THE FULL MONTY; OR, THE WHOLE HOG WITH **HERMES**

Gerardo Coello-Coutiño

Computing Department, Institute of Cellular Physiology, Universidad Nacional Autónoma de México. Apartado Postal 70-600, 04510 México, D.F. Mexico. Email: gcoello@ifisiol.unam.mx

Ana María Escalante-Gonzalbo

Computing Department, Institute of Cellular Physiology, Universidad Nacional Autónoma de México. Apartado Postal 70-600, 04510 México, D.F. Mexico. Email: aescalan@ifisiol.unam.mx

Shirley Ainsworth*

Library, Biotechnology Institute, Universidad Nacional Autónoma de México. Av. Universidad 2001, Cuernavaca, Morelos, 62250, Mexico. Email: shirley@ibt.unam.mx *corresponding author

A solution to reduce user frustration with multiple databases, confusing interfaces, bewildering query languages, assorted search protocols, and differing results formats. We present **Hermes**, a tool designed to integrate information resources, from searching in disparate bibliographic databases, to accessing the full text articles online via one seamless common interface. Hermes aims to reduce some of the artificial boundaries that exist in the information space, the barriers between user need and gratification. The system has been developed in Perl, on a Linux platform and includes a MySQL database. At the present it searches 12 bibliographic databases on three different platforms and delivers results from more than 6,000 full text electronic journals. We also discuss the obstacles which exist for the development of this kind of tool, due to the lack of acceptance and implementation of standards, and we propose new utilities that could enrich these types of tools, if the standards are implemented.

Frustration

There has been an explosive and rather chaotic growth in information services available on the internet in the last 10 years. Most of the service providers have migrated their database search platforms to the web, and the majority of academic journal publishers have developed electronic versions of their publications (Grogg & Tenopir, 2000). This growth, like that of many other applications which have migrated or been developed on a web platform, has been highly disorganized, and given the lack of rules and previously established standards, each supplier and publisher has decided independently on both the formats and the access protocols for storage and search. As a result we have a tangle of sites, passwords, protocols and formats. The poor user is bewildered and is unable to use all the information services effectively. He/she is lost in the information space. So great is the information scatter, that it can be a daunting task to get from searching in several databases to the full text. These problems have been recognised before, both by academic experts in information retrieval and information service providers.

Referring to searching in electronic databases Schneiderman, Byrd & Croft (1997) say "Current user interfaces for textual database searching leave much to be desired: individually, they are often confusing, and as a group, they are seriously inconsistent" and continues "many of the current text-

search interfaces -- especially on the World Wide Web -- are neither simple nor clear: they are often needlessly complex, and they very often obscure key features. The result is confusion, frustration, and failure for intermediate and advanced users as well as novices." Payette & Reiger (1997) comment on the same point, "Users often encounter frustration in their efforts to discover relevant sources, negotiate connections, learn resource-specific user interfaces, and search using a variety of inconsistent query languages and semantic conventions."

Two basic problems can be identified: information scatter makes it difficult to select and explore the range of available resources, and the user needs training in the use of heterogeneous sytems to search and retrieve the information. As a first step we need to integrate and homogenize; that is, group together the greatest number possible of services in one place, with a single consistent interface. In an article about the creation of a Digital Library Federation, Liu et al (2001) state clearly "The usefulness of the many on-line journals and scientific digital libraries that exist today is limited by the inability to federate these resources through a unified interface." As far as Rudner (2000) is concerned, he insists "Digital resources must be developed with expert intermediaries and contain pre-selected resources if they are to be a service."

There have many different attempts, both by service providers and academic groups, to reduce the dispersion of resources; however, none of them has been completely successful.

1.- In the case of attempts at integration initiated by service providers, the fundamental problem is that the alternatives they offer will always be, by reason of their origin, partial solutions, with their own databases and their own journals (or the ones they have licence agreements with). In no way can they cover the whole range of available information. One cannot fail to ask if it is ingenuity or underestimation of the users' needs, when one reads on the web pages of information aggregators phrases such as .."Here you can find all the information you need, without having to change interface…" or, "... if you use this system, you won't need to go elsewhere to obtain all the information…" This is very limited *one-stop shopping*.

2.- In the case of the more academic solutions, the programmer must confront the underlying heterogeneity of the information, specifically the storage formats, search protocols and results formats. Paepcke et al (2000) state "The problem in creating such applications is that no generally agreed upon programmatic interface exists for accessing information sources. Rather than focusing on innovative user level facilities, programmers must expend effort on accommodating unnecessarily different information source access methods, or even resort to screen-scraping of Web pages in order to retrieve information," and continues, "There is, then, a need for what we call "search middleware". This term refers to protocols and associated software packages that enable information application writers to access information sources easily. Of these, the Z39.50 standard is the most widely known, but it is a somewhat complex and detailed protocol. Not all databases support the Z39.50 standard, and attempting to tease out the information contained within proprietary databases can be a taxing activity.

And so we can observe that the integrated information solutions that have been developed are still incomplete or insufficient, due to the great amount of programming obstacles which must be overcome and the lack of cooperation between some of the service providers, who refuse to incorporate standard protocols into their search systems. Tennant (2001) has compiled a representative list of cross-database searching tools in existence or in development.

The problem of heterogeneity also affects, though to a different degree, the second stage of integration, that of direct access to the full text of the search results. Few systems have attempted to include this functionality, and none has been completely successful, given that apart from the problem of the multiplicity of protocols and storage schemes as implemented by each publisher or even electronic journal, must be added the complexity involved in the control of electronic journal subscriptions. Some are placed directly with the publisher, others through agents such as *Swets* or *Ebsco* who provide their own search interface, still more through aggregators and there are also collections maintained on local servers (also known as the *appropriate copy* problem (Caplan & Arms, 1999)). Some work has begun on the development of standards in this area, which facilitate the identification and obtention of specific articles taking into account standard information obtained from the citation. However, much work remains to be done on the acceptance and implementation of standards, before they can be effectively functional.

The Full Monty- Gratification, Hermes Style

Hermes is a tool designed to integrate information resources, from searching in multiple bibliographic databases, to accessing the full text online via one seamless common interface. In its present stage 12 databases on three different platforms are integrated: *Current Contents (CCC)* from ISI, *Pubmed* from the National Library of Medicine, and 11 scientific databases from *Silver Platter*.

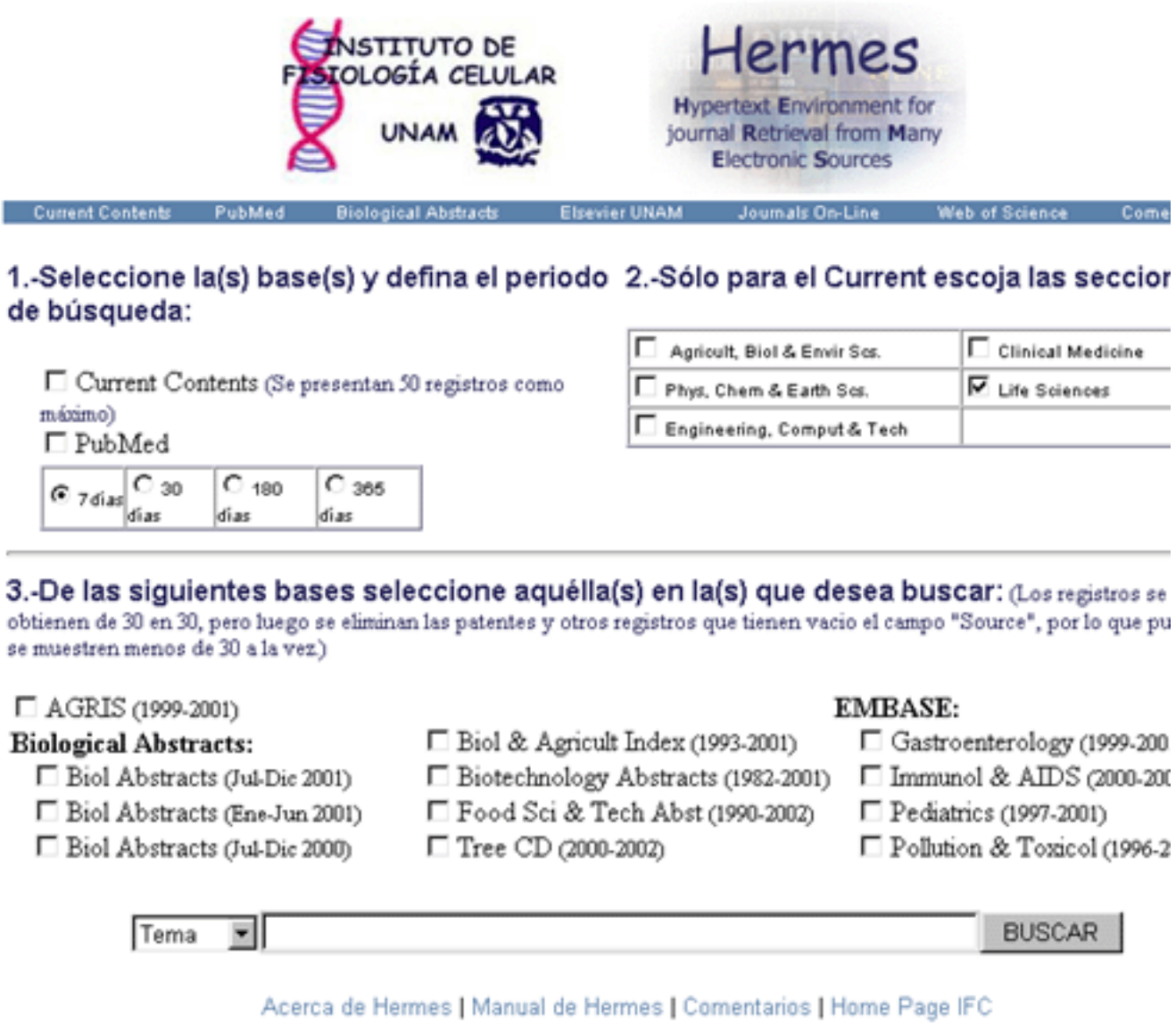

Figure 1. Shows **Hermes** main search screen, with available databases and search input form.

The user can select all or several of the available databases for simultaneous searching, on the fields of subject, title, author or journal title. Searching can be done by keyword, phrase or multiple terms connected by boolean connectors (Fig. 1).The search results are presented in a uniform consistent format, and access to the full text journals is controlled through a database which stores the URLs for the journals subscribed to by the institution. **Hermes** represents a specific solution for the *Universidad Nacional Autónoma de México (UNAM),* but small modifications in the content of the locally held databases would allow it to be adapted to the conditions of any institution.

The main problems faced during its development were caused precisely by the lack of standards implementation by the different service providers, in addition to the efforts of at least some of them to

maintain their information as obscure and inaccessible as possible. The result has been that the system is not as efficient as we would like, and that access to the journals in most cases is not yet at the article level; but it is a completely functional system that operates very satisfactorily.

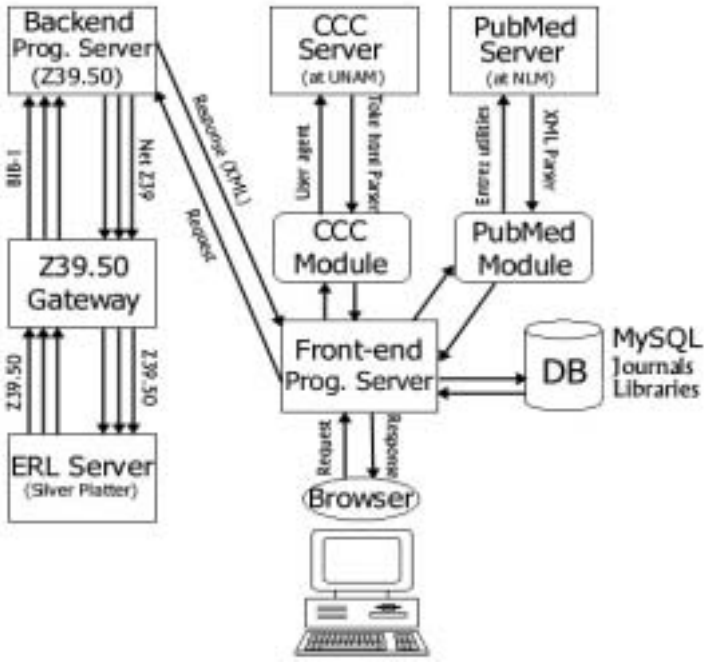

Figure 2 Demonstrates the **Hermes** system architecture

The system has been developed in Perl, on a Linux platform, and the user interface is in html. The system architecture (fig. 2) is based on a main front-end program server, which controls the information flow in the system: it receives the search query, and sends parallel agents to the different search modules and/or backend program server, which take charge of obtaining information from the different databases. It receives the replies from the search modules, filters the references obtained through the database of full text journals available for the institution, and formats the results. The search modules format the request appropriately for each specific search engine (that is, *Current Contents* or *Pubmed*), establish communication with the remote data base, send the request and extract the results. In the case of *Pubmed*, the request is made using the *entrez utilities,* designed by the NLM specifically to facilitate online search requests, and the process is very efficient. In the case of *Current Contents Connect*, we have to simulate a non-interactive Web search session, due to the lack of support for standards and online access utilities by this company. The process is slow and inefficient, and the results are limited to 50 references for considerations of time.

The backend program server controls the requests to databases which support the Z39.50 protocol (at the moment the Silver Platter databases). It receives requests from the front-end program server, converts the search query to RPN (Reverse Polish Notation) and sends parallel requests to the Z39.50 Gateway for each database selected by the user. It receives data from the ERL server (from Silver Platter), through the Gateway, translates the data from BIB-1 to XML, stores the abstracts in a temporary file and sends the results to the main front-end program.

Locally two databases are maintained in MySQL, the database with details of the full text electronic journals accessible from the UNAM, which contains, amongst other things, the URLs to

connect to the journals; and a database of libraries and librarians in the UNAM, which resolves article requests automatically by ip address of the client machine.

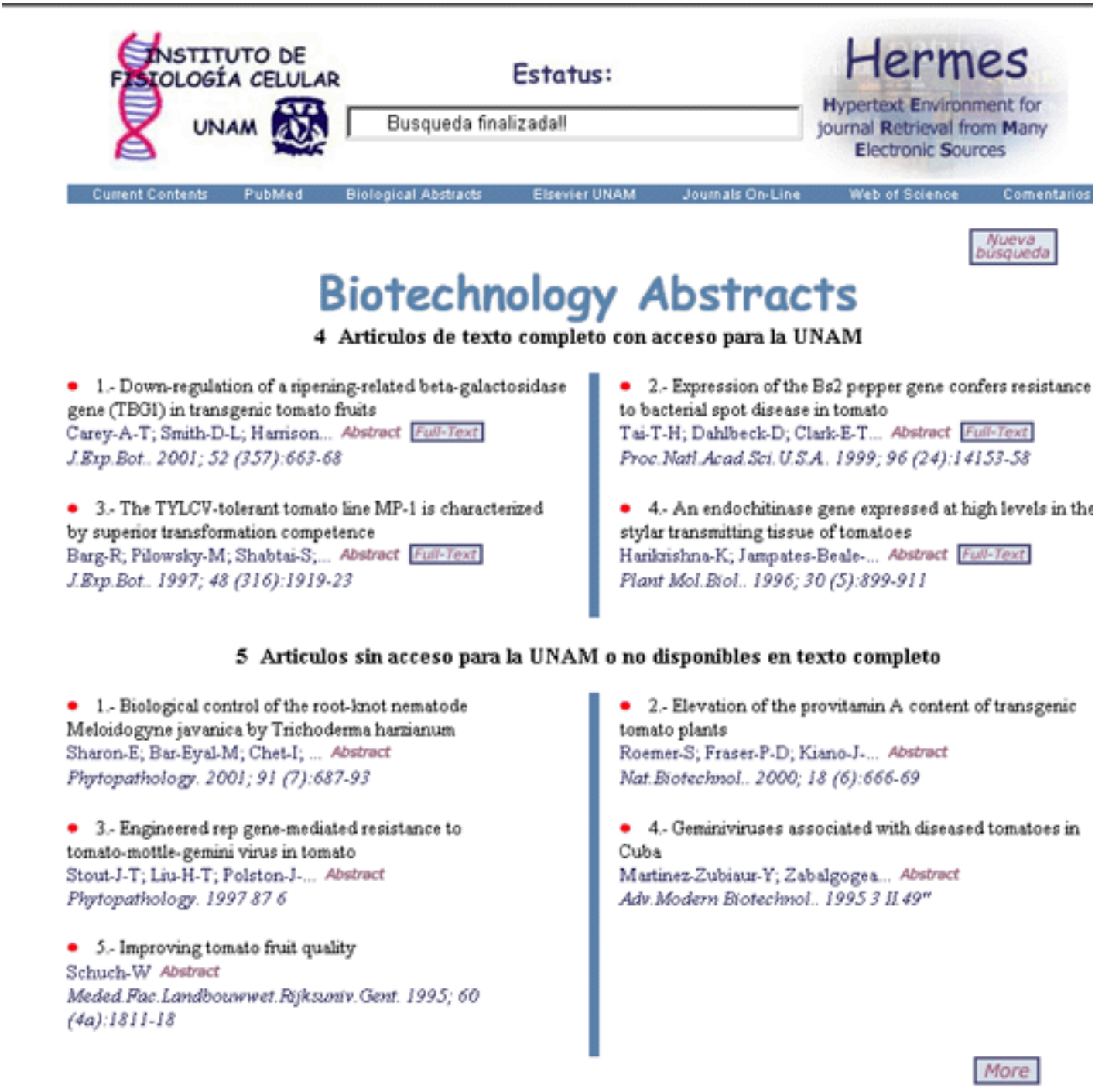

Figure 3. Results screen from a Hermes search.

Hermes displays the results organized by database, and separated into two sections (Fig. 3): first, the references for which the users can access the full text of the articles, each with its corresponding link, and then the results for which there is no online access. For this second group of references, we have implemented a system which can send an ILL request to the appropriate departmental librarian. (The UNAM has a decentralized system of some 168 different libraries spread over 7 campuses.) All the results include a button allowing the viewing of the article full record and/or abstract when this is available.

Looking to the future

Hermes is a fairly complete solution to the problem of integrating services, but much work remains to be done. Some of the projects we are working on for its further development and more efficient functioning are:

1.- The incorporation of search engines from other database systems (*Ovid Technologies, Cambridge Scientific Abstracts, Proquest* etc), in order to integrate more resources and cover more areas. This implies redesigning Hermes as a portal organized in different subject areas.

2.- A *public* version based on free databases and free electronic journals, such as the free backfiles of many journals offered by Highwire Press and others, and the Biomed Central journals, so we can promote the use of these valuable resources.

3.- Hermes is a project that has been developed entirely with free software, and we hope to be able to include it as Open Source. To that end we will need to restructure it as a modular system, with standard output formats, which can be invoked independently by other systems similar to Hermes.

4.- We are beginning to incorporate additional services into the results, based on metadata, such as that proposed by Van de Sompel (1999) "to create added value by linking related information entities, as such presenting the information within a broader context estimated to be relevant to the users of the information". Development of the **LEO** system (*Linking External Objects*) which similar to *OpenURL* (Van de Sompel, 2001) and its commercial manifestation *SFX*, will provide the user with: complete reference, abstract, multiple ways to access the full text (there are many *appropriate copies*), links to related articles and genome and sequence databases when appropriate, amongst other things. LEO will have a philosophy similar to OpenURL (Van de Sompel, 2001): “the OpenURL framework as an architecture that allows a user to escape from the metadata plane in which default links relating to a referenced scholarly work are delivered by information services. The architecture gives the user the freedom to reach into an overlaying service plane and ask a service component to deliver additional/alternative/appropriate service-links that relate to the referenced scholarly work”.

5.-Finally and working closely with the LEO system, we intend to concentrate on improving access to the full text at article level. Initially we shall have to depend on ad-hoc solutions, using the *links-to-services* that some publishers are beginning to provide. The great disadvantage of course lies in the fact that each publisher or even journal constructs their own system for linking in. Our OpenURL type of system will generate the full text links on the fly, taking as a basis the article citation supplemented by information stored in a database with each specific publisher routine.

In the long run, solutions will be implemented based on the OpenURL and DOI (Digital Object Identifier) standards, (Rosenblatt, 1997) as soon they are accepted and they have widespread usage among publishers and database providers. We cannot delay the development of **Hermes** until these standards are consolidated, especially given that experts affirm that OpenURL and DOI are not yet accepted standards and that much agreement still needs to be made among the active players (Hitchcock, 1998). In the words of Van de Sompel (1999) "Straightforward progress … is highly dependent on the cooperation of the information industry. Many established players might be reluctant towards such an idea since it requires far-reaching openness of their services. Proprietary solutions are part of a traditional strategy aiming at the minimization of competition."

To conclude, we would call upon the companies which produce bibliographic databases and the journal publishers to cooperate in the integration of standards, which damages no-one, and permits the development of much more creative and ambitious tools, providing better quality information to the user in a shorter time.

REFERENCES

Caplan, D., & Arms, W. (1999). Reference linking for journal articles. *D-Lib Magazine* 5(7/8). Retrieved May 2nd 2002 from http://www.dlib.org/dlib/july99/caplan/07caplan.html

Grogg, J., & Tenopir, C. (2000). Linking to full text in scholarly journals: here a link, there a link, everywhere a link. *Searcher* 8 (10). Retrieved May 2nd 2002 from http://www.infotoday.com/searcher/nov00/grogg&tenopir.htm

Hitchcock, S., Carr, L., Hall, W., Harris, S., Probets, S., Evans, D., & Brailsford, D. (1998) Linking electronic journals. Lessons from the Open Journal Project. *D-Lib Magazine*, December. Retrieved May 2nd 2002 from http://www.dlib.org/dlib/december98/12hitchcock.html

Liu, X., Maly, K., Zubair, M., & Nelson, M. (2001) Arc - an OAI service provider for digital library federation. *D-Lib Magazine* 7(4), April 2001. Retrieved May 2nd 2002 from http://www.dlib.org/dlib/april01/liu/04liu.html

Paepcke, A., Brandriff, R., Janee, G., Larson, R., Ludaescher, B., Melnik, S., & Raghavan, S. (2000). Search middleware and the simple digital library interoperability protocol. *D-Lib Magazine* 6(3). Retrieved May 2nd 2002 from http://www.dlib.org/dlib/march00/paepcke/03paepcke.html

Payette, S., & Rieger, O. (1997) Z39.50: the user´s perspective. *D-Lib Magazine*, April. Retrieved May 2nd 2002 from http://www.dlib.org/dlib/april97/cornell/04payette.html

Rosenblatt, B. (1997). The Digital Object Identifier. Solving the dilemma of copyright protection online. *The Journal of Electronic Publishing* 3(2). Retrieved May 2nd 2002 from

http://www.press.umich.edu/jep/03-02/doi.html

Rudner, L. (2000). Who is going to mine digital library resources? And how? *D-Lib Magazine* 6(5). Retrieved May 2nd 2002 http://www.dlib.org/dlib/may00/rudner/05rudner.html

Shneiderman, B., Byrd, D., & Croft, B (1997). Clarifying search. A user-interface framework for text searches. *D-Lib Magazine*, January. Retrieved May 2nd 2002 from http://www.dlib.org/dlib/january97/retrieval/01shneiderman.html

Tennant, R, (2001). Cross-Database Search: One-stop Shopping. *Library Journal*, October 15th Retrieved May 2nd 2002 from http://libraryjournal.reviewsnews.com/index.asp?layout=article&articleid=CA170458&display=Digital+LibrariesNews&industry=Digital+Libraries&industryid=%industryid%&verticalid=151

Van de Sompel, H., & Hochstenbach, P. (1999). Reference linking in a hybrid library environment. Part1. Frameworks for linking. *D-Lib Magazine* 5(4). Retrieved May 2nd 2002 from http://www.dlib.org/dlib/april99/van_de_sompel/04van_de_sompel-pt1.html

Van de Sompel, H., & Beit-Arie, O. (2001). Generalizing the openURL framework beyond references to scholarly works. The Bison-Futé model. *D-Lib Magazine* 7(7/8). Retrieved May 2nd 2002 from http://www.dlib.org/dlib/july01/vandesompel/07vandesompel.html