 2, pp. 144--173. 2000.

14. Brin S. and Page L. The Anatomy of a Large-Scale Hypertextual Web Search Engine. Proceedings of the Seventh World Wide Web Conference (WWW7), Brisbane, also in a special issue of the journal Computer Networks and ISDN Systems, Volume 30, issues 1-7. 1998.
15. Noronha, Norman, and Campos, João P. And Gomes, Daniel and Silva, Mário J., and Borbinha, José Luís. A Deposit for Digital Collections. Proceedings of the European Conference on Digital Libraries, ECDL, pp. 200-212. 2001
16. Gomes, Daniel and Silva, Mário J. Tarântula - Sistema de Recolha de Documentos da Web. CRC'01 - 4ª Conferência de Redes de Computadores. 2001
17. Campos, João. Versus: a Web Repository, Master Dissertation. Faculdade de Ciências da Universidade de Lisboa. Technical report [online]. 2002.<http://www.di.fc.ul.pt/tech-reports>
18. Costa, Miguel and Silva, Mário J. Ranking no Motor de Busca tumba!. CRC'01 - 4ª Conferência de Redes de Computadores, Covilhã, Portugal. 2001 (in Portuguese).
19. Costa, Miguel (2003). SIDRA: Web Indexing and Ranking System. Master Dissertation, Faculdade de Ciências da Universidade de Lisboa, (in preparation).
20. LEAF. Linking and Exploring Authority Files [online]. <http://www.leaf-eu.org/>
21. MALVINE. Manuscripts and Letters via Integrated Networks in Europe [online]. <http://www.malvine.org/>
22. TopicMaps.org [online]. <http://www.topicmaps.org/>
23. W3C. Resource Description Framework (RDF) [online]. <http://www.w3.org/RDF/>
24. LOC. ZING, Z39.50-International: Next Generation [online]. <http://www.loc.gov/z3950/agency/zing/>