Information Research, Volume 5 No. 3 April 2000
Search features of digital libraries, by Alastair G. Smith Location: http://InformationR.net/ir/5-3/paper73.html © the author, 2000.
Last updated: 21st February 2000

Search features of digital libraries

Alastair G. Smith
School of Communications and Information Management
Victoria University of Wellington
New Zealand

Abstract

Traditional on-line search services such as Dialog, DataStar and Lexis provide a wide range of search features (boolean and proximity operators, truncation, etc). This paper discusses the use of these features for effective searching, and argues that these features are required, regardless of advances in search engine technology. The literature on on-line searching is reviewed, identifying features that searchers find desirable for effective searching. A selective survey of current digital libraries available on the Web was undertaken, identifying which search features are present. The survey indicates that current digital libraries do not implement a wide range of search features. For instance: under half of the examples included controlled vocabulary, under half had proximity searching, only one enabled browsing of term indexes, and none of the digital libraries enable searchers to refine an initial search. Suggestions are made for enhancing the search effectiveness of digital libraries; for instance, by providing a full range of search operators, enabling browsing of search terms, enhancement of records with controlled vocabulary, enabling the refining of initial searches, etc.

Introduction

What search features will be required for users to make effective use of digital libraries? While the term "Digital Library" for collections of electronic materials is relatively new, such collections have been in existence for some time, for instance in large commercial on-line search services such as Dialog, DataStar and Lexis. Over the years, a range of search features - boolean and proximity operators, truncation, etc - have been developed to facilitate access to these collections, and Library and Information professionals have become skilled in their use. This paper attempts to draw on the experience built up by searchers of "traditional" search services such as Dialog, and to suggest the features that should be present in digital library systems.

Why should we be concerned about the search features of digital libraries? Evaluation of electronic information has in the last few years become an important issue in library and information services (e.g. Tillman 1998, Smith, 1997). Librarians and other information professionals, in building collections and advising users of information, will need to be evaluate the strengths and weaknesses of different digital libraries. A previous paper (Smith, 1998) considers evaluation criteria for digital libraries, including searchability.

There are at least three possible models for searching in a digital library. One is that of the on-line Public Access Catalog (OPAC) used in traditional libraries. Another is that of the Internet search engines, such as Alta Vista and Inktomi. A third is that of the traditional on-line search services such as Dialog, Data-Star and Nexis. OPACs are primarily concerned with the searching of bibliographic records, and searching to the level of items, rather than of the full text content of the items. Internet search engines enable searching of content, but across a wide range of documents with no standard structure. Traditional on-line search services, while they started by providing searching of bibliographic databases, have moved increasingly to full text databases (Tenopir & Berglund, 1993) but continue to include standardised bibliographic content such as subject descriptors. Of the three models, traditional on-line search services provide the closest model for the searching of digital libraries, which will be both full text and usually have a standardised bibliographic structure.

This paper will examine features that are desirable for searching in a digital library, and compare these with search features in some current typical digital libraries.

The digital libraries surveyed were electronic collections that have been described as digital libraries in locations such as *Digital Library Research & Development* at the Berkeley Digital Library Sunsite or IFLA's *Digital Libraries: Resources and Projects*, and which had substantial information resources accessible over the Internet, rather than descriptions of information resources. A summary of the search features available at each digital library is presented in Appendix 1.

Search features of digital libraries

The literature of on-line searching includes many treatments of search features. Walker & Janes's (1993) basic text on on-line searching includes many features, including boolean operators, controlled vocabulary, proximity searching, etc. Schwartz (1993) presents a detailed checklist for CD-ROM search products, including different modes of searching, field restrictions, indexing, etc. There are many comparisons of search features of Internet search engines, such as that maintained by Winship (1999), and critical reviews by Poulter (1997) and Schwartz (1998).

Boolean logic

Information professionals have a long history of using boolean logic in search services, and while raw boolean operators can be daunting to inexperienced users and are counterintuitive (Tenopir, 1997), most search systems include a form of boolean logic. Boolean operators appear in search systems in a number of different forms:

- requiring that all the terms entered are present in retrieved items (implied AND) or that at least some of the terms entered are present (implied OR)
- marking terms with + (term must be present, equivalent to AND) or - (term must not be present, equivalent to NOT)
- dialogue boxes into which terms may be entered; the contents of windows being connected by AND, OR, NOT (the specific operators possibly being fixed, or possibly being selectable by drop down menus etc)
- A command mode, where operators are typed into a search statement. This is most flexible, but requires more familiarity by the user.

Nesting of operators (e.g. by brackets in command mode) is an important tool for experienced on-line searchers, for instance to enable different variations of a terms to be OR'd together, then AND'd with the terms representing another concept; but is difficult to implement in other than command mode.

Of the eleven surveyed digital libraries, four had full boolean searching, but none had no form of boolean searching.

Information professionals place some importance on boolean searching because of the sense of control it affords. However it should be noted that boolean searching is best suited to bibliographic databases, and that for full text databases (which is what most digital libraries are) proximity operators are desirable. ("bicycles AND marketing" in a bibliographic database is likely to produce records that relate to marketing of bicycles; in a full text database it may produce records where the terms are widely separated and unrelated).

The OR operator performs a useful function in free text searching - it provides the ability to specify a number of variations of terminology for a concept. One of the barriers to effective searching is that for a given topic, people will use a wide range of terms (Bates, 1998: 1188), so it is important that searchers have the ability to incorporate as many variant terms as possible into their search.

Phrase and Proximity Searching

Phrase searching means that words must be together in a specified order; proximity searching is where a looser linking is specified, for instance that the terms must appear within n words of each other. Many systems implement phrase searching, for instance by placing quotes around the desired phrase, or by the user selecting "exact phrase" from a menu. Proximity searching is less common, but is important in full text databases since it can be used to find records where the search terms occur and are related, e.g. "bicycles NEAR marketing" will find records where the terms occur closely to each other, and are hopefully related. If the extent of proximity is controllable, the user has the flexibility to adjust the extent of linking between search terms.

Five of the surveyed digital libraries had both phrase and proximity searching, four had phrase searching, and only two had no form of proximity or phrase searching (one of these libraries featured relevancy ranking, though.

While proximity operators are powerful, it is worth remembering that proximity is usually a factor in relevancy ranking, so a system with relevancy ranking may *de facto* be allowing proximity.

Relevancy ranking

Newer search systems, especially those of Internet search engines, present results ranked by "relevancy" as determined by an algorithms in the search software.

Relevancy can be based on:

- Frequency of search terms
- Position of search terms

Interestingly only six of the surveyed libraries used relevance ranking, despite its popularity in the Internet environment. In general though, it was difficult to determine how the relevancy was arrived at. This echoes studies of Web search engines, where commercial considerations mean that information about ranking algorithms are not widely available (Courtois *et al*, 1999).

Relevancy ranking performs some of the functions of a boolean AND and of proximity operators: usually the highest ranked records will be those where the terms appear in the same record, and close together. However practical experience shows that some form of NOT is desirable for effective searching on relevancy ranked systems: a search for "cycling" in the sense of pedal cycling may be more effective if the items with the term "motor" can be eliminated.

Browsing of indexes

Analyses of search success (Drabenstott & Weller, 1996) usually find that a significant number of errors are due to mistyping of search terms. This issue can be addressed by allowing the searcher to browse and select terms from alphabetical indexes of terms in the database - this is particularly useful for names of people, where numerous variations of spelling and forms of name can occur. In Dialog this is implemented with the EXPAND command, and is widely used by experienced searchers.

Only one of the surveyed digital libraries provided the ability to browse indexes, although four others provided limited implementations, for instance within a specific collection, or browsing of author/title lists. Given the contribution that index browsing makes to search effectiveness, it is surprising that this is not more widely implemented.

Truncation

Another way for searchers to address variations in forms and spelling is truncation. This can be both internal (wom?n to find the terms "woman" and "women") and external (comput? for computing, computer, etc). Systems can allow automatic truncation, so a search for "comput" will retrieve all terms starting with that term, or use algorithms that search for related plurals, variations, etc from a dictionary. Automatic truncation is useful, but can result in unwanted records, for instance a search for "factor" in mathematics may usefully retrieve "factorial", but less usefully retrieve "factory".

Six of the surveyed libraries provided truncation, and three provided a form of automatic truncation.

Field searching

The traditional bibliographic record is structured into fields - author, title, descriptor, etc. The ability to search these fields separately is important, for two reasons:

- Different types of data require different tactics - titles are suited to keyword searches, while an author search may be better performed by browsing an index.
- Restricting a search to a field can make the search more specific - a search for "Wells" in the author field will separate items relating to HG Wells from those relating to holes in the ground.

Seven of the surveyed libraries enabled searching in specific fields, though one of these was limited. Searching of specific fields is difficult to implement in digital libraries that consist of collections with disparate structures; for instance American Memory allows field searching in specific collections, but does not allow field searching in an overall search of the digital library.

Extent of searching

The extent of material in a database or collection that is searchable may vary. A system may allow searching only of bibliographic data, or of the full text of documents. If a printed resource has been digitised, there can be variations in the material that is available for searching, for instance articles in a journal may be searchable, but not letters, advertisements etc. Searching on commonly occurring words (stopwords) may not be possible. In some cases this can create problems: at one stage in the development of a search system that the author was involved with, users were frustrated in getting access to a magazine entitled *More* because the term "more" was a stopword.

It can be desirable to limit the extent of searching - by restricting to bibliographic data, the relevancy of retrieved records may be higher.

Six of the surveyed libraries searched the full text of documents, three were primarily databases of images with descriptions that were fully searchable, and only two did not have full text searching, in both cases because the text of the documents was distributed, or in a non-text format like PDF. An interesting approach to the problem of full text searching of distributed collections of full text is the NZ Digital Library, where searching indexes are built from the document, but the actual documents are left at their original sites (Witten *et al* 1996.

In two of the digital libraries surveyed (*Early Canadiana on-line* and *Making of America*) a distinction needs to be made between images of full text and text produced by Optical Character Recognition (OCR). The OCR'd text was not displayed, because of its lack of quality, but was used for search purposes. While this extended the potential of full text searching, it raised the issue that errors in the OCR'd text might be influencing retrieval accuracy.

Case sensitive searching

Case sensitive searching is uncommon in bibliographic systems, but is often implemented in systems designed for file searching (e.g. in UNIX where case is significant). For this reason it is sometimes available in digital library systems. Three of the surveyed digital libraries implemented case sensitive searching.

Controlled vocabulary

In bibliographic systems the use of subject headings or descriptors from a controlled vocabulary (a thesaurus or list of subject headings), to describe document content is common. By searching on a well designed and implemented controlled vocabulary, users can improve results. However there can be problems if the terms used in the controlled vocabulary are not obvious to users, and there are problems with incorporating new and changed terminology into a controlled vocabulary system. While there has been ongoing debate in the information retrieval community about the relative effectiveness of controlled vocabulary searching versus natural language, in practice both approaches complement each other and are required for effective searches (Bates 1988).

Five of the surveyed digital libraries implemented a controlled vocabulary, and one included a limited implementation.

Language translation

In systems that incorporate materials in different languages, or accommodate users who speak different languages, translation may be incorporated. Technical problems mean that this has not been widely implemented in the past, language translation is likely to be more common in the digital libraries of the future, since networks make materials available in different language regions.

Three of the digital libraries addressed language issues: *Early Canadiana on-line* had both English and French subject headings, effectively enabling multilingual searching. Perseus stored material in English and in Greek (displayable in Greek characters with special software), and NCSTRL contained material in several languages, but did not standardise subject terms, so that an exhaustive search would require searching using all the possible languages.

Date/range searching

A common way of restricting a large search set is to restrict to a particular date range. Most commonly this is used to restrict to the more recent materials.

Range searching is clearly a useful feature for digital libraries that contain historical materials. Five of the eleven digital libraries surveyed had some form of date searching or limitation.

Refining of initial search

Few searches can be completed in one command. Usually a heuristic process takes place, with the searcher refining their search in response to the results obtained. In a sophisticated search, a searcher may build up a number of sets which are used as "building blocks" in developing flexible search strategies (Quint, 1991).

Given the value of this approach, it is surprising to discover that none of the digital libraries surveyed allowed the development of a search with sets. Of course, in most cases it was possible to return to the search window and add extra commands, but this does not permit the development of sophisticated search strategies, and limits the overall effectiveness of searching.

Related items

If the user finds a relevant item, it can be useful if they can search on like items, for instance those that use keywords from that item. This is an implementation of the "pearl-growing" strategy (Quint, 1991).

Only two of the eleven libraries implemented some form of related item searching.

Multimedia searching

Searching for images, audio, and video usually takes place by searching text descriptions. However the "holy grail" of multimedia searching lies in the direct searching of images, for example for a defined shape, or sound. These are mostly found in experimental systems, for example, Chang *et al*, 1997, McNab *et al*. 1997. In the surveyed digital libraries, the only true multimedia search feature was the NZDL's music search.

Advanced and basic search facilities

Providing search facilities for users of differing abilities involves compromises. A fully featured search interface for experienced users is likely to confuse less experienced searchers, while a simple search interface will not provide the power required for complex queries. While it is possible to have a simple query interface (a box with "enter search terms here") which power users can extend by commands, it is generally preferable to have defined basic and advanced search interfaces.

Only six of the digital libraries had different modes of searching, which probably reflects the fact that most were not offering a great range of search features. If digital libraries increase the range of search features, it will be necessary to introduce different modes of search interface.

Display features

As noted above, searching is a heuristic process, and part of the search process is the display of records. Effective searching requires several different display formats. When the search is being developed it is useful for a searcher to see records that are brief, so that many can be scanned quickly to determine the success of the search, and which include terms such as descriptors that can be fed back into the search. Once a useful search result set is obtained, a format is required, for instance with an abstract or other indication of content, that enables selection of items for viewing in full. Some research (Bates 1998 p.1198) indicates that the human search process works in steps of 30: first selecting titles of documents, then abstracts that are about 30 times the size of the title, then to documents that are about 30 times the size of the abstract. This implies that designers of digital library systems should provide display formats that enable users to word from titles to full text in steps of about 30 times the previous display format.

The order of display of items can be important - a standard default on many systems is to display most recent items first, except where relevancy retrieval is used. However in other circumstances author or title order can be useful.

In the Internet environment, the ability to bookmark the URL of a specific document or section of a digital library is useful, and is provided by, for example, *Early Canadiana on-line*. In some other systems a URL is generated on the fly, and is not bookmarkable.

All the digital libraries surveyed provided more than one format of display, however it was not clear that this had been designed from the point of view of facilitating search development, or of selection of items for full viewing.

Where a digital library contains original documents that have been digitised, a choice lies between displaying the image or the OCR'd text. Both have their advantages: the image is more authentic but has bandwidth costs, the text can be more easily read and manipulated, and transferred more rapidly over networks. Some of the digital libraries sampled used OCR text for searching, but displayed images, since it was felt that the OCR was of insufficient quality for display (e.g. Early Canadiana on-line). However a searcher might find it useful to have the choice of which format to display, both for the reasons alluded to, and to aid the development of search strategy.

Help and documentation information

Like other software systems, search systems require help information to be available to the searcher. Issues in help information for the digital libraries surveyed included cases where the help information referred to features that were no longer present, and there are some areas where insufficient information is given, for instance as to how ranking is determined.

All the surveyed libraries had help information.

Conclusion

Traditional on-line search services such as Dialog have developed a wide range of search features, that allow users to perform sophisticated searches for a variety of queries. Although developments in search engine technology (ranking, automatic term expansion, etc) have made it easier for novice users to perform searches without using these sophisticated features, experienced users are likely to require a greater range of search features for carrying out more complex searches. In developing search interfaces for digital libraries, the objective should not be to find the "ideal" search features, but to provide a range of search features allowing different types of searches, and accommodating different search styles and levels of searcher experience. As digital collections grow larger, the level of sophistication required to perform effective searches is likely to increase.

While the eleven digital libraries surveyed provided overall a wide range of search features, none provided the wide range of features that traditional on-line services provide. Digital library designers should consider providing a wider range of features in future versions of their software.

Features that do not appear to be widely available, but which have proved their worth in traditional on-line services, are:

- Full boolean searching, in particular OR operators and nesting, to facilitate the incorporation of alternative terminology for concepts. The NOT operator should be an option, particularly in conjunction with relevancy ranking.

- Proximity, as well as phrase searching. Phrase searching assists finding specific terms, but proximity searching is an effective way to specify that terms are closely related in a full text digital library.
- Browsing of terms in the database. This offers a convenient way for searchers to establish the variant forms of a search term, and may overcome many problems with misspelling.
- Searching of specific fields. This enables searchers to use tactics suited to the nature of different types of information.
- Controlled vocabulary and other metadata. While this is expensive to implement, and there is ongoing debate in the retrieval community as to its effectiveness, at least some classes of searches are enhanced by subject descriptors. The Dublin Core and other metadata projects are leading the way in this area.
- Refining of an initial search. The "building block" strategy has been shown to be effective in on-line databases, but requires the ability to create and save a range of search sets, which can then be refined.
- Related items. The ability to "pearl-grow" searches, feeding terms from relevant items back into iterations of the search strategy, is valuable, and can be facilitated by a related items feature in a search interface.
- Advanced and basic search interfaces. If a useful range of search features is provided, it will be necessary to provide different search interfaces for different types of user and different types of search.
- A range of display modes should be available, to facilitate development of search strategy, and to enable selection of items for further viewing or printing. Searchers should be able to examine the text that was searched, in the case of OCR'd text.

Multilingual indexing and searching are likely to become an issue as digital libraries become a global resource. Ideally a searcher should be able to enter terms in one language, and find relevant documents in the library written in other languages. This can be done by multilingual indexing, such as is done in ECO, but this is expensive, and automatic translation and indexing tools are likely to have to be applied here. It is worth noting in this context that assigning useful subject descriptors in one language can facilitate access to the material by a searcher using who is not familiar with the language.

As digital collections become more common, there is likely to be demand for the search interface be standardised so that users can search a wide range of collections through the same interface. Z39.50 client server approach has been proposed as a solution to this, although none of the digital libraries surveyed appeared to be Z39.50 compliant. In the 1980's the Common Command Language (Klemperer 1987) was proposed as a solution to the problem of differing commands on different on-line search services, but did not achieve success, partly because adherence to a standard risked limiting the development of new search features. The same problem is likely to be an issue in digital library development.

The digital library research community has achieved a wide measure of success in a short time. However there are still lessons to be gained from the "traditional" digital libraries, the on-line search services.

References

- Bates, M.J. (1988) "How to Use Controlled Vocabularies more Effectively in on-line Searching" *On-line* **12**(6), November, 45
- Bates, M.J. (1998) "Indexing and access for digital libraries and the Internet: human, database and domain factors". *Journal of the American Society for Information Science* **49**, 1185-1205.
- Chang, S., Smith, J.R., Meng, H.J., Wang, H. and Zhong, D. (1997) "Finding Images/Video in Large Archives: Columbia's Content-Based Visual Query Project". *D-Lib Magazine*. February. Available at: http://www.dlib.org/dlib/february97/columbia/02chang.html [Accessed 17 Feb 2000]
- Courtois, M.P. and Berry, M.W. (1999). "Results Ranking in Web Search Engines". *On-line* **23**(3) Available at: http://www.on-lineinc.com/on-linemag/OL1999/courtois5.html [Accessed 17 Feb 2000]
- Drabenstott, K.M. and Weller, M.S. (1996) "Handling spelling errors in on-line catalog searches". *Library Resources & Technical Services* **40**(2), 113-32
- Klemperer, K. (1987) "Common Command Language for on-line interactive information retrieval". *Library Hi Tech* **5**(4), 7-12
- McNab, R.J., Smith, L.A., Bainbridge, D., and Witten, I.H. (1997, May) "The New Zealand Digital Library MELody inDEX". *D-Lib Magazine* Available at: http://www.dlib.org/dlib/may97/meldex/05witten.html [Accessed 17 Feb 2000]

- Poulter, A. (1997). "The design of World Wide Web search engines: a critical review". *Program* **31**(2), 131-145.
- Quint, B. (1991, July) "Inside a Searcher's Mind: The Seven Stages of an on-line Search - Part 2" *On-line*. **15**(4), 28-36.
- Schwartz, C. (1993) "Evaluating CD-ROM products: yet another checklist". *CD-ROM professional* **6**(1), 87-91.
- Schwartz, C. (1998) "Web search engines". *Journal of the American Society for Information Science*. **49**(11), 973-982.
- Smith, A.G. (1997) "Testing the Surf: Criteria for Evaluating Internet Information Resources". *The Public-Access Computer Systems Review* **8**(3). Available at: http://info.lib.uh.edu/pr/v8/n3/smit8n3.html [Accessed 17 Feb 2000]
- Smith A.G. (1998) "Criteria for Evaluation of Internet Resources in a Digital Library Environment". In: *Proceedings of the First Asia Digital Library Workshop, Hong Kong, 6 - 7 August 1998*.
- Tenopir, C. and Berglund, S. (1993) "Full-text searching on major supermarket systems: DIALOG, DATA-STAR and NEXIS". *Database* **16**(5), 32-42.
- Tenopir, C. (1997). "Common end-user errors". *Library Journal* **122** (8) 31-32
- Tillman, H. (2000) *"Evaluating Quality on the Net"*. Available at: http://www.tiac.net/users/hope/findqual.html [Accessed 17 Feb 2000]
- Walker, G. and Janes, J. (1993). *On-line Retrieval: a dialogue of theory and practice*. Englewood, CO: Libraries Unlimited.
- Winship, I. (1999). *Web search tool features.* Available at: http://www.unn.ac.uk/features.htm [Accessed 17 Feb 2000]
- Witten, I.H., Cunningham, S.J. and Apperley, M.D. (1996, November). "The New Zealand Digital Library Project". *D-Lib Magazine* Available at: http://www.dlib.org/dlib/november96/newzealand/11witten.html [Accessed 30 July 1999]

Appendix 1: Search features of digital libraries - a selective survey

Digital Libraries surveyed

American Memory	http://memory.loc.gov/ammem/amhome.html
Early Canadiana On-line	http://www.canadiana.org/
ETDC: Electronic Thesis and Dissertation Collection	http://www.theses.org/vt.htm
Everglades Digital Library	http://everglades.fiu.edu/library/
Making of America	http://www.umdl.umich.edu/moa/
NCSTRL: Networked Computer Science Technical Reference Library	http://www.ncstrl.org/
NZDL: New Zealand Digital Library	http://www.nzdl.org/
On-line Archive of California	http://sunsite2.Berkeley.EDU/oac/
Perseus Project	http://www.perseus.tufts.edu/
SCRAN: Scottish Cultural Resources Access Network	http://www.scran.ac.uk/
Timeframes	http://timeframes.natlib.govt.nz/