
Genesis of an electronic database expert system

*Wei Ma and
Timothy W. Cole*

The authors

Wei Ma (weima@uiuc.edu) is Assistant Reference Librarian and Assistant Professor of Library Administration, and Timothy W. Cole (t-cole3@uiuc.edu) is Interim Mathematics Librarian and Associate Professor of Library Administration, both at the University of Illinois at Urbana-Champaign, Urbana, Illinois, USA.

Keywords

Academic libraries, Databases, End-users, Expert systems, Web applications, World Wide Web

Abstract

This article reports on the creation of a prototype, Web-based, expert system utility that helps end-users better navigate the range of library databases available at the University of Illinois at Urbana-Champaign (UIUC). Both librarian-assigned database descriptors and terms drawn from the controlled vocabularies of the databases themselves are used to thoroughly characterize resources. End-users then utilize keyword searches and/or menu selections to identify resources most relevant to their information needs. In addition to reporting on the UIUC prototype and the work done to create it, the concerns that gave rise to the project are discussed. Previous work and research elsewhere are summarized, and the more common approaches currently in place in academic libraries today are noted. Plans for testing the UIUC prototype with librarians and end-users, for evaluating the results of those tests, and for iteratively refining the tool based on those evaluations are then briefly described.

Electronic access

The current issue and full text archive of this journal is available at
<http://www.emerald-library.com>

Challenges and issues

The past decade has seen an explosion in the numbers and types of electronic databases available to identify and retrieve articles and other information relevant to a scholar's research needs. Students and faculty are faced with a confusing overabundance of choices. This situation challenges libraries and demands that they develop new mechanisms to facilitate and better inform end-user selection of electronic databases and search tools. The challenge is that much greater because the ready availability of online information resources enables scholars to conduct much of their library research remotely over the campus network – thus reducing the frequency and depth of their in-person contact with librarians. The process has been further complicated with the rapid development of information technology and the proliferation of digital information resources. Libraries and their users are faced with the following challenges.

The rapidly increasing number of electronic databases

In 1991 online journal article index databases were made available to end-users campus-wide at the University of Illinois at Urbana-Champaign (UIUC) for the first time. The databases made available that year were Current Contents (Institute for Scientific Information), ERIC, and seven general and subject-specific databases from H.W. Wilson. A small number of CD-ROM databases were also available in selected departmental libraries at that time. By the fall of 1999, the list had grown to 97 general and subject-specific online databases available campus-wide. Over 220

The authors wish to acknowledge the Campus Research Board of the University of Illinois at Urbana-Champaign (UIUC) and the Research and Publication Committee of the UIUC Library, which provided support for the completion of this research. Special acknowledgement also goes to this project's consultant, Professor Linda C. Smith of the Graduate School of Library and Information Science of UIUC. She also helped review resource characterization data collected. Also Professors F.W. Lancaster and Linda C. Smith of the Graduate School of Library and Information Science of UIUC[1].

CD-ROM databases are now available in specific UIUC departmental libraries. The number of electronic databases available to end-users at UIUC has grown by more than a factor of 20 in less than eight years, with more than half of that growth having taken place in the last three years.

Constant changes

Owing to software migration, updates, product changes, vendor mergers, changes in database producer-vendor agreements, and changes in database availability or cost, there has been constant change in, for example, search interfaces, database coverage, database features, and selection of databases licensed by the campus. The look and feel of our library databases change regularly. Which databases are available at any time and the way they are made available change all too frequently. It is difficult enough for library staff to keep up with such changes; it is unrealistic to expect end-users to keep up with such changes over time.

Different access points

The UIUC Library includes more than 40 different public service units spread out across more than 20 buildings. Often CD-ROM information resources are available in only a single library, sometimes only from one workstation within that library. It is nearly impossible for users to be aware of the existence of all the electronic databases available in the campus library system.

Heavy demand and use of the databases

In terms of number of searches performed, the combined use of UIUC campus-wide online journal article index databases, non-existent prior to 1991, has consistently exceeded total use of the online public access catalog by UIUC Library patrons for the past several years.

User response to the problem

In the spring of 1998, the UIUC Library did a user needs assessment survey. As might be expected, given the factors detailed above, difficulty in deciding/choosing appropriate databases to search in order to find desired information was near the top of the list of the

concerns expressed by respondents, both students and faculty. Though recognizing the problem, users deal with these difficulties in less than optimum ways. Library user observations reveal that patrons who used certain databases before have a tendency to select just those familiar databases, ignoring others that could be more appropriate for their current topics (Meyer and Ruiz, 1990). New users may feel confused if professional assistance is not available. Informal observations by library staff reveal that many users waste hours searching irrelevant databases. Even some professional librarians with previous experience in selecting and searching resource databases now may have difficulty maintaining their expertise (Zahir and Chang, 1992; Thornburg, 1987). Because of this difficulty, much of the professionals' advice leads to heavy use of general-interest databases, overlooking others (e.g. more specialized or in different locations) that might be more appropriate and relevant to the topics of the patrons (Hightower *et al.*, 1998). As a result, many valuable specialized databases appear to be underutilized (Hightower *et al.*, 1998).

These generalizations were borne out by an analysis of transaction logs recorded at UIUC during April 1999. During that month, detailed transaction logs were taken regarding user database selections from all public-access workstations (about 70 in total) located at UIUC's Undergraduate Library, Central Circulation and Reference, and Grainger Engineering Library Information Center – the three most frequently visited UIUC Library public service units. Selections made by users from a list of just over 90 Web-accessible bibliographic databases were logged. By default, the list from which these selections were made was organized into six subject categories. Optionally, users could view the list organized alphabetically by database title or by search system/vendor (e.g. Ovid Technologies, FirstSearch, SilverPlatter). They also had available an option to view a subject-organized subset menu listing only those databases that contained at least some pointers to article full-text online. A total of 16,635 database selections were logged during April 1999.

As suggested by the research cited above, general-interest databases were the resources most frequently selected. The top four

databases were Readers' Guide Abstracts (from H.W. Wilson, searched via a local implementation of Ovid), Periodical Abstracts, Wilson Select Full Text (both via OCLC FirstSearch), and Expanded Academic Index (from IAC via their Web site). Together, these four databases accounted for nearly one-third of all resources searched. Search system vendor recognition also seemed to play a role in determining which databases were selected. Databases available via OCLC FirstSearch or Ovid (35 databases in total) accounted for more than two-thirds of all resources selected. Excluding Expanded Academic Index, the 14 databases which each provided their own unique search interface (i.e. not searchable through third party aggregator or vendor such as OCLC, Ovid, or SilverPlatter) accounted for less than 10 per cent of selections made. This was the case in spite of the fact that among these 14 databases were multidisciplinary resources such as the Web of Science (Institute for Scientific Information) and JSTOR.

Previous work

While there has been considerable effort in recent years to enhance end-user interfaces to individual electronic information resources, the work to automate and simplify the navigation process among information resources has been more limited. Vendors with multiple databases provide some tools, but these tools consider only the specific collection of the particular vendor. Web search engines have been developed to simultaneously search all unrestricted resources available on the Web, but because of their breadth they are too generic and are unable to effectively filter the information requested. Additionally, resources only available by license are not included in most Web search engines. Much of the research that has been done (see below) has required the user to select from a very limited range of terminology and descriptors as chosen by librarians and information professionals.

Early trials (pre-Web)

Research on electronic database selection has been ongoing for more than a decade. Several systems were developed to help people decide

which online commercial databases were most likely to satisfy a particular information need. However, for various reasons, many of them did not move beyond laboratory or prototype settings. Among these systems:

- Thornburg (1987) described an interactive menu-driven system, called the "rule based expert system", which was used to guide a user in the selection of electronic databases. A user had to select from seven different menu layers, such as the "general subject" areas (e.g. agriculture, business), the "type request" (e.g. subj-oriented, thesis/diss, citation), "the slant of the request" (e.g. clinical, applied), level of exhaustivity of the search (e.g. comprehensive, selective), "special items" (e.g. agronomy, enon_enton), "special topics" (e.g. 1st_naming, obesity), "specific pub.types" (e.g. journal, non-journal, govt_docts), before the system would suggest a database and would calculate the probability if this database was the best choice among the 18 available. The system was not otherwise searchable at any level.
- Trautman and von Flittner (1989) described a prototype database selection system, which used an open-ended classification system for commercial online database searching with nine different attributes. Each of the nine attributes assigned to a database had a controlled vocabulary schedule of categories. The first eight attributes involved miscellaneous coverage features and the ninth was a new subject classification using a three-level viewpoint schedule. Every category of each attribute was assigned a rank and subject coverage, including time spans, languages, and target audience. The prototype system was written using GURU (an artificial intelligence development shell by Micro Data Base System, Inc.). Six internal experts were used to filter information needs:
 - (1) user model;
 - (2) question clarifier;
 - (3) searcher;
 - (4) ranker;
 - (5) evaluator; and
 - (6) browser.

This system was more sophisticated than

Post-Web developments

The advent of the Web has influenced developments and has led to the creation of several other relevant tools:

- The most prominent work in this area has been the development of powerful and robust Internet search engines. There are many different kinds of Internet search engines in use. Each uses its own algorithms and usually produces different search results for the same queries (Feldman, 1998). According to Susan Feldman, developing trends in Web search engine design focus on areas such as query formulation, customization and flexibility in interface design, and displaying search results (Feldman, 1998). Few of the search engines employ in a meaningful way the information-filtering experience and strategies used by librarians. Users are frequently overwhelmed by too many results, most of which are only marginally relevant to the topic (Balas, 1999). In addition, Internet search engines deal with open registered Web sites only, excluding most commercially developed licensed databases.
- DialogSelect (<http://www.dialogselect.com/main.html>) is the Web version of the Dialog system. The vendor developed this product to access its own family of databases only. The system requires a user to select from a broad subject area to a more specific subject area, and then further identify a specific subject term or a search criterion before the user could get to the final search page to search individual database. This arrangement creates difficulties for the user with a cross-disciplinary topic, or when the user does not immediately recognize the subject category most relevant to their topic.
- The University of California, San Diego, has recently developed an article database selection program, "Database Advisor" (<http://scilib.ucsd.edu/Proj/dba/dba-public.html>), to help users decide which of more than 25 science article databases to search. The system adapted the structure of DialIndex. Search results, again, only represent the keywords and selected subject categories, and display with the most number of "hits" (articles) at the top of the list (Hightower *et al.*, 1998).
- University of Washington's Information Gateway (<http://www.lib.washington.edu/search/>) is another unique approach. The Information Gateway serves as a pointer to a huge variety of electronic resources available at the University of Washington, including online catalogs, electronic dictionaries and encyclopedias, locally developed databases and instruction pages, useful Web sites, electronic journals, and indexing/abstracting databases. It provides two search options: Simple Search and Advanced Search. Simple Search allows a user to choose one of the publication types from the field of Resource Type (e.g. dictionaries, directories, Web sites, encyclopedias, and then combine with one of the Subject Areas (e.g. Accounting & Taxation, Aeronautics, Arts). The only way for a user to select article databases to search is to select "Indexes: Finding Articles" from the field of "Resource Type", and choose one of the general subject terms (limited and again determined by developers) from the field of "Subject Areas". A list of the article databases on the chosen subject area will appear. Advanced Search allows keyword searches in the areas of the author, title, and subject of the individual database (however, the database author search function is not helpful in this context, since article databases do not have authors). Information Gateway moves one step further by allowing keyword searches in the areas of database titles and a limited range of general subject terminology as chosen by librarians and information professionals. It does not, however, allow topic keyword search and other database characteristics searches.

Common library practices today

There are several approaches currently in use to facilitate user selection of online databases. For remote users (users who access the library resources from office or home), these approaches rely heavily on the knowledge and sophistication of the user, often to an unrealistic extent.

(1) Library gateway menu system

Many libraries use list type menu systems (e.g. Ohio State University: <http://www.lib.ohio-state.edu/Tools/subjects.html>, the University of Houston Library: <http://info.lib.uh.edu/remote/access.htm>, UIUC: <http://www.library.uiuc.edu/resource/article.asp>) to assist library users in identifying and selecting resource databases appropriate for their topics. Such organized list menu systems serve well as a centralized point of information about library resources and services available, but as a navigation aid to direct library users to the appropriate resource databases, these menu lists reach their limit when an overabundance of choices are available. Also, many library users come to the library with cross-disciplinary topics, or an inability to place their topics within the existing subject categories. For these users, the menu system is not as helpful as might be desired. Faced with the overabundance of choices on the menu, students have a hard time deciding which database to search. Many users have to browse information and entries on different menu pages, guess by reading the database names, or try searching different databases if assistance is not available.

(2) Recommendation list for special user groups

Most academic libraries are organized into separate units by discipline. Such units then can highlight the subject-specific databases for the specific academic department(s) they serve. For example, a UIUC Biology Library's Web page (<http://www.library.uiuc.edu/bix/electron.html>) lists the databases most relevant to biology. This method, similar to the library Gateway Menu system mentioned above but more manageable because of the discipline focus, works well with scholars with focused interests who know the subject area well. For other users who have cross-discipline information needs, or are less familiar with the organization of a particular discipline, it remains difficult to identify the relevant databases if assistance is not available.

(3) Library instruction workshops or Web page tutorials

This method is good for teaching specific database searching skills. When it comes to database selection, however, it is usually beyond the scope of the tutorial to introduce so

many databases and the specific subjects covered by each database. According to informal user observations and interviews, at best users can come away from such a tutorial only remembering the few databases of interest to them at the time of the tutorial. They will not remember the other databases later when they have a new and different information need.

(4) Annotated paper handouts or annotated Web pages

This is the usual method for describing the databases in detail. However, it is not practical to expect users to read and assimilate page after page of descriptive materials in order to select one or two appropriate databases.

(5) One-on-one reference services advice

This is currently the best way to provide resource database selection advice. However, there are three limits on these services. First, Reference Desk services are not always available while the library is open. For example, the UIUC Undergraduate Library is open 17 hours each weekday (during the normal semester), while reference services are only provided eight hours per day. Second, as mentioned earlier, the quality varies depending heavily on the specific professional experience and expertise of the person working at the Reference Desk. Third, users must be on site in the library or have access to a phone when the Reference Desk is staffed. When the Library or the Reference Desk is closed, no assistance is available.

Building a database selection expert system

Based on the limitation of the methods mentioned above, the authors researched, designed and developed a prototype Web-based database selection expert system. A fully functional version of this prototype has been created and is now being tested by UIUC librarians and end-users. The prototype system is designed to assist users in selecting from among the article databases, directories, statistical and government document electronic resources presently available at UIUC (including Web and CD-ROM bibliographic/full-text resource databases, but excluding the online catalogs, electronic books,

encyclopedias, and dictionaries). Users are able to select resources by any of three methods:

- (1) searching for relevant resources using free-text keywords and phrases;
- (2) browsing available databases by subject categories; or
- (3) choosing desired database characteristics from menus (e.g. material types indexed, chronological coverage).

The Smart Database Selector user interface is shown in Figure 4.

For each of the first two approaches, selections can be narrowed by, for example, broad subject areas, user levels, database types, material formats, content types, and languages.

The project's objectives

- to identify and better understand computer-assisted techniques for guiding users to the databases most suitable for their particular information needs;
- to pioneer innovative methods to improve user services given the proliferation of electronic databases and digital information within individual institutions; and
- to improve the ease of access to information by taking advantage of the Web and related protocols and technologies.

The project involves

- a diverse collection of electronic databases from different information vendors and producers, and from different data-carrier media (e.g. Web and CD-ROM);
- resource identification techniques that allow the user to enter keywords which are then compared to the extensive database descriptors derived in part from the vocabulary of the resources themselves;
- information filtering techniques of professional librarians to narrow the selection to resources most likely to be relevant; and
- exploration of the most up-to-date techniques to deal with the new challenges and issues in today's new electronic information retrieval environment.

Describing and automating the database selection process

Preparatory research work for this project began in the spring of 1998. The search patterns of users seeking information were analyzed; the strategies which reference librarians used to respond to users' information requests were studied; and the databases currently available at the UIUC Library were analyzed for content, format, levels, and more. A database selection algorithmic formula, which attempts to replicate the professional librarians' database selection strategy, was developed in the summer of 1998 from the above studies, and from the discussions and inputs of several UIUC librarians consulted.

Database selection algorithm

Topic/subject category term +
/broad subject +/Information Type +
/level +/database type => database

This formula states: topic keywords or specific subject terms combined with broad subject area, format of information sought, type of content sought, level of information need, and/or database type will determine the databases relevant to the topic.

The following process demonstrates how a librarian's decision-making model is translated to an algorithm for database selection:

- (1) *Topic keywords*: a user who needs to search electronic databases should have a topic or specific subject to research. He or she can often express that topic in appropriate free-text keywords or noun phrases. For example, a user would express his or her topic, "The position of women and the attitudes towards women in Hispanic American community", using the keywords "women" and "Hispanic American community".
- (2) *Subject category terms*: if the user has difficulty expressing the topic in free-text terms, a browse list of the detailed subject terms is an appropriate alternative.
- (3) *Broad subject*: when a librarian is approached for assistance, the librarian usually first tries to determine the broad subject area of interest to the user. For example, the topic "cloning" can have many aspects, such as technical/medical

- (how to clone), social effects of cloning, and legal aspects of cloning.
- (4) *Information type*: the librarian has to determine in what format(s) and/or type(s) of content the patron is interested, such as newspaper articles, magazine articles, journal articles, conference proceedings, government documents, patents, dissertations, and trade publication, and/or biographical information, current events, historical information, and statistical information.
 - (5) *Level*: often the librarian also will try to characterize the research level of the patron's interest, e.g. the patron needs to do a brief paper for Freshman rhetoric; the patron is working on a senior class design project; the patron is starting research on a thesis topic; the patron is a faculty member researching an upcoming article.
 - (6) *Database type*: based on the information obtained from the reference interview and knowledge of what is available in the library, the librarian decides which type of database is suitable for the patron, such as bibliographic, full-text, directory, or statistical.
 - (7) *Database (the result)*: combining all the information obtained from the above steps, the librarian will be able to point the patron to the database(s) relevant for the topic.

Having analyzed the decision-making process, our next step was to convert the above database selection algorithm formula into a computer decision process, and to identify the data needed as input to that process. The specific strategy developed is shown graphically in Figure 1.

Data gathering

Having settled on the overall design of the prototype tool, our next step was to gather data about each resource and assign characteristics. We used the following approach to collect the data required.

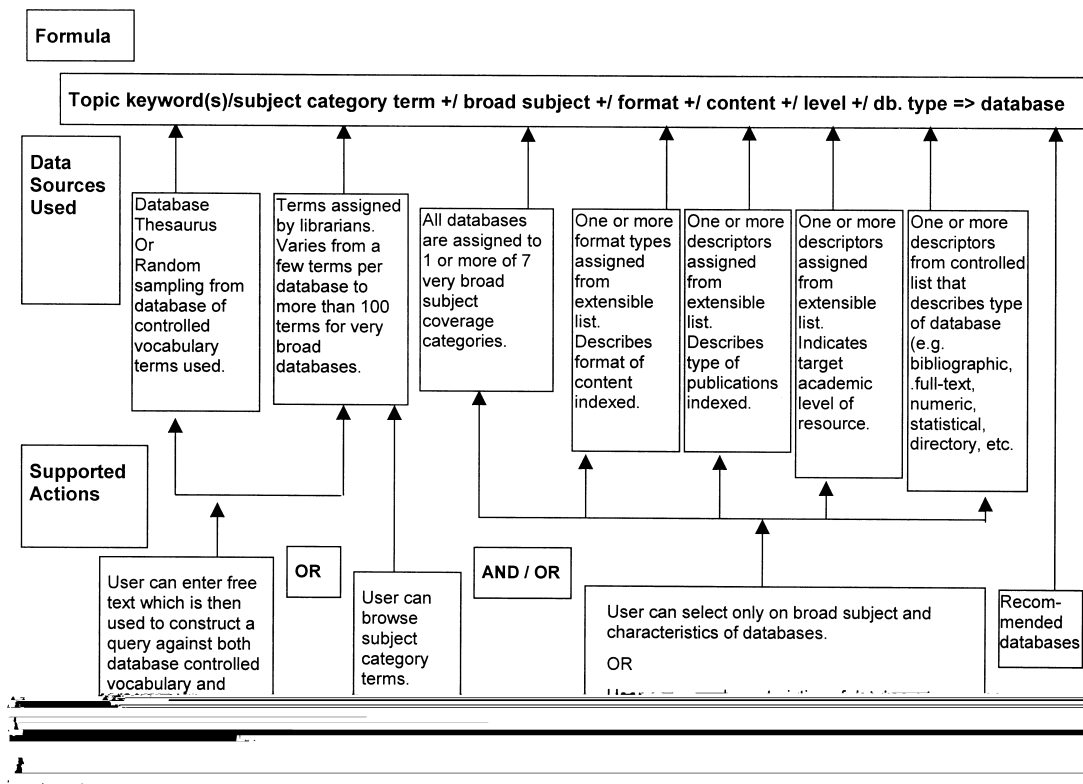
- (1) *Assigning characteristics to each database*
Every database has unique features – different subject and time period coverage, varied material formats, content, data types, target audience (academic levels). For example, Contemporary Women's Issues covers different

subject areas from the Public Affairs Information Services (PAIS); Statistical Universe indexes numeric data and information, while PAIS includes bibliographic citations; PAIS includes various materials types, such as maps, conference proceedings, newsletters, book chapters, pamphlets and research reports, while Statistical Universe covers only statistical information sources and government documents. How could we equally and objectively describe each database in a cost-effective and efficient manner?

We carefully studied and characterized selected databases, settled on standard criteria, developed controlled vocabulary for the database characteristics of interest, and created a template to facilitate data gathering. Figure 2 shows the most current version of the template used. In the Database Type field of the Template, we tried to include the types of databases available. In the Academic Level field, we presented selection options involving both academic levels and academic subject areas. For the Special Contents field, we included the categories which represented the types of content users usually asked for. The Special Formats field outlined most types of publications included in various databases. For the field of Broad Subject Areas, we thought it was important to keep the list simple, clear and distinguishable. The controlled vocabulary for each field was based on what patrons are likely to use (rather than database classification purposes). This process helped keep the consistency of the database reviewing process, while at the same time, simplifying the workflow – the process of mass data collection and data entry. Additionally we consulted publications that described the databases available at UIUC – most notably the Gale Directory of Databases (Kumar, 1999) and DialogWeb (<http://www.dialogweb.com/index.html>). These established authorities facilitated the process of characterizing the databases and helped to insure consistent treatment of the resources.

As noted in Figure 2, multiple terms could be assigned for most database characteristics, such as database types, academic level of the database, special contents indexed, and special formats indexed. We also developed documentation that was attached to the

Figure 1 Database selection computer decision process



Database Review Template (not shown). This documentation prescribed in detail the definitions and the methods through which we assigned terms describing database characteristics. Every database to be included was evaluated carefully based on the criteria on the Databases Review Template.

We encountered some difficulties during the database review process and, in particular, noted inconsistencies in basic database information provided by individual database producers. The basic database information we needed for each database included subject categories, publication types, and the types of information (contents) indexed. Some database producers did not provide any publication type information, such as OCLC FirstSearch's ContentsFirst and ArticleFirst. OCLC provides minimal subject category information as well. Some database producers, such as Cambridge Scientific Abstracts, provide more extensive information on the subject categories, contents indexed and publication types. If we could not find adequate information from a producer's product documentation, we would turn to the

Gale Directory of Databases to supplement. If the *Gale Directory of Databases* did not provide sufficient information, we would further supplement by searching or browsing the databases and extracting needed information. The documentation in support of the Databases Review Template specified procedures for performing these tasks.

With the template and procedures well defined, the process of actually gathering and inputting the data was done by graduate assistants from the UIUC Graduate School of Library and Information Science under the supervision of the authors. A Web data entry form was created to facilitate entry of the data gathered into a structured query language (SQL) database. The authors checked all inputs before data were actually entered into the SQL database.

(2) *Getting thesaurus or controlled vocabulary*
As shown in Figure 1, we wanted to supplement librarian-assigned descriptors and characterizations with terms from the database thesaurus or controlled vocabulary of each

version of the database's thesaurus or controlled vocabulary. Concern was expressed about providing access to these data, even in the context of this research project. Initial responses from the database producers were slow. Smaller producers seemed more agreeable, perhaps because of a recognition that this kind of research might help improve the visibility of smaller, more specialized databases. As we began to receive electronic copies of controlled vocabularies from some producers, other producers became more amenable. We also found it important to be precise and clear about the ways in which these data would be used and to emphasize the fact that the controlled vocabulary would not be directly browsable or downloadable through our prototype tool. When requested we provided this assurance in writing.

We gained useful experience in this process. Among the lessons learned:

- Talk to the right person in the producer organization. If possible, try to reach to the top level of decision-makers.
- Emphasize the need for the data. Give detailed examples of how the problem affects user access to databases, which will eventually affect user awareness and usage of their products.
- Inform producers of the benefit of the project to the producer's products.
- Follow up with e-mail after the initial telephone contact. The e-mail, outlining the objectives of the project, the method we would apply to achieve the goal, and our attitude toward their proprietary data should be sent right after the first phone call. This e-mail provides details in writing and can be forwarded within the company if necessary.
- Follow up with reminder phone calls. These phone calls remind the producer that we are serious about our business and we need their support.
- When possible raise the issue of availability of an online copy of the database's controlled vocabulary (thesaurus) at the time when purchase of a license to the database is being negotiated.

In the end, we generally received a high level of cooperation and support from database

producers. Almost all companies who provided us electronic copies of their controlled vocabularies had to put in some time and human effort to extract the data from their database, since, as mentioned earlier, we were among the first ever to ask for such data. Many producers set up file transfer protocol (FTP) sites, preparing for future updates of their thesauri. Generally this research area is of interest to many producers and many of them are already working on similar projects.

(3) Alternatives to obtaining complete controlled vocabulary

Some databases (e.g. Current Contents from Institute for Scientific Information) do not have true controlled vocabularies. In other cases, an electronic version of the controlled vocabulary is simply not available or not available in a useful format. In other cases the controlled vocabulary is derived from some other source (e.g. Library of Congress Subject Headings) and therefore encompasses a much broader range of topics than does the database itself.

In these cases and when practicable and agreeable to the database producer, we developed sampling approaches to automatically derive a representation of subject headings or subject heading equivalents as actually used in the database. For databases containing a subject descriptor field of some kind and also accessible via Z39.50 we were able to create simple software that connected to the database via Z39.50 and sampled some percentage of the records indexed in a given year (1998 was used). An analysis was performed to look for a point of diminishing return. For most databases accessed, the growth in number of unique subject headings per record sampled began to drop sharply after sampling about 5 to 10 percent of the records in a given year. Further sampling continued to find some unique subject terms, but the return per record sample was reduced. For this prototype, we settled on a 10 percent sampling rate as sufficient. We have not yet extended the analysis to determine change in controlled vocabulary over time, though clearly such work is needed. Also further analyses to determine types of databases that might warrant more extensive sampling should be done.

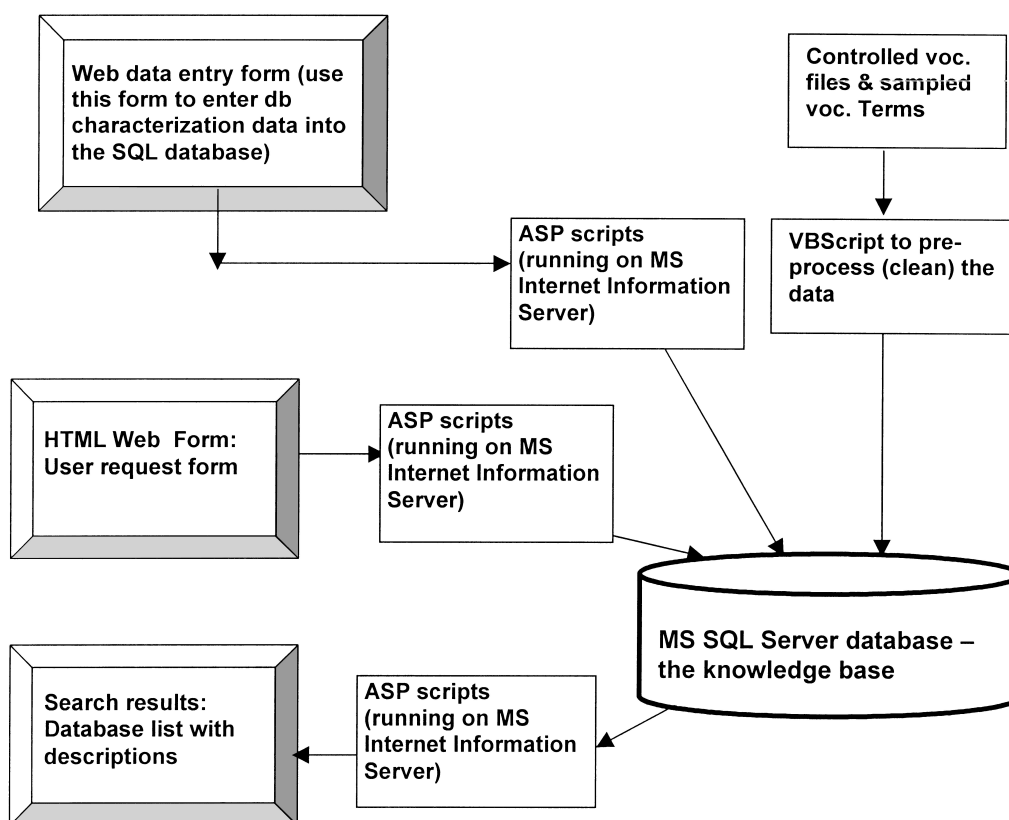
To address Web-accessible databases not also accessible via Z39.50 we are currently porting our Z39.50 scripts to HTTP-compatible scripts. This approach is more problematic since the Web interaction must be customized for each database. Still preliminary work indicates this is a viable, if initially time-consuming approach, and we anticipate adding controlled vocabularies for several more resources using this approach. For Web or Z39.50 accessible databases without any form of a controlled vocabulary at all, we considered using state-of-the-art data mining techniques to create a representative vocabulary for the database from the actual abstracts or other text of the records contained in the database. After assessing the difficulty of this process we decided it was outside the scope of this initial phase of the project. Such databases are only represented in the prototype by their characteristics and the subject category terms assigned by the researchers. Similarly CD-ROM databases for which we were unable to obtain a controlled vocabulary from the producer are also only represented in the prototype by the subject categories that have been assigned and by their general characteristics as entered in the prototype tool database.

Overall, the approach of automatic random sampling of records to obtain a representation of a database's controlled vocabulary has two attractions. First, it tends to result in a selection of controlled vocabulary terms that are actually in use. Second, once the sampling software has been written and verified on a particular resource, it is fairly easy to maintain and update the collection of controlled vocabulary terms over time.

Prototype tool

As described above, a design criterion for this prototype tool was that it would be Web-based and widely accessible (interpreted as browser-neutral).

Figure 3 The architecture (components) of the prototype system



The result screen

Figure 5 shows the top portion of the result screen for a Boolean keyword search for electronic databases appropriate to a particular information need. The topic area of interest to the end-user was summarized as: "Teaching methods employed in distance education". The user was interested in finding journal articles. The search result recommended six databases on the topic. All of the databases recommended were judged as relevant, though clearly some were more relevant than others. Three of the databases recommended by the Smart Database Selector might easily have been overlooked for this topic: National Technical Information Service, Cumulative Index to Nursing and Allied Health, and British and Australian Education Indexes. In particular, British and Australian Education Indexes is a stand-alone CD-ROM resource only accessible at the Education and Social Science Library and therefore not well-known to many library patrons and staff outside of that departmental library.

Conclusion and future plans

We have developed a prototype Database Selection System, which helps users navigate the range of available electronic resources databases and provides advice on the selection of the databases. We have collected knowledge from human experts, and from printed and electronic resources, developed a database selection algorithm – a decision model – and incorporated the algorithm into a computer decision process. This project has given us the opportunity to look into how intermediaries select databases for users by keywords and phrases from their topics, and into other selection criteria that influence database selection decision making.

The prototype system pioneers a new use of database thesauri/controlled vocabularies, combining them with librarian-assigned resource descriptors to facilitate resource discovery from a diverse collection of electronic databases from different information vendors

Figure 4 Database selector HTML Web search forms

The screenshot shows a web browser window titled "Database Selector Project - Microsoft Internet Explorer". The page content is as follows:

Smart Database Selector

[About the project](#) | [Help](#) | [Feedback](#) | [Search for articles](#)

Use one of these forms to suggest or help select the database(s) most suited to your information need.

Select by Keyword(s) (Type in keyword(s) from your topic; also you may specify other criteria that apply)

Keyword(s): AND OR

Other Criteria (optional)

Choose a Broad Subject Choose database level

Choose Type of Information Choose a Database Type Choose Languages

Browse List of Database Topics (You may limit your browse by specifying additional criteria that apply)

Click a letter to browse list of database topic terms: [A](#) [B](#) [C](#) [D](#) [E](#) [F](#) [G](#) [H](#) [I](#) [J](#) [K](#) [L](#) [M](#) [N](#) [O](#) [P](#) [Q](#) [R](#) [S](#) [T](#) [U](#) [V](#) [W](#) [X](#) [Y](#) [Z](#)

Choose a Broad Subject Choose database level

Choose Type of Information Choose a Database Type Choose Languages

Select by Type of Information (Choose a type of information; then specify other criteria)

Choose a Broad Subject Choose database level

Choose Type of Information Choose a Database Type Choose Languages

Type in a time period: After Before e.g 770, 1930

and producers, and on different data-carrier media, such as the Web and CD-ROMs. Additional research has been done to assist in obtaining subject headings or subject heading equivalents as actually used in the databases, for which a thesaurus or controlled vocabulary is not available or not in a usable format. Further analysis needs to be done in this area to establish whether this use of controlled vocabulary is effectual. Automated vocabulary switching methodologies may be needed to better map end-user free-text queries to controlled vocabularies. Alternatively, full-fledged data mining may be required to supplement the use of controlled vocabularies in this way.

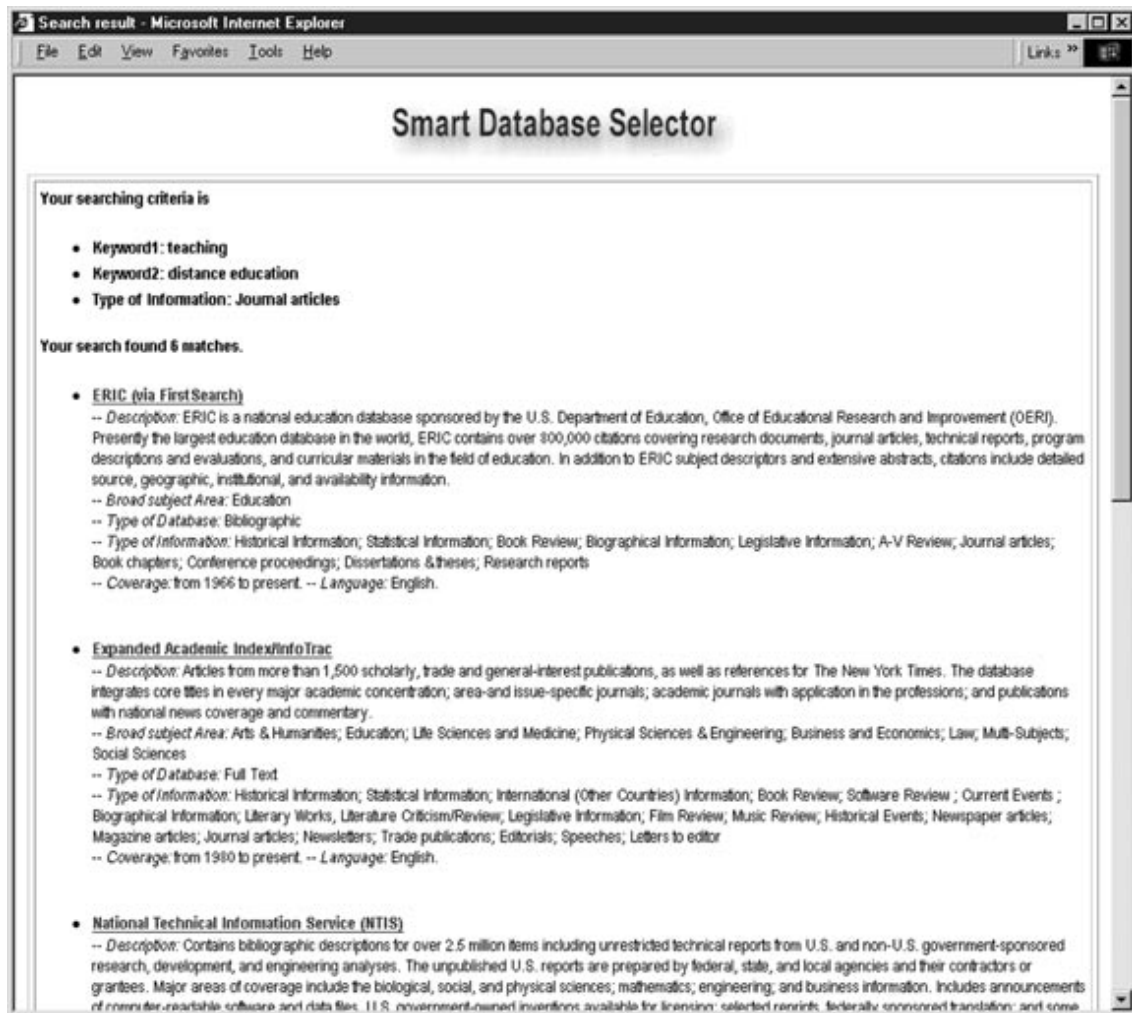
The prototype system architecture has proven robust even when dealing with relatively large data sets (nearly 2 million controlled vocabulary terms are included in the prototype's resource characterization database). The architecture has proven flexible, allowing for easy updating. New database characteristic values could be

added or modified without fundamental changes in design.

An expert system must be evaluated both by experts and users before it can be widely accepted (Zahir and Chang, 1992). We are currently testing the System with professional librarians and staff at UIUC. Included in this testing are usability tests where database recommendations suggested by the human experts are compared with recommendations suggested by the prototype system. We also are limited end-user beta testing and usability testing.

We will compare database transaction logs before and after the implementation of the prototype system to see if database access patterns and usage have been affected. We will also collect information about the nature of the searches performed, including what keywords and phrases are used and what subject terms or other selection criteria were selected to search or refine the searches. These evaluation processes will allow us to perform detailed

Figure 5 Typical search results



system analysis and to make recommendations for improvement and development. We are looking at incorporating features of this prototype in our production database selection utilities at the UIUC library as early as the fall of 2000.

Note

- 1 Professors F.W. Lancaster and Linda C. Smith of the Graduate School of Library and Information Science of UIUC generated a report in May 1997 to the Special Libraries Association (Lancaster and Smith, 1997). The report, entitled "Potential applications of artificial intelligence, expert systems and related technologies within the special library environment", describes the historical development and applications of the database selection expert system technology and was of great help to this study.

References

- Balas, J. (1999), "Exploring some new search tools for librarians", *Computers in Libraries*, Vol. 19 No. 5, pp. 34-7.
- Feldman, S. (1998), "Web search services in 1998: trends and challenges", *Searcher*, Vol. 6 No. 6, pp. 29-39.
- Hightower, C., Reiswig, J. and Berteaux, S.S. (1998), "Introducing database advisor: a new service that will make your research easier", *College and Research Libraries News*, Vol. 59 No. 6, pp. 409-12.
- Hu, C. (1987), "An evaluation of an online database selection by a gateway system with artificial intelligence techniques", Doctoral dissertation, Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign, Urbana, IL.
- Hu, C. (1988), "An evaluation of a gateway system for automated online database selection", *Proceedings of the 9th National Online Meeting, Learned Information*, Medford, NJ, pp. 107-14.
- Kumar, L. (1999), *Gale Directory of Databases*, The Gale Group, Detroit, MI.

- Lancaster, F.W. and Smith, L.C. (1997), "Potential applications of artificial intelligence, expert systems and related technologies within the special library environment", a report to the Special Libraries Association, Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign, Urbana, IL.
- McCarthy, M.V. (1986), "InfoMaster: a powerful information retrieval service for business", *Online*, Vol. 10 No. 6, pp. 53-8.
- Meyer, D. and Ruiz, D. (1990), "End-user selection of database. Part I: Science/Technology/Medicine", *Database*, Vol. 13, pp. 21-9.
- Morris, A., Tseng, G. and Drenth, H. (1994), "CIDA: the expert company information adviser", *Journal of Information Science*, Vol. 20, pp. 247-59.
- O'Leary, M. (1988), "EasyNet revisited: pushing the online frontier", *Online*, Vol. 12 No. 5, pp. 22-30.
- Smith, A.G. (1991), "Kiwinet Advisor: a knowledge base for the selection of online databases", *LASIE*, Vol. 22 No. 1, pp. 4-17.
- Thornburg, G.E. (1987), "LOOK: implementation of an expert system in information retrieval for database selection", Doctoral dissertation, Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign, Urbana, IL.
- Trautman, R. and von Flittner, S. (1989), "An expert system for microcomputers to aid selection of online databases", *Reference Librarian*, Vol. 23, pp. 207-38.
- Tseng, G. et al. (1994), "Expert selection of databases for UK company news", *Online Information 94*, Learned Information, Medford, NJ, pp. 75-85.
- Tseng, G. et al. (1995), "The selection of online databases for UK company information", *Journal of Librarianship and Information Science*, Vol. 27, pp. 159-70.
- van Brakel, P.A. (1988), "EasyNet: intelligent gateway to online searching", *South African Journal of Library and Information Science*, Vol. 56, pp. 191-7.
- Zahir, S. and Chang, C.L. (1992), "Online-Expert: an expert system for online database selection", *Journal of the American Society for Information Science*, Vol. 43, pp. 230-357.