MEDINFO 2004
M. Fieschi et al. (Eds)
Amsterdam: IOS Press

The NLM Gateway: a Metasearch Engine for Disparate Resources

Lawrence C. Kingsland III, Maureen F. Prettyman, Sonya E. Shooshan

U.S. National Library of Medicine, Bethesda, Maryland, USA

Abstract

The U.S. National Library of Medicine (NLM) has created a metasearch engine called the NLM Gateway at the URL "gateway.nlm.nih.gov". The Gateway allows the user to issue one search that takes place on multiple NLM retrieval engines. A composite result set is presented in several categories of information: journal citations; books, serials and audiovisuals; consumer health; meeting abstracts; and other collections.

Keywords

Information Storage and Retrieval; Medical Informatics Applications; Medical Informatics Computing; Databases, Bibliographic; Systems Integration; Automatic Data Processing; Systems Theory; Computing Methodologies

Introduction

The U.S. National Library of Medicine offers an increasing number of Internet-based information resources, each with its own Web address and its own user interface. We believe an intelligent gateway system may help some of NLM's users by letting them initiate searches in multiple retrieval systems from one interface at one address. The NLM Gateway is such a system. An important point to note is that the individual systems still have their own interfaces, for those who do know where to look and who wish to do focused searches in a single place.

The target audience for the new system is the Internet user who comes to NLM not knowing exactly what is here or how best to search for it. The user audience includes a wide variety of persons with differing training, backgrounds and experience. New NLM systems draw in users with diverse information needs. The MedlinePlus site for consumer health information, built for the public end user, is likely to create a strong influx of users who wonder "So what else is there?" More recent sites such as *ClinicalTrials.gov* and Toxnet on the Web are likely further to increase the number of new users accessing NLM information for the first time. Finally, a category of users sometimes given short shrift is machines: programs running on computers elsewhere on the Web that must access our data through Application Program Interfaces (APIs).

The goals of the Gateway system are to

- Provide "first-stop shopping" for an increasing number of NLM information resources,
- Help lead users to information they might not have known was present,
- Offer citations, full text, video, audio, and images,
- Ultimately, offer user profiles to guide searches in clusters of databases.

The Gateway project directly addresses the problem of users not knowing where to look and not wishing to learn a dozen new interfaces. It provides a single address with a single look and feel, allowing users to state a query and have it reformulated and sent automatically to multiple retrieval systems having different characteristics but potentially useful results. Results from the target systems are presented in categories (e.g., journal citations; books, serials and audiovisuals; consumer health; meeting abstracts; other collections) rather than by database. Access to multiple collections within a single category of results is offered when the organization of the information makes this useful.

We expect that users will come to the Gateway for an overview scan of NLM's resources. Those who immediately find what they need will be pleased. If they note that one resource such as PubMed or MedlinePlus has information they'd like to know more about, they can go straight to that resource for a focused search using its native interface. Direct links to other major NLM resources are provided on the Gateway's search screen. This combination of a single point of access for an overview scan coupled with focused searches for a second phase of inquiry should help improve user access to information offered at NLM's increasingly comprehensive series of Web sites.

Methods

Resources Accessed by the Gateway

The NLM Gateway accepts the user's query and translates it as appropriate for different retrieval systems. Specific examples will be shown later. The current version of the NLM Gateway issues simultaneous searches in

- MEDLINE (14+ million journal article citations 1953-present, in PubMed using the PubMed API; includes access to LinkOut and PubMed Central)
- LOCATORplus (approx. 1,100,000 books, monographs, serial titles, audiovisuals, other cataloged information; searched in the Voyager ILS using Z39.50 for the initial search and JDBC to Oracle for specific elements of individual records)
- MedlinePlus Health Topics (650+ health topics, searched using the MedlinePlus API with hot links to MedlinePlus itself after search)

- MedlinePlus Drug Information (for approx. 9,000 drugs, searched as with Health Topics)
- MedlinePlus Medical Encyclopedia (approx. 4,000 articles on diseases, tests, symptoms, injuries and surgeries, searched as with Health Topics)
- MedlinePlus Current Health News (late-breaking stories about medicine and health, searched as with Health Topics)
- MedlinePlus Health Tutorials (easily understood animated graphics and sound to explain conditions and procedures, searched as with Health Topics)
- *ClinicalTrials.gov* (approx. 9,100 clinical trials, searched using the *ClinicalTrials.gov* API)
- DIRLINE (approx. 9,000 records on health organizations, searched using the DIRLINE API)
- Meeting Abstracts (approx. 72,000 HIV/AIDS abstracts, 8,600 Health Services Research abstracts, 900 Space Life Sciences abstracts, all searched in the Verity retrieval system component of the Gateway)
- HSRProj databank (approx. 6,100 Health Services Research Project records, searched in Verity)
- OMIM (Online Mendelian Inheritance in Man, catalog of approx. 15,000 human genes and genetic disorders, searched using the OMIM API)
- HSDB (Hazardous Substances Data Bank covering approx. 4,700 hazardous substances, searched using the Toxnet API)
- UMLS Metathesaurus (3,000,000 names for 1,060,000 concepts from 100 vocabularies in 15 languages, searched using custom Gateway code)
- Document delivery (through NLM's Loansome Doc facility).

Gateway Architecture and Transaction Flow

The Gateway is fully object oriented, written primarily in Java and using the CORBA (Common Object Request Broker Architecture) distributed object computing infrastructure. The current design deliberately does not use "cookies" and can run without JavaScript, since both are sometimes banned in institutional settings for security reasons. Users can set preferences to adapt some aspects of the interface to their needs. The user can specify which record elements are to be presented in the brief and expanded displays. A "locker" is used to store selected results the user wishes to keep across sessions. Use of the locker requires a login. Users can order documents and can email results to themselves or to colleagues.

Where appropriate, internal data interchange is done in XML. External data interchange is XML-preferred, but output is offered in XML, HTML and ASCII text. There are Application Program Interfaces (APIs) to major Gateway functions to make them accessible to other programs. In particular, the Metathesaurus browser component of the new system will provide standalone functions accessible to other programs. The commercial Resonate system provides virtual IP services that allow load sharing and failover among multiple machines. MySQL is used for the database of user IDs and passwords that allows users to log in and be matched with their preferences and their lockers.

A sample transaction details how the Gateway system works. The user enters query terms and presses the Search button on the Search Screen. The user's browser sends a request to the Command Broker. The Command Broker parses the input data, translating the request into a CORBA method call on the User Interface Manager.

The User Interface Manager receives data from the Command Broker and parses the user input search string into its component parts. It loads these into the Document Object Model (DOM) server as a search document. The User Interface Manager then tells the DOM to create an empty result document. The User Interface Manager next processes user preferences and the input from the browser to determine which databases to query, how many items to retrieve, and what fields the user wishes to display in brief or in expanded records. Using this information, it finishes filling in the search document stored in the DOM. The User Interface Manager then sends the search document pointer and the result document pointer to the Data Source Broker (DSB), which reads the search document from the DOM.

Working with this information, the DSB translates each term of the search in accordance with Gateway rules for each of the databases. The resulting translation is copied back into the DOM. Searches for each of the databases requested in the search document are spawned as separate threads in Sun's Solaris operating environment and sent in parallel. As each database returns its results, the DSB transforms the data from the database's native format into the Gateway's internal format, designed to hold result sets from any arbitrary database. As each record is translated, it is stored in the DOM's result document. Any errors or warnings are also stored in the DOM's result document.

When all data have been received, the DSB returns control to the User Interface Manager. The User Interface Manager checks to see whether all the requested data was received. If not, it sets one or more error flags. It then passes the search document and the result document pointers and control to the Formatter. The Formatter has been informed by the User Interface Manager which page the user is on. It reads that page template from permanent storage. Following our template design macros, it knows to extract and display various pieces of data that have been passed to it by the User Interface Manager. Eventually it reaches the part of the template that tells it how to format a citation, at which point it requests the citation record data from the DOM.

Following the processing instructions in the template, in several configuration files and in the user preference data, the Formatter generates an HTML page containing everything received from the DOM. The completed HTML page is returned to the User Interface Manager, which returns it to the Command Broker, which sends it to the user's browser using the Apache web server.

Several other interactions happen behind the scenes. When the browser first communicates with the User Interface Manager, if no session ID is present or the session ID is stale, the User Interface Manager requests a new session ID from the Security Manager.

Database	PubMed	LOCATORplus	Meeting Abstracts (Verity)
Type of Search	*Program Action* [LC = LowerCase]	*Program Action* [SH = subheading(s)]	*Program Action*
AUTHOR (AU)	LC(Echo), append [AU]	NA	Parse; (LC(last name) <EXACT> <IN> LNAME AND LC(initials) <IN> INITIALS AND LC(suffix) <IN> SUFFIX) <IN> AUTHOR
MeSH (MH)	If no SH: LC(Echo), append [MH]. With SH: Entry for each SH separated by OR operator; LC(Echo) / SubheadingLookup(Echo)append [MH]	If no SH: MeSHLookup (Echo) /u=25; t=1 (MeSH Subject). With SH: Entry for each SH separated by OR operator; MeSHLookup(Echo) /u=25; t=1 (MeSH Subject) AND SubheadingLookup(Echo) /u=25; t=1 (MeSH Subject)	If no SH: MeSHLookup(Echo), remove parentheses, append <EXACT> <IN> KEYWORD_LIST. With SH: Entry for each SH separated by OR operator; MeSHLookup(Echo), remove parentheses, append <EXACT> <IN> KEYWORD_LIST
SUBJECT (SU)	Parse into phrases or user quoted phrase; Entry for each phrase separated by AND operator; LC(phrase)	Parse into phrases or user quoted phrase; Entry for each phrase separated by AND operator; LC(phrase) /u=1016 (KeyWordAnywhere) OR MeSHLookup(phrase) /u=25; t=1 (MeSH Subject) OR SubstanceNameLookup(phrase) /u=1016 (KeyWordAnywhere)	Parse into phrases or user quoted phrase; Entry for each phrase separated by AND operator; LC(phrase) <IN> TITLE OR LC(phrase) <IN> ABSTRACT OR LC(phrase) <IN> KEYWORD_LIST OR MeshLookup(phrase) <EXACT><IN> KEYWORD_LIST OR SubstanceNameLookup(phrase) <EXACT> <IN> KEYWORD_LIST OR LC(phrase) <IN> SPACE_FLIGHT_MISSION

Figure 1 - Small Fraction of the NLM Gateway Search Translation Maps

If the user invokes the Locker or Preferences button from the display sidebar, he is given an opportunity to log in. Once the user logs in, a call is made to the Preference Manager. The Preference Manager supplies the user's current set of preferences, which override corresponding data in the default preferences configuration for that session. Many elements are configurable in the user preferences files.

At each important event, a log message is sent to the Alarm Panel to record the progress and operation of the system. These logs will be used later to provide statistical reports of system usage.

Core of the Gateway: Search Translations and Results Display

At the heart of the Gateway is its ability to accept a user's search and translate it into the series of search statements needed by each of the retrieval systems the Gateway can access. This process is accomplished by the Data Source Broker. The search translation algorithms are created by a trained medical librarian on the Gateway team in consultation with the NLM experts responsible for the retrieval systems. Some of the search algorithms build on mapping techniques first applied by NLM in the earlier Coach [1] and Internet Grateful Med systems. They are then tested carefully to confirm that the Gateway achieves search results comparable to those of the native interface whenever feasible. A very small fraction of the NLM Gateway search translation maps is shown in Figure 1.

For a gateway system dealing with disparate retrieval systems, it may be as important to display results coherently as it is to translate searches correctly. Each of the retrieval systems presents results in its own format, with its own look. As noted above, the Gateway's Data Source Broker accepts these results, parses them into a canonical Gateway form in XML, and stores them in the Document Object Model, the DOM. The Formatter reads them from the DOM in XML canonical form, reformats them according to configuration-table defaults and the user's preferences, and sends them to the User Interface Manager for display by the user's browser.

Result formatting can also be seen in a broader view. Navigation of the results from searches across multiple systems opens many avenues of exploration. Soyeon Park [2] defines three types of user interaction with the results of multiple database searches.

1. Separate interaction: the user is connected to each system involved and must understand how to use the capabilities of multiple systems and also merge the results from multiple system searches;
2. Common interaction: a single interface for searching different resources so that users do not need to know specific query languages and techniques specific to each source but still need to repeat actions as they move from one source to another as well as integrate the results; and
3. Integrated interface searches: the interface searches multiple sources and the user interacts with the integrated results but is unable to take advantage of characteristics unique to the individual sources.

The results of Park's study indicate that users prefer Type 2 and Type 3 to Type 1, with Type 2 having a slight edge. The preference for Type 2 was largely based on the fact that users could select the databases to search once they were familiar with the options. However, the preference for Type 3 was based on the utility of the integration of the results. Interestingly, we have some of the characteristics of all three types of user interaction in the systems discussed in this paper.

NLM's former Internet Grateful Med was a Type 2 system (common interaction) with some aspects of a Type 1 system. The user chose a database and searched in that database individually, but from a series of search screens that had a deliberately coherent

look and feel. Results were presented from only one system at a time, with a concerted effort made to carry across search terms and other search elements from the previous search as appropriate when a user changed databases.

Another system, NLM's HSTAT (Health Services/ Technology Assessment Text) [3] currently is also a Type 2 (common interaction) system. That is, the user may select multiple sources to search and the HSTAT client reformulates the query as necessary for other systems. The results, however, are not integrated. The HSTAT results are followed by the results from each additional site selected. To view the expanded results from another site, the user is taken to that site (and then has the option of refining the search using site-specific capabilities). HSTAT efforts in agent technology [4] have addressed the desirable goal of integrating search results from multiple sources.

The Gateway is more nearly a Type 3 (integrated interface) system. The interface searches multiple sources and the user receives an integrated results summary count showing numbers for each document collection. This saves time and saves load on the target retrieval systems; the user can call up results from any document collection with a single mouse click, but may not always choose to do so.

The search translations and the results formatting happen behind the scenes, but they are the keys to the Gateway user's experience in seamless simultaneous searching of multiple retrieval systems.

Results

The NLM Gateway has been in production since October 2000. It should be noted that the component nature of the Gateway makes it readily adaptable for the addition of new features. Alternate ways of interpreting and translating searches input by the user are possible, as well as additional help modules. One that has been discussed has been a search page that would perform as a "digital librarian" or "search wizard", walking the user through a set of forms comprising a reference interview and building the search from that input. Other similar new modules, readily pluggable into the component system, offer significant opportunities for using the NLM Gateway system as an infrastructure for retrieval systems research.

With the initial public release of the Gateway accomplished, we have some important enhancements in the planning stages for follow-on versions. Links to NLM's Profiles in Science, Genetics Home Reference, and Images from the History of Medicine are planned. Additional enhancements may stem from the possibilities discussed in the next section.

Discussion

The creation of the NLM Gateway involves research in areas such as user interfaces for naïve users; search formulation given multiple resources; transaction log analysis including data mining; system adaptation to user actions; and deduplicating, ranking and presenting the results of multiple simultaneous searches in several retrieval systems. Many hooks into system functions will create an environment that facilitates experimenting with aspects of the user interface. Various options on both the input side (helping users choose among clusters of databases) and the output side (presenting clear options for displaying results from searches in multiple databases) are worth exploring.

Where appropriate, the new system will be used to explore agent technology as a means of creating multiple searches, executing them, and integrating their results. There are interesting research issues in parsing user queries to provide command statements appropriate to different retrieval systems, user profiling, and source selection to highlight collections most likely to have relevant responses. Domenig and Dittrich have done interesting work in some of these areas [5], as have French and Viles [6]. Soergel has concisely stated useful goals for powerful search functions that combine information across databases in [7]. We will track and test developments in methods for maintaining state when necessary in Web transactions. We will continue to explore means of load testing and load balancing. We will consider load quantifying and throttling when necessary to detect and minimize the effects of denial-of-service attacks.

Finally, with NLM's increasing emphasis on providing information to patients, their families and the public, it becomes even more important to continue exploring creative methods of transforming the queries of consumer-level users to those medical terms more likely to get good retrieval from professionally indexed databases.

The system's design makes a concerted effort to build on the strength of the intelligent gateway concept: the ability to offer value-added capabilities the user did not know to ask for. Links within and across databases will be vigorously exploited. Searches in clusters of databases or simultaneously in all sources accessible to the gateway will be offered. In user preferences, the user can specify fields to display in particular result sets.

Evaluation

Information science researchers have been evaluating retrieval systems searching individual databases for decades. The evaluation of systems that simultaneously search multiple databases on multiple retrieval engines is somewhat different, and necessarily more complex. Harter notes that documents come and go, schemas evolve, and retrieval algorithms change [8]. One could perhaps do precision and recall tests on each of the databases searched, then produce a composite measure. The dynamic, regularly updated nature of the databases searched by a production system such as the Gateway adds to the challenge.

Deciding what to evaluate is not always straightforward. Robertson suggests that one approach is to consider three properties of a retrieval system: its effectiveness, how well it satisfies the designer's intent; its benefits or usefulness; and its efficiency or performance [9]. There are other elements still more difficult to quantify. One is the convenience of having single-point access to multiple knowledge resources. This might be balanced against the possible lack of specificity caused by having one common-denominator search screen that may not take advantage of some of the specialized characteristics of some of the databases.

One measure of the utility of any system like the NLM Gateway is usage. If we find that people seek it out and use it for an overview scan rather than going serially to the native interfaces of each of the systems it accesses, the Gateway is doing its job. User satisfaction can be ascertained through user comments and responses to brief surveys. There is another utility that intuitively seems important but is more difficult to measure. This is the fact that the Gateway will be able to get the information in NLM's less-known databases before the public: there is real worth in presenting useful information the user never knew existed, from a database he didn't know to try.

Conclusion

We have presented the background, the current status and the development plans for an NLM Gateway that offers single-point access to multiple NLM knowledge resources. It allows users an overview scan of several NLM systems and will serve as a platform for exploring new functionality in federated database searching.

Follow-on releases of the Gateway will provide access to increasing assistance with query formulation and to additional NLM knowledge resources.

Acknowledgments

We acknowledge with gratitude the intellectual contribution and dedicated efforts of a talented group of software developers in the design and development of the NLM Gateway. Jeffrey Holmes* and Donald White** played key roles as technical leads. Ajay Kanduru*, Manish Inala*, Xiaonan Ju**, Sharada Jayanna**, Xiaocheng Luan** and Lee Mericle** made significant contributions to the evolving system. Our colleagues in the MEDLARS Management Services section of NLM's Division of Library Operations were extremely helpful in suggesting and refining functionality of the user interface and in performing usability testing. (* Taj Technologies, ** Aquilent)

References

[1] Kingsland LC III, Harbourt AM, Syed EJ and Schuyler PL. Coach: applying UMLS knowledge sources in an expert searcher environment. Bull Med Libr Assoc 1993 Apr;81(2):178-83.

[2] Park S. User preferences when searching individual and integrated full-text databases. In: Proc Fourth ACM Conference on Digital Libraries, 1999; pp.195-203.

[3] Prettyman MF, Antonucci R, Lynch P and Mericle L. Electronic publication of health information in an object-oriented environment. In: Proc 1999 ASIS Annual Meeting. Washington, D.C. Nov. 1-4, 1999.

[4] Luan X, Prettyman MF, Antonucci, R. System expansion and integration with agents in HSTAT. In: Proc World Scientific Conference. Hong Kong, December 14-17, 1999.

[5] Domenig R and Dittrich K. A query based approach for integrating heterogeneous data sources. In: Proc CIKM 2000, Conference on Information Knowledge and Management, McLean, VA, USA, pp. 453-460.

[6] French JC and Viles CL. Personalized information environments: an architecture for customizable access to distributed digital libraries. D-Lib Magazine 5:6, June 1999.

[7] Soergel D. A framework for digital library research. D-Lib Magazine 8:12, December 2002.

[8] Harter SP and Hert CA. Evaluation of information retrieval systems: Approaches, issues, and methods. In Martha E. Williams (Ed.) Annual Review of Information Science and Technology, Vol. 32. (pp.3-94). Medford, NJ: Information Today, 1997.

[9] Robertson SE. The methodology of information retrieval experiment. In: Karen Spark Jones, ed. Information Retrieval Experiment. London: Butterworths, 1981, pp 9-94.

Address for Correspondence

Lawrence C. Kingsland III, Ph.D.
Assistant Director for Applied Informatics
National Library of Medicine
8600 Rockville Pike
Bethesda MD 20894
USA
kingsland@nlm.nih.gov