# Text retrieval software for the Web


## Ernest Abadal



Citación recomendada: Ernest Abadal. *Text retrieval software for the Web* [en linea]. "Hipertext.net", num. 3, 2005. <http://www.hipertext.net> [Consulted: 12 feb. 2007]. .


## 1. Introduction

Text Retrieval Systems (TRS) are a well-known type of program in the sphere of information and documentation, especially as they are designed for the retrieval of text and cognitive documents. The main characterisitcs can be summarized as follows: they have a flexible record model (variable length fields, multiple value fields, etc.), they facilitate access to logs through reverse indexing, contain a varied set of data-recovery features, and are provided with diverse instruments for terminology control. Some of the best known and prevalent systems are CDS/ISIS, FileMaker, Knosys, and Inmagic DB/Text.

Global theories of approach have been developed about them, among which we can highlight that of Sieverts and other Belgian researchers (1991-93), authors of a series of very complete and exhaustive articles that describe the characterisitics of these types of programs, elaborating a typology and presenting a very detailed evaluation of some thirty products. Subsequently, William Saffady twice (1995) (2000) realized an updated approach of TRS. In Spain, various diverse works of a global nature have been published, the most recent of which being a monograph by Abadal and Codina (2005) and *Directorio español de software para la gestión bibliotecaria, documental y de contenidos ( the Spanish software directory for library, document and content management ) (* 2003), which contains descriptive data of 135 computer programs in the field indicated by its title. At a different level, we can point out the gateway CMS-Spain (www.cms-spain.com) , which contains diverse reports and studies of content management programs, among which are included references to text retrieval.

TRS have enabled small and medium-sized organisations to be able to create reference-type document databases allowing users of these centres to locate and consult in depth (this refers to books, magazine articles, photographs or other types of documents).

Lately, these programs we refer to have commercial computer applications (vulgarly named Web gateways) that alow consultation, from a Web browser, of the databases created with them. This enables them to significantly extend the spectrum of potential users of the databases as it is now unnecesssary to utilise local area networks to share the use of these databases, nor travel to the physical location where they are locate d.

The aim of our text is to present precisely the current situation in the market of these applications, to evaluate them, and point to future tendencies. To do this, we will in the first place summarise the basic functions of these Web gateways, and then will evaluate comparatively those which are most widely used in the Spanish market.


The author particularly wishes to thank the co-operation of Lluís Codina, and also that of Jordi Casadellà, César de los Santos y Carlos Valmaseda, for their help in the analysis and evaluation of these programs.


## 2. Publishing databases

Until a few years ago, the producers and distributors of databases (the latter in particular) were used to their products having a specialised nature and, because of this, tthey had a potent business structure. This situation has changed radically with the bloom of the Internet and the development of different easily configurable and adaptable tools that make it possiblefor small and medium sized centres of information and documentation, and even personal users, to be able to to convert themselves into producers and distributors of databases.

Small and medium organisations which have created document databases and which we did not refer to in the last section, are bringiing about a general process of publication of their content on the Web. This means that users only need a browser to be able to access up-to-date records which have, in the majority of cases, the same query and exploitation features that text retrieval systems have when they are consulted locally or through local area networks.

## 2.1. Elements

For this method of access to be possible, it is necessary to have, on the server side, a program or suite of programs that permits the communication between two environments which are, at first glance, incompatible or different: the database managed by the TRS, for one, and, secondly, the Web server, which is what serves the browser users and which is only capable of interpreting html pages sent via the http protocol. These programs are usually called Web gateways as they act as intermediaries between the database records and the html coded data that comes from the query form filled in by a user.

The following diagram shows the basic elements that are involved in this process and its functioning:

**Figure 1. Operating diagram**


User (Browser)

page query result listing

httpd Server (Apache, IIS, etc.)

ASP, JSP, etc.

CGI Program

Database

Next, we will explain in more detail the two elements which are most important to the aims of this text: the Web gateway and the query interface

- **Web gateway**

This is the software that connects the Web server (Apache, IIS, etc.) with the TRS. That is to say, it is a program that is capable of reading and interpreting the requests that are sent from an html form, some of them introduced by the user (i.e. the search criteria) and others corresponding to general parameters (i.e. The location of the program and the database on the server, the display format, the number of documents to view, etc.) They are then executed and the result is transferred to the user in html format (this is the listing with the result).

These programs can carry out diverse protocols or systems of communication with the Web server. The oldest is the CGI protocol, but there also exist the protocols ASP (developed by Microsoft), JSP (from Java, and which is open source code) and currently in preparation, the technology .NET (an evolution of Microsoft's ASP and VisualBasic).

The CGI protocol is used to connect a Web server (httpd) with external programs and works by incorporating, from the web page, a call to an executable file (the CGI program) that is located in the cgi-bin or equivalent directory of a server, and which is capable of processing the data that is sent with the page (that is to say, a query to a database). This operating model is less than satisfactory for many webmasters because it requires the installation of different CGI programs on the server (one for each TRS or specific application that needs to connect to the Web server), and because it is not known if these will be entirely compatible with each other or the quantity of the machine's resources they will use. The performance of ASP or JSP is different, in that the scripts are included in the Web page and are executed in the same server before the page is sent , this being a more robust system because it is better integrated with the Web server. To this extent, if a Web server and a TRS supports this standard technology, they can communicate with each other directly without the need to install a separate CGI program. On the other hand, these systems utilise a standard programming language, the same ASP or JSP, that makes data handling easier.

- **The interface**

The interface is the set of pages that facilitate the performance of the query on the user's behalf and which displays the presentation format of the results. This is built with its own programming language, which is included in the gateway, in CGI's case, or with standard ASP or JSP code, intermixed with html code. It basically features three elements: query forms, display format of results (listing); and display format of the complete document. The benefits that are analysed in the next section are, principally those which help to generate this code, assisted by the user.

In a recent publication (Abadal, Codina, 2005) we have written at length of the analysis of indicators used to make and evaluate database interfaces ; we can also take a look at the the book of M.Carmen Marcos (2004) for general considerations regarding the interface.

# 3. Market

In this section we are going to analyse the principal features of the three Web gateways which have made notable inroads and which have a consolidated place in the Spanish market, followed by a table of evaluation, using some basic indicators.

The evaluation has been undertaken according to the actions that an end user might undertake using only the assistant – the utility that allows the assisted generation of a basic query interface. That is to say, the set formed by a query page and the indicators that generate the listing and / or display the document – without resorting to the use of programming language. The tests were undertaken using the versions indicated in the descriptive notes.

It has to be remembered at this point that gateways are no solution to the limitations that a determined text retrieval system may have. As such, if Knosys for Windows only permits the indexing of the content of the fields word by word and CDS/ISIS or Inmagic, on the other hand, allows also the indexing of groups of words, this will not be able to change the fact that some gateways are better than others. In this way, when the field indexes are shown, in the first instance they can only display, at the most, unique terms, and in the second instance, groups of words can also be viewed.

## 3.1. GenIsisWeb (WwwIsis)

Producer: Pierre Chabert ( http://perso.wanadoo.fr/pierre.chabert/ ), developer of the assistant (GenIsisWeb), and Bireme ( http://www.bireme.br/wwwisis.htm ), which is really the CGI (WwwIsis).

Price: free (GenIsisWeb uses version 3, which is also free).

Examples: Figure 3 includes a test undertaken by the author. You can find real examples managed by WwwIsis, although not necessarily created with Genisis, in Bireme's databases ( www.bireme.br/bases ) and also at other Web sites ( http://www.bireme.br/wwwisis/I/listsites.htm ).

**Figure 2. GenisisWeb Assistant**

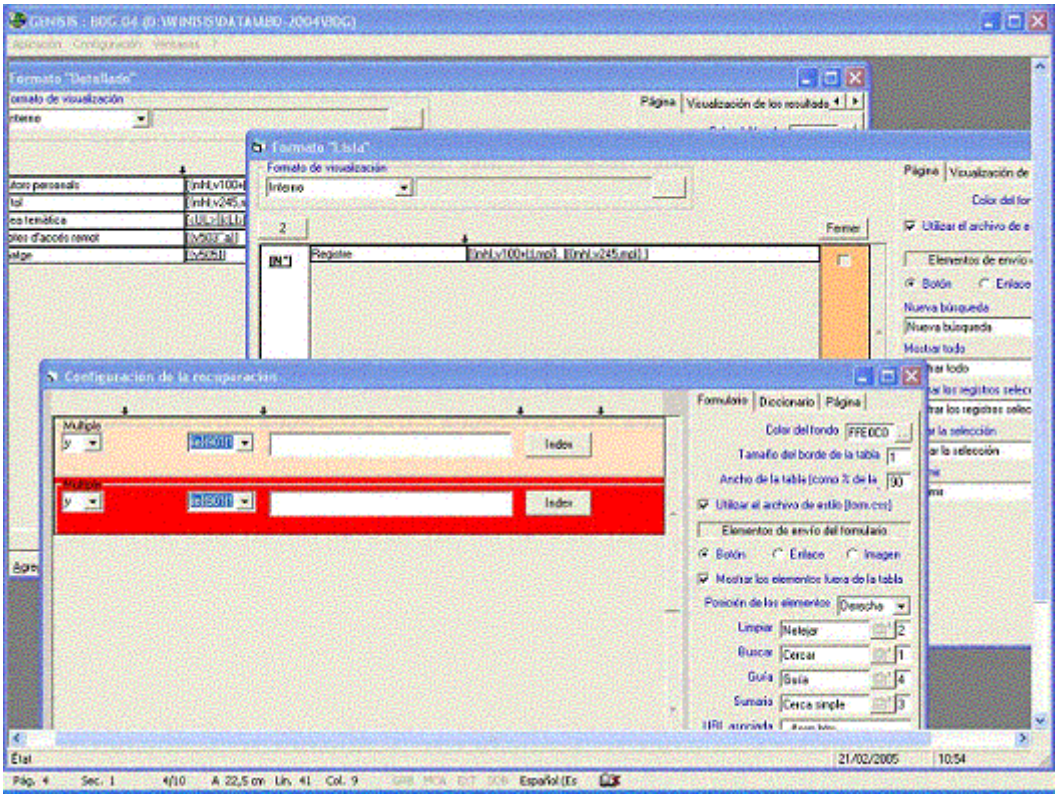
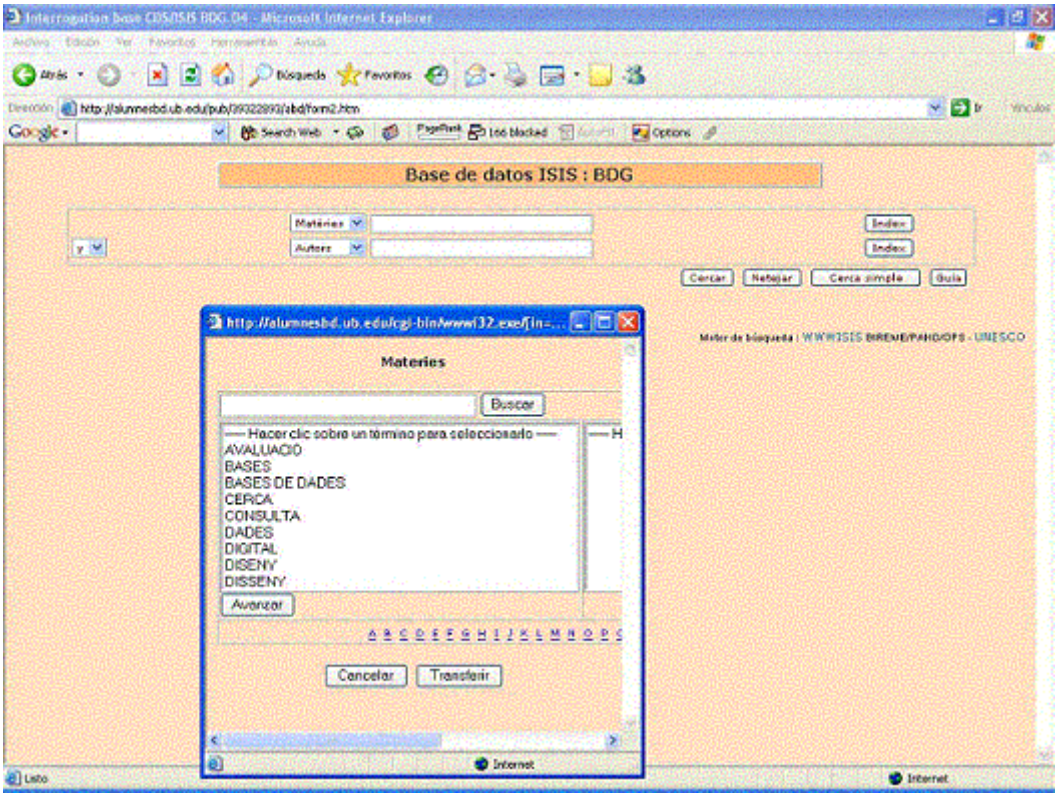


**Figure 3. Test Database created with GenisisWeb**




GenIsis is a very thorough assistant that has been developed by a French programmer, and which incorporates version 3 of WwwIsis (free), a CGI developed by the Brazillian organisation Bireme. This is a free program but it does not meet the requirements of free software (or of open source

code).

The operation of the assitant is relatively simple and allows the preparation of a basic interface in a few minutes. With a little more time, it has also the majority of characteristics required of a standard configuration. The program is structured in three different display screens, each one of which groups the functions that are used in making the query page, the listing or the display of the entire registry.

As far as what is referred to as the preparation of the query page, the following characteristics can be highlighted:

- Display field indexes.

When a field is indexed field it is possible to place a button or link next to each one of the search fields, that allows searching from within the field indexes.

- Use of operators among fields and also from within the same field

A page can be prepared so that the user chooses which is going to be the boolean operator to use among different fields, and also from within the same field (i.e. If in the "Title" or "Summary" field more than one term is indicated).

- Search of different fields using drop-down menu.

It is possible to include a list of search fields in a drop-down menu and even show the indixes associated with each one of these. This feature is especially useful for advanced queries in that it enables all combinations in a small space. This is a feature that the other two programs do not offer.

As far as the listing or document pages are concerned, the enabling of relating records should be highlighted: that which enables the relation of records with image files or other file types, allowing the importing of display indications from CDS/ISIS and through which it is possible to establish automatic relationships among the terms of the database (and in this way, for example, enable the user to navigate between authors or materials that appear in the listing or in the records display). As a counterpart, it has to be noted that the terms which are included within brackets ("<>") are not interpreted correctly, because it is a notation that Winisis uses to indicate the indexing of database terms and so, the assistant confuses them with html tags.

It also has a very similar application that allows the querying of a database on an optical disk .

There exists a review by Sergi Chávez and Noemí Alcázar published in the journal *BiD* (Chávez, 2003).

## 3.2. KnosysInternet

Producer and distributor: Micronet ( http://www.micronet.es )

Examples: Jurisprudence database of the 'Ilustre Colegio de Abogados de Málaga' (the school of law in Malaga) ( http://www.icamalaga.es/juris1.htm ). There are more examples in the 'Clients' section of the old pages about KnosysInternet ( http://www.micronet.es/menu/prof/mki.htm ).

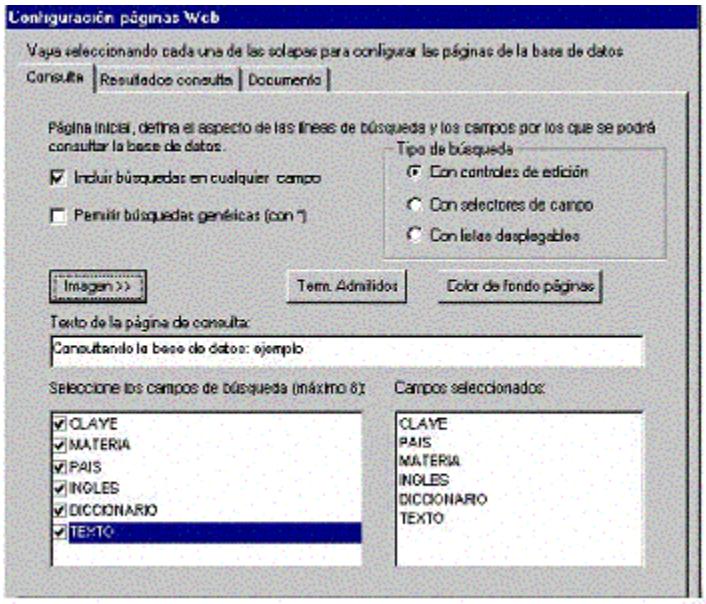
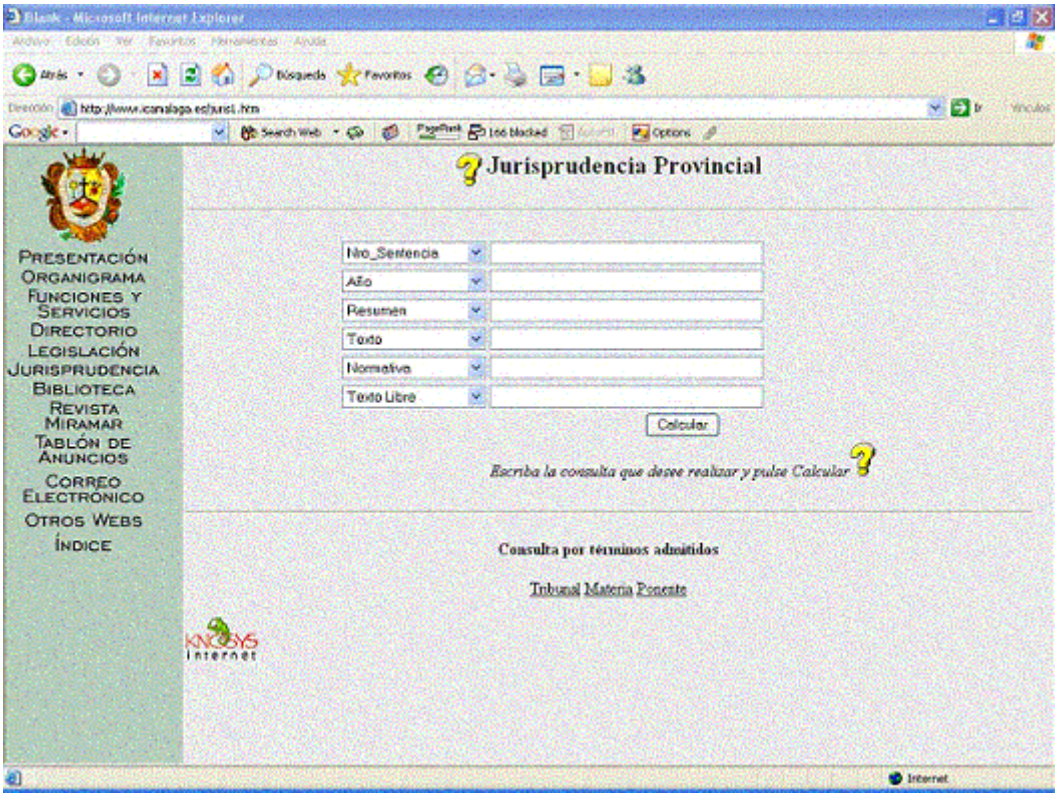



**Figure 4. KnosysInternet Assistant (query page)**

**Figure 5. Jurisprudence database of the 'Ilustre Colegio de Abogados de Málaga'**




Knosys 2004 is the new denomination of the well known document management program

KnosysWindows ( http://www.knosys.net ). This new version follows the client-server model (that which enables online simultaneous updates in the same local network) and has facilities for importing and exporting in XML.

As far as the Web gateway is concerned, it no longer uses the CGI protocol, using JSP instead, although at the moment it does not have an assistant for the automatic generation of database query forms. In the new version of this gateway (which the company has annouced for the end of 2005), an assistant for generating Web pages that is very improved with respect to the version that is reviewed below, and that allows the realisation of of maintenance of the database via the Web, is planned for inclusion.

At any rate, the analysis has been undertaken with the version of KnosysInternet that is currently on the market. This is an assistant that is very easy to use (in three minutes it is possible to have a database available for consultation on the Web) although, as is detailed below, the configuration that is allowed is very basic and a little limited.

With respect to the query page, the following should be highlighted:

- No more than one type of data collection system can be used.

That is to say, if you choose to use text boxes, these cannot be combined in the same page together with drop-down menus, or check boxes, for example.

- The Boolean operator cannot be selected.

The operation that operates between different systems of data retrieval, be they text boxes, drop down menus, etc. This will always be the sticking point. It cannot be configured so that the user can choose the operator that best suits him.

- The field indexes cannot be consulted

The system does not allow the display of terms that form part of each one of the fields. There is a field, however, which shows what Knosys calls "Admitted" and which is none other than a field evaluation listing .This listing, however, has to be introduced manually in a text record that Knosys has prepared.

On the positive side, we have to highlight the incorporation of a statistics module that enables the control of completed queries (it indicates ip address, date, query) and also does a tracing of the general administrative operations that are undertaken with the database (open, close, etc.). This data may be very useful, at a later date, to contribute to the evaluation of use of the supported databases.

Appraisals of this program can be found in the articles of José V. Rodriguez (1998) and of E. Abadal y R. Martínez (2000).

## 3.3. WebPublisher

Producer: I nmagic ( http://www.inmagic.com )

Distributor: Doc6 ( http://www.doc6.es )

Price: €10,000-12,000 (Web with CS Workgroup), and from16.000 to 45.000 euros (for CS standard or Enterprise)

Version: 7 (in Spanish), 8 (in English)

Example: University of South Africa: A select online bibliography of South African history

(SOBiBSAH) ( http://bibinf.unisa.ac.za/dbtw-wpd/textbase/history/menu4.htm ). A list of examples can be consulted at: http://www.andornot.com/links.asp

**Figure 6. Webpublisher Assistant (from DB/Text)**

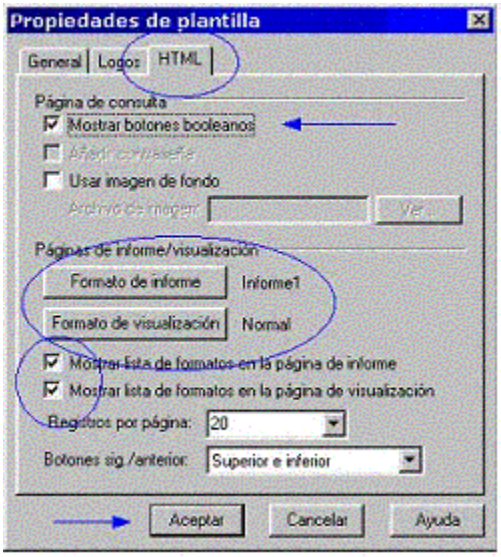
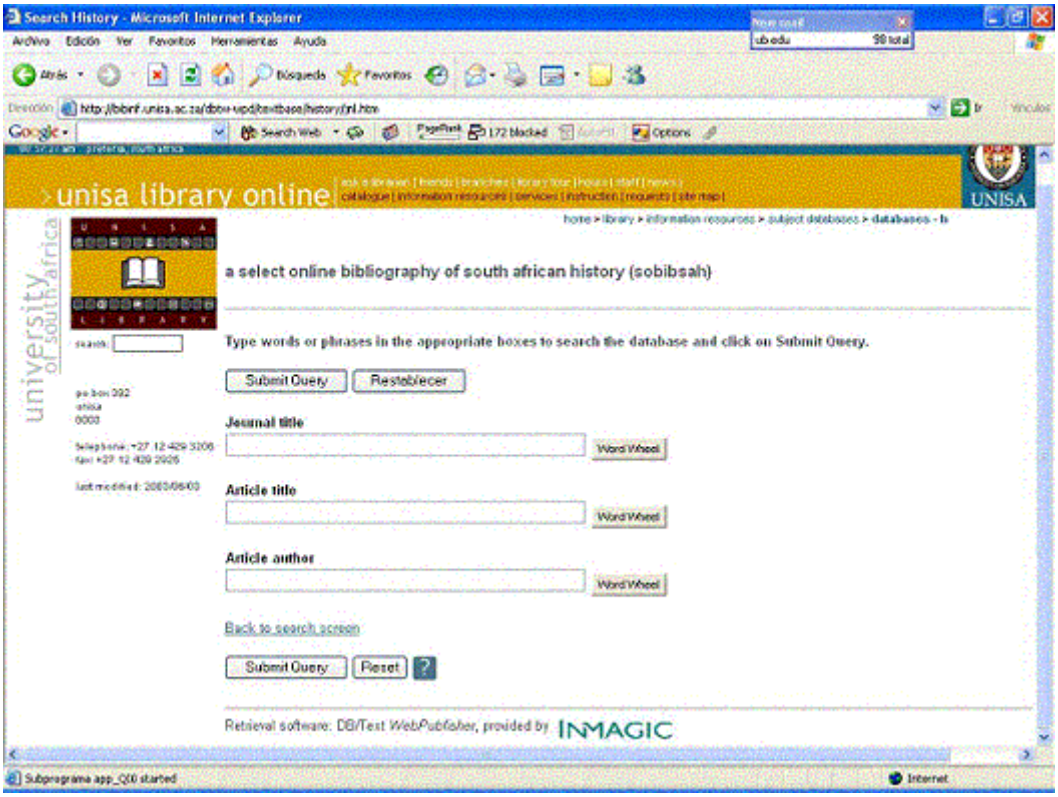



**Figure 7. SOBiBSAH Database**




This is a very comprehensive program. The functioning structure comes from the page design that is undertaken in DB/Text and that is later exported to WebPublisher (in xml). With this, the tasks the CGI gateway really has to perform are very small.

You can indicate in the DB/Text templates what will be the display characteristics fore each field on the Web, and can also specify some elements for the entire template (i.e. on the query page you my or may not indicate if it is going to use boolean operators, to choose the number of records, etc.).

In what is referred to as the query page, the following aspects can be highlighted:

- Display of indexes.

The Wordwheel tool displays the field indexes, no matter if they are word to word or the different values of entire fields.

- Text boxes for searching.

They can point to more than one field. For example, a text box can be created below the sign "Themes" that will search in the indexes of the fields Title, Subject and Abstract.

- It allows you to choose a number of records and display formats.

- You cannot choose the operator from the same field.

Therefore, if we want to include more than one word inside the text box of the field Journal title or Abstract, the user has to indicate with the corresponding operator (&, /, etc.) the operation that is going to be processed.

- A drop down menu cannot be created with various fields.

It is not possible to create, therefore, an advanced search that allows the combination between fields using only two text boxes, as the drop-down menu cannot include field names and start a search on their content.

With respect to the listing, the feature of using more than one presentation format for data has to be highlighted. It is also possible to incorporate images and adjust them to determined display characteristics (width, height, etc.) as well as establishing hypertext links between values of a determined field (subjects, or authors, for example)

In this evaluation two fundamental aspects that notably differentiate it from other programs need to be highlighted:

- Establishing relationships between databases.

Query forms and / or listings that include fields from different related databases can be prepared.

- Making query forms for data entry.

This is the main innovation of version 7. It is a very interesting and useful function for facilitating maintenance work of the database in a distributed way. On the other hand, it is also possible to consult the check lists that a determined field may have via the Web.

There is also a version that incorporates what is called Content Server, the program that enables information management within relational tables, managed by SQLServer.

## 4. Global valuation

Finally, we are going to present a comparative table and some comments about different aspects of the programs we have analysed.

**Table 1. Principal features and characteristics of the programs**

| | GenIsis | Knosys | Webpublisher |
|---|---|---|---|
| Producer | P. Chabert | Micronet | Inmagic |
| Version | 3 | | 7 |
| Price | Free | Mid | High |
| Consult more than one database | No | No | Yes |
| Global querying on all fields | Yes | Yes | Yes |
| Annotate query in a single field | Yes | Yes | Yes |
| Display field indexes | Yes | No | Yes |
| Choose boolean operators among different fields | Yes | No | Yes |
| Choose boolean operator within the same field | Yes | No | No |
| Use and control of text boxes, drop-down menus, check boxes and radio buttons on the same page | Yes | No | No |
| Select number of records | Yes | Yes | Yes |
| Select display format | Yes | No | Yes |
| Hypertext linking between terms in indexes | Yes | No | Yes |
| Link to external image, text, etc. files | Yes | Yes | Yes |
| Record entry | No | No | Yes |
| Display of check lists | No | No | Yes |
| Search statistics | No | Yes | No |

We find ourselves in front of three consolidated products that have been notably improved in subsequent versions. All of them allow a non-expert user to be able to publish a document database on the Web in a short period of time although they allow a different level of features in each case. It cannot be forgotten, nonetheless, that the majority of options that are indicated by a "No" can be resolved by writing directly in the html pages that make up the query interface, although this is an option that is not within reach of most end users.

Webpublisher is, without doubt, the most comprehensive of the three programs that have been analysed. To the features that enable the generating of a very accurate query interface are added two important benefits, such as the possibility of being able to undertake maintenance work on the database (introduce, modify, erase records) and also the power to relate databases together. In order to obtain similar results, not only Genisis, but also KnosysInternet have to resort to programming as these are options which are vetoed by the assistant. The major drawbacks are, on

the one hand, the price, as its cost is quite or substantially above the other two products and, on the other hand, the need to know DB/Text in depth in order to generate html templates of the query interface.

KnosysInternet, for its part, is a user-friendly program but it has different limitations in order to include features in the interface used exclusively by the assistant. In its favour, we have to point out that the price is substantially more accessible than the previously mentioned software, and also, the inclusion of an interesting statistics module. The appearance of a new assistant is forthcoming, which in the light of the substantial changes effected by Knosys2004 supposes an important qualitative leap.

Finally, Genisis offers an array of similar features to WebPublisher in the query section (excluding data entry), with an ease of use that is similar to the other two programs, and with the great benefit of it being a free program.

## 5. Tendencies

In this section, beginning with the innovations that are incorporated into the programs and from the study of the needs and requirements of users, we will try to indicate what are the subsequent areas of development.

One of the most asked for aspects by users is the possibility of having maintenance functions of the database, not only query options, accessible from the Web. This has been possible for some time but it has not been until the latest version presented by WebPublisher that this set of features configurable from the assistant, has been incorporated. It is supposed that the rest of the programs will also incorporate this in the short term, as Knosys has announced, for example.

Secondly, the integration with text retrieval systems is also going to improve, in such a way that the programs we have analysed are going to develop features in order to also facilitate the management of changes, versions, etc. for the documents generated by an organisation, and so better integrating the external document control that hardly ever changes, and also that of internal documentation, which is exposed to constant changes and to manipulations brought about by different users.

Thirdly, they are also going to explore the use and integration within relational systems. To this extent, for example, WebPublisher has developed a module (Content Server) that is based on SQLServer and allows the administration of the whole system using relational tables.

Finally, there is a move evolving from CGI protocol towards JSP, an open standard which is more robust and more compatible with servers. Knosys has already initiated this path and the rest will probably follow.

## 6. Bibliography


Abadal, Ernest (2002). "Elementos para la evaluación de interfaces de consulta de bases de datos". *El profesional de la información* , sept.-oct., 11: 5, p. 349-360.

Abadal, Ernest; Codina, Lluís (2005). *Bases de datos documentales* . Madrid: Síntesis.

Abadal, Ernest; Martínez, Raúl (2000). "Distribució de bases de dades en el web amb Knosys Internet", *BiD: textos universitaris de biblioteconomia i documentació* , jun., n. 4. ( http://www.ub.es/bid/04abadal.htm ). [Consulted: 1/02/2005]

Baeza-Yates, R.; Ribeiro-Neto, B (1999). *Modern information retrieval* . New York: Addison-Wesley. 513 p.

Chávez, Sergi; Alcàzar, Noemi (2003). "GenIsisWeb: la herramienta para publicar en Internet las bases de datos CDS/ISIS". *BiD: textos universitaris de biblioteconomia i documentació* , dec. 11.
( http://www2.ub.es/bid/consulta_articulos.php?fichero=11chave2.htm ) [Consulted: 27-01-2005].

Codina, Lluís; Abadal, Ernest (1992). "Gestió documental amb microordinadors: característiques, estructura i tecnologia dels sistemes de gestió documental". *Item* , 11, p. 72-100.

*CMS-Spain: portal sobre tecnologías para la información y la colaboración.* ( http://www.ecm-spain.com/home.asp ). [Consulted: 14/02/2005]

*Directorio español de software para la gestión bibliotecaria, documental y de contenidos* . Grupo Activa. SEDIC. Coordinación: Luis Rodríguez Yunta, Carlos Tejada Artigas. Madrid: Consejo Superior de Investigaciones Científicas, 2003.

Gillman, Peter (ed.) (1990). *Text retrieval: the state of the art* . London: Taylor Graham. 208 p.

Kemp, A. (1988). *Computer-based knowledge retrieval* . London: Aslib. 399 p.

Kimberley, Robert (1990). *Text retrieval: a directory of software* . 3rd ed. Aldershot [etc.]: Gower.

Marcos, Mari Carmen (2004). *Interacción en interfaces de recuperación de información: conceptos, metáforas y visualización* . Gijón: Trea.

Nieuwenhuysen, P. (1980). "Criteria for the evaluation of text storage and retrieval software". *The electronic library,* jun., 6: 3, p. 160-166.

Rodríguez Muñoz, José V.; Saorín, Tomàs (1998). "Modelado documental de servicios de información en web". *El profesional de la información* , septiembre, 7: 9, p. 10-18.

Saffady, William (1995). "Digital libray concepts and technologies for the management of library collections: an analysis of methods and costs". *Library technology reports* , may-june, 31: 3, p. 221-230.

Saffady, William (2000). "Text retrieval products for libraries". *Library technology reports* , march-april, 36: 2, p. 7-16.

Sieverts, E.G. et al (1991). "Software for information storage and retrieval tested, evaluated and compared: Part 1 – General introduction". *The electronic library* , 9: 3, p. 145-154.

Sieverts, E.G. et al (1991). "Software for information storage and retrieval tested, evaluated and compared: Part 2 – Classical retrieval systems". *The electronic library* , 9: 6, p. 301-316.

Sieverts, E.G. et al (1992). "Software for information storage and retrieval tested, evaluated and compared: Part 4 – Indexing and full-text retrieval programs". *The electronic library* , 10: 4, p. 195-206.

Sieverts, E.G. et al (1993). "Software for information storage and retrieval tested, evaluated and compared: Part 6 – Various additional programs". *The electronic library* , 11: 2, p. 73-89.

Tenopir, Carol; Lundeen, Gerald (1988). *Managing your information: how to design and create a textual database on your microcomputer* . New York: Neal-Schuman. 226 p.