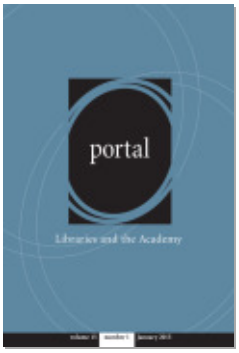
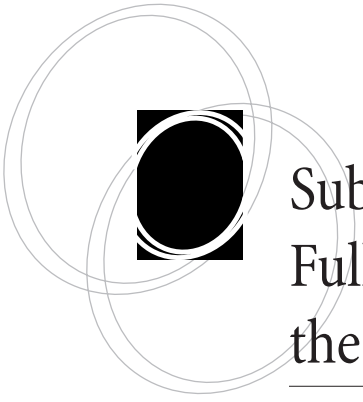
Subject Retrieval from Full-Text Databases in the Humanities

John W. East



portal: Libraries and the Academy, Volume 7, Number 2, April 2007, pp. 227-241 (Article)

Published by Johns Hopkins University Press
DOI: 10.1353/pla.2007.0018



Subject Retrieval from Full-Text Databases in the Humanities

John W. East

abstract: This paper examines the problems involved in subject retrieval from full-text databases of secondary materials in the humanities. Ten such databases were studied and their search functionality evaluated, focussing on factors such as Boolean operators, document surrogates, limiting by subject area, proximity operators, phrase searching, wildcards, weighting of search terms, limiting by type of document, controlled vocabulary indexing and ranking, and display of search results. The author suggests ways in which full-text searching might be improved, whether by enhancement of database records, by introduction of enhanced search functionality, or by the education of searchers in more effective search techniques. The conclusion is that current digitization projects are not producing databases that meet the needs of scholars.

Introduction

The development of full-text databases in the humanities antedates the Web browser by many years, but there is no doubt that the Web has significantly accelerated development in this area. A recent study reports that "in the humanities, there is a growing appreciation of the ability to interrogate the full text of large corpora, especially in literary, linguistic, and cultural fields of inquiry."[1] Most of the significant early developments were in the digitization of primary materials. However, in recent years, large collections of secondary materials have also become available, and this is a trend that seems likely to continue. The purpose of full-text databases of primary materials is to provide texts that would otherwise be hard to access and also to facilitate linguistic and thematic analyses of those texts in ways that have not previously been possible—or only possible by expending considerable amounts of time manually searching and collating results. These primary databases vary enormously, both in the nature of the materials that they contain and the ways in which they are used by researchers.

portal: Libraries and the Academy, Vol. 7, No. 2 (2007), pp. 227–241.

It would be difficult to make general comments about the search functionality that is appropriate for such databases; the *Duke Database of Documentary Papyri* will not be searched in the same way as *The New York Times Digital Archive* or *Early English Books Online*. Jeffrey Garrett has provided an overview of the problems involved in searching full-text collections of primary sources, and the digital libraries surveyed by Alastair Smith were also mainly collections of primary materials.[2]

However, full-text databases of secondary materials serve much the same function in all disciplines. At the most basic level, they make available in convenient electronic format journal articles and books required by researchers, and that was usually the main purpose behind their creation. Digitization has also created opportunities for resource discovery (searching), and the database creators have recognized this and developed search engines with varying degrees of sophistication.

These full-text secondary databases have mushroomed in recent years, and even larger projects are in the pipeline. For example, Google has begun work on digitizing the collections of large libraries, the publisher Macmillan has announced plans for its own searchable repository of digital book content, MSN Book Search is working with the British Library to digitize books from that collection, and the Open Library Consortium is also working on large-scale digitization. Yet despite all these ambitious digitization projects, little attention has been paid to the question of how researchers will be able to effectively search these large electronic collections of secondary materials.

Yet despite all these ambitious digitization projects, little attention has been paid to the question of how researchers will be able to effectively search these large electronic collections of secondary materials.

As Helen Tibbo pointed out more than 10 years ago, "Computers and automatic concordance programs have certainly not solved the problems of subject indexing and library cataloging for the humanities. As the body of humanistic scholarship grows at an unprecedented rate, the storage and retrieval challenges facing indexers, catalogers, and users of humanistic materials are exacerbated."[3]

The problems of information retrieval from full-text databases have been well documented since at least the 1980s,[4] and, yet, there still seems to be an assumption that full-text searching is the solution to all information retrieval problems. As Peter Jackson and Isabelle Moulinier warn, "Exaggerated claims for these technologies, which suggest that computer programs can somehow 'understand' the meanings of words or the intentions of users, are counterproductive."[5]

The focus of this paper is on the issues surrounding subject retrieval from full-text databases of secondary materials in the humanities. Although many of these issues are applicable to other disciplines, it has long been recognized that subject retrieval in the humanities is inherently more difficult than in the sciences. As Tibbo puts it, "Humanists face the greatest information retrieval challenges of all scholars. In addition to being large, their literatures are almost always dispersed in subject matter, format, and age, and characterized by 'fuzzy,' often metaphorical language."[6] As a starting point, a number

of existing full-text databases of secondary materials are examined in order to assess the effectiveness of the subject retrieval mechanisms currently available. This will be followed by a discussion of the ways in which full-text searching might be improved, whether by enhancement of database records or by introduction of enhanced search functionality or by education of searchers in more effective search techniques.

Methodology

The databases that were chosen for study fall naturally into two categories: those containing digitized journal articles and those containing digitized monographs (books). There are obvious differences between full-text databases that contain smaller, discrete units of text (journal articles) and those that contain lengthy units of text (books), broken up into discrete sub-units (chapters). One of the aims of the study was to establish how these differences affected information retrieval.

A number of subject searches were conducted across all the databases in order to compare their performance. As the databases have differing content, it would be impossible to compare retrieval using any quantitative measures. It was only possible to test the search features available and attempt to assess their usefulness.

The journal article databases studied were JSTOR, Infotrac Expanded Academic, Oxford Journals, Blackwell Synergy, and Cambridge Journals Online. The monograph (book) databases studied were Google Book Search, A9.com, Oxford Scholarship Online, NetLibrary, and Ebrary. All of these databases contain full-text secondary materials covering a wide range of subject areas. The purpose of the study was to focus on the usefulness of the databases to the humanities researcher, but many of the conclusions will be relevant to researchers in other areas also.

Search Functionality: Current Situation

There are some queries that can be reduced to uncommon words or phrases, and such search terms can be retrieved very effectively from a full-text database. Indeed, it is here that the full-text database comes into its own, offering the researcher the opportunity to find needles in haystacks in a way that would have been simply impossible for previous generations of scholars. However, in many cases, the researcher can only express a query in words that will occur hundreds of thousands of times in a full-text database. How can such searches be conducted effectively?

Boolean Operators

Information professionals would normally consider that, for effective searching of a database, it is essential to be able to construct full Boolean searches, with the standard AND, OR, NOT operators and the ability to bracket (nest) these terms. All of the journal article databases and most of the book databases in the study permit the use of standard Boolean operators, although only some allow the construction of more complex bracketed searches.

However, the usefulness of Boolean operators to the average researcher is debatable. George Buchanan and colleagues found that only seven of the 18 humanists whom they

studied had any awareness of Boolean searching.[7] Deborah Shaw found that some of the graduate literature students in her study "were uncomfortable with focused queries involving Boolean AND, as they felt this approach failed to retrieve all relevant items."[8] One student commented, "I just feel like if I limit to two terms that I could easily lose out on a lot of information and then not know what I missed out on."[9]

Database designers understand that most users have little understanding of Boolean operators, so they tend not to advertise this search strategy.

Database designers understand that most users have little understanding of Boolean operators, so they tend not to advertise this search strategy. A more sophisticated approach is to provide implicit Boolean searches, with search menus offering options like those on JSTOR: "all of these words" and "at least one of these words." Google Book Search uses a similar approach on its advanced search screen.

Document Surrogates

Although it may be useful to retrieve documents in which specified search terms appear, if those documents are large (for example, a whole book), the relevance of many of the retrieved references may be low or nil. On the other hand, occurrence of the search terms in a summary of a document may be a higher indication of relevance. For this reason, most of the journal article databases in this study offer the option to limit the search to article abstracts only. However, many journals in the humanities do not contain author abstracts, so this feature is often of limited use. Of the journal article databases studied, Infotrac is the only one that creates its own abstracts when no author abstract is available.

Of the book databases, Oxford Scholarship Online offers "specially commissioned" abstracts at both book and chapter level, and these can be separately searched. Interestingly, the quick search option on that database actually ignores the full text and searches only the abstracts, titles, and keywords, reflecting the assumption that searching document surrogates provides more useful results. The other book databases contain no abstracts, but it is interesting to note that the default basic search in NetLibrary does not search full text but restricts itself to the data that would be available in a library catalog record: title, author, and the Library of Congress (LC) subject headings. The library catalog record is, of course, a very basic document surrogate for a book.

Perhaps the most useful surrogate for a scholarly book is a review in a scholarly journal. If this is the case, the book reviews in the full-text journal article databases may act as an effective, if indirect, tool for discovering relevant monographs in the full-text book databases.

Limiting to Sources in Broad Subject Areas

In large, multidisciplinary full-text databases, many search terms will retrieve irrelevant results from other disciplines, especially when a wildcard has been used to truncate terms. Limiting the search to source documents identified as belonging to the broad

discipline in which the researcher is interested can at least mitigate this problem. There is the danger of missing relevant material in other disciplines, however, as research today is increasingly cross disciplinary.

To offer this search facility, the database creator must have some way of assigning documents to subject areas without laborious and expensive intellectual input. In the case of a database of journal articles, this may not be too difficult since most scholarly journals are explicitly associated with one or possibly more subject disciplines. JSTOR, Blackwell Synergy, Oxford Journals, and Cambridge Journals Online all provide the ability to limit searches to broad subject areas, although the breakdown of disciplines is more detailed in some databases than others.

In the case of the book databases, the problems are far greater. Oxford Scholarship Online is a relatively small database of recent publications from one publisher, restricted to four broad subject areas, each of which can be searched separately. With the other book databases, there is no facility to limit searches to specific disciplines.

Both NetLibrary and Ebrary allow the user to construct Boolean searches that combine terms from the full text of books with terms from the Library of Congress subject headings assigned to the books in the database. In some cases, this may provide a useful way of limiting a full-text search to a broad subject area; but, in order to be used effectively, it requires a better understanding of LC subject headings than most searchers would possess.

Proximity Operators

When searching for two terms using a Boolean AND operator, it seems obvious that, if the search terms appear on the same page of a document, the document is likely to be more useful than one in which the terms appear five pages apart. Of the journal article databases studied, all except Oxford Journals and Blackwell Synergy support proximity operators. With book databases, in which the documents are so much larger, proximity becomes even more important. Ebrary is the only one of the book databases studied that supports explicit proximity operators, but Google Book Search, A9.com, and Oxford Scholarship Online will retrieve documents in which search terms appear on the same page—an implicit use of proximity operators. NetLibrary does not support proximity operators, and the results of Boolean AND searches on that database are often low in relevance.

Phrase Searching

The ability to search bound phrases is really just a specialized example of the use of proximity operators. It is a facility that all of the databases examined offer and can be very effective in cases in which a search can be expressed as a bound phrase—or even in a series of phrases, each being a variant of the other.

JSTOR prompts the less sophisticated user by providing a search box labelled "the exact phrase" on its advanced search screen, and Google Book Search has a similar feature. In other databases, the searcher needs to understand the use of quotation marks to construct phrase searches, so only the more sophisticated searcher is likely to use this feature.

Wildcards

Truncation symbols (wildcards) are a familiar feature of databases. They are a fairly crude tool since truncation of a search term may retrieve many unrelated terms that begin with the same group of letters. In large full-text databases, these problems are aggravated by the sheer number of words being searched.

Of the databases studied, Google Book Search, Ebrary, and Blackwell Synergy are the only ones that do not support wildcards. However, with Blackwell Synergy "the Search Engine uses natural language analysers to determine a word's stem. Stems are not just truncated versions of words, i.e. tin is not the stem of tint. If you enter a search term containing the word nursing, the Search Engine will return results containing nurse as well as nursing."[10]

Some of the other databases also offer morphological searching (stemming). In Infotrac Expanded Academic, the relevance search option does not accept wildcards but instead searches for variant spellings. Oxford Scholarship Online also uses an automatic stemming algorithm. In NetLibrary, the searcher can use the double asterisk, which searches for all forms of a term (for example: drive** searches for drive, drove, driving, and driven).

Weighting ("Boosting") Search Terms

Some databases allow the searcher to differentiate between the values of search terms. The searcher can ask to retrieve documents containing term A and term B but specify that term A is more important than term B. The search engine then ranks the results and displays first the documents in which term A occurs most frequently. Of the databases studied, JSTOR and Cambridge Journals Online (which both use the Apache Lucene search engine) offer this facility. Thus, when constructing the search, the researcher is able to specify the relative importance of search terms by means of a numerical value; for example, *colonialism^4 AND france* tells the search engine that, although both terms are required, the first term is four times more important than the second.

Limiting by Type of Document

Most of the journal article databases display some ability to limit searches to specified types of documents. JSTOR offers the most useful range of options since it can limit to article, review, opinion piece, or "other." Infotrac Expanded Academic can limit searches to refereed journals (but not to the type of article within those journals). Oxford Journals offers the option to limit to review articles only, whereas Cambridge Journals Online has an option to exclude book reviews from results.

It is questionable whether researchers in the humanities would want to exclude book reviews from their searches, given that book reviews are such an important form of scholarly communication in the humanities. As Tibbo has pointed out, "Historians have relied upon the book reviews in…journals as a source of bibliographic information on new monographs. …Critical, scholarly reviews are useful for evaluating the merits of individual works while placing the texts within particular traditions and schools of thought."[11] However, the ability to exclude news items and other minor notices is certainly a useful feature.

In full-text databases containing both scholarly and non-scholarly publications, the ability to limit to scholarly, academic material is an extremely useful feature. This is an issue that arises in book databases that include non-scholarly material, such as a9.com, Google Book Search, and NetLibrary; and these databases offer no facility to limit the search to scholarly content.

Controlled Vocabulary Indexing

Previous studies have established that the most effective way of searching databases in the humanities is to combine free-text searching with the use of controlled-vocabulary indexing.[12] Unfortunately, of the journal article databases under examination here, Infotrac Expanded Academic is the only one that assigns controlled vocabulary indexing terms to each document.

Of the book databases, Oxford Scholarship Online assigns specially commissioned keywords at book and chapter level, but there is no indication that these keywords are based on any controlled indexing system (thesaurus). NetLibrary and Ebrary both provide Library of Congress subject headings, but these are assigned to the book as a whole and not to individual chapters or sections.

Automatic Ranking of Results

In very large, full-text databases, many searches will inevitably retrieve a large number of results. Most databases will rank the results in some way to assist the searcher. JSTOR and Cambridge Journals Online both use the Apache Lucene search engine, and both databases adopt a similar approach to ranking. The method used by JSTOR can be summarised as follows: relevance is based upon the number of occurrences and frequency of occurrence of the search terms, weighted to give higher rankings to matches of uncommon terms and higher rankings to search terms occurring in the title or abstract of the article. Additionally, results are weighted by article type in the following descending order of preference: full-length articles, book reviews, editorials, and news/miscellaneous items.

Sample searches of JSTOR suggest that the relevance rankings are indeed helpful, but there are many questions that could be asked. Is it effective to give higher weighting to matches of uncommon terms? Does "uncommon" mean uncommon across the whole multi-disciplinary database or uncommon within the set of retrieved documents? Is it helpful to give higher weighting to occurrences of terms in the title, given that the titles of many documents in the humanities are not intended to be literally accurate descriptions of the document content? What is the numerical value of these weightings, and on what empirical evidence are they calculated?

Infotrac Expanded Academic has a separate relevance search option. When this option is selected, results are ranked according to where the words appear in the article (title, text, and so on) and how closely they match the original search terms (exact match, word variant, and so on) and by how often the search term is mentioned in relation to the length of the article.

Oxford Journals uses a ranking system that also considers number and frequency of occurrence of search terms, but it then apparently gives preference to articles in which any

one of the search terms appears in the title or abstract. This does not seem helpful in the humanities, in which many articles do not contain abstracts and titles are not necessarily descriptive of content. Blackwell Synergy ranks results by relevance but does not give any explanation of the criteria used. With Google Book Search, it appears that proximity of search terms and frequency of their appearance are the factors used in ranking. The ranking system used by A9.com is not obvious and does not seem very helpful.

Oxford Scholarship Online offers a relevance ranking weighted according to the number of matches for the search term in significant parts of the text. It treats each printed page as a separate document, however, and makes no attempt to associate those pages with the chapters or complete books to which they belong. The results, therefore, do not indicate the chapters (or books) that contain the highest number of page matches. NetLibrary defaults to ranking by date of publication, but it offers an option to rank by top matches. This seems to be a simple count of the number of occurrences of the search terms in each book. Ebrary uses a similar ranking system as its default option.

Displaying Results

When the search engine has retrieved matching documents and applied ranking algorithms to sort them, the final step is to display those results to the searcher. At this point, we reach a crucial stage in the human-computer interaction—the searcher finally has to assess which documents are likely to be worth pursuing. It is essential, therefore, that the display of the results facilitates rapid and informed assessment. If not, researchers may become discouraged and not venture beyond the first screen of results.

At this point, we reach a crucial stage in the human-computer interaction—the searcher finally has to assess which documents are likely to be worth pursuing.

Some of the databases studied calculate the relevance of each item as a percentage and display this figure with the results. There is no indication of how this percentage is calculated, and it is highly questionable whether these figures convey anything to the searcher. All databases display the titles and authors of the documents and also usually the source, whether this is a journal or a publisher. All of this information is useful in assessing the relevance of the item, although the point has already been made that titles in the humanities are not always descriptive of subject content.

The next step in assessing relevance is to start viewing the full text of retrieved documents. At this point the researcher will probably be keen to go to the section of the text where the search terms appear, preferably highlighted for quick scanning. Oxford Journals provides in its initial results display several lines of text for each document, showing occurrences of search terms in context, but it is the only one of the journal databases to offer this facility. None of the journal article databases displays search terms highlighted in the full-text articles, although JSTOR lists the page numbers on which search terms appear.

Google Book Search shows one occurrence of the search terms in context, and there is a link that will display other occurrences of the search terms in the book. This makes it

easy to establish how relevant the book is to the search. When full details of a book are displayed, it is possible to view a publisher's synopsis and also the contents and index of the book—much like the process of browsing books on the library shelves.

A9.com has a similar approach and can also display the table of contents and index. By linking to the entry for the book on the Amazon.com database, it can often provide a synopsis and excerpts from reviews. The records on Amazon.com will also display statistically improbable phrases and capitalized phrases. These are an attempt to use automatic text analysis to highlight subject content, but it is not certain that they serve any very useful purpose. Oxford Scholarship Online allows the user to view the abstract of the chapter in which the search terms appear.

Examining retrieved documents in NetLibrary is particularly cumbersome since the user has to select an item from the results list and then search it again to see where the terms appear. This second search is a very basic function, which does not accept Boolean operators. When the results from an individual book are retrieved, however, there is an option to display them by chapter, which can help the searcher to pinpoint the most relevant sections of the book. When results are displayed in Ebrary, the searcher must select an item of interest, and this will display the first page where the search terms appear. There is no indication of how many pages in the book contain matches, but clicking on an icon will take the user to the next relevant page.

Search Functionality: Suggestions for Improvement

Having examined the search facilities currently available in full-text secondary databases, let us now look at what might be done to enhance those facilities to better serve the needs of researchers in the humanities.

Record Enrichment

As mentioned above, previous studies indicate that controlled vocabulary indexing significantly improves database retrieval in the humanities. However, the challenges are enormous. As Tibbo points out, "The monograph remains the dominant scholarly vehicle in the humanities. This means that effective bibliographic indexes in the humanities must not only provide coverage of journal contents but should also index monographic sources as well. This is a much greater task both in terms of locating all relevant items and analyzing the content of much larger and more complex works."[13]

It is clear that with large-scale retrospective digitization projects it is impossible to undertake the extremely expensive process of metadata creation that would be required to provide suitable indexing at this level. There are, however, more modest possibilities based on incorporating existing metadata as we can see in NetLibrary and Ebrary, which include Library of Congress subject headings in their database records. Of course, a few subject headings assigned to a book as a whole cannot provide the depth of analysis that researchers require, but they may be helpful in limiting the search range.

An extension of this approach would be to harvest classification numbers from library catalogs and add them to the metadata for books in digitized collections. This would allow broad categorization of the subject content of books, which could also be useful as a limiting device to refine searches.

Document surrogates have played an essential role in computerized searching of journal articles for over 20 years now. It is a familiar process. The researcher uses keywords to interrogate a database of document surrogates, in which the surrogate for the full article is the article title, an abstract of the article, and probably subject indexing terms as well. Because the searching of document surrogates is a long-established practice, we assume that it is an effective means of information retrieval, even if we have no real empirical evidence to prove this. The fact that the journal article databases in this study normally offer the option to restrict the search to the article abstracts is further evidence of the belief that searching surrogates may be more effective than searching the full text.

If we accept, for argument's sake, the value of searching document surrogates as opposed to the full text, we can immediately see the problems that confront us. Many journals in the humanities have no author abstracts, so document surrogates for journal articles can only be made available if considerable intellectual input is invested in writing abstracts, which is clearly not possible when long back runs of journals are being digitized. In the case of full-text books, the problem is even more extreme. To be useful, document surrogates would need to be created for each chapter.

Are there any ways of creating useful document surrogates without expensive intellectual input? Automatic text summarization already offers the possibility of producing useable extracts of large bodies of text,[14] and this technology might be useful in producing summaries of journal articles and book chapters. Another approach that might be used with scholarly books is to use chapter headings and subheadings and terms in the index to create a basic document surrogate that could be extracted from the digitized text.

Search Engine Enhancement

The TREC (Text REtrieval Conference) workshops, held annually since 1992, have provided an important forum for evaluating and sharing new technologies designed to improve information retrieval from large bodies of text.[15] More recently, the fierce competition between Web search engines to produce the most useful results for searchers has resulted in an unprecedented amount of research into improving database performance. Much of this research depends upon natural language processing and has been usefully summarized by Jackson and Moulinier.[16] The databases surveyed for the present paper utilize, to varying degrees, the developments resulting from this research. As these technologies continue to develop, what features can we expect to find in the full-text databases of the future?

Where it is available, morphological searching (stemming) is still a fairly crude tool in the databases surveyed, often producing irrelevant results. This is particularly problematic if the database offers no facility to turn off the stemming. Intelligent search systems that can identify variant forms of a given word and also distinguish the different meanings that can be assigned to the same word should become more widespread and more efficient.[17] However, the challenges presented by scholarly texts in the humanities, with their specialist and often imprecise and "woolly" terminology, will not be solved easily.

Search engines that can automatically identify synonyms for requested terms will require more than a standard thesaurus if they are to cope with the vocabulary of the humanities. As Stephen Wiberley found, the terms used by humanists "are very imprecise: their definitions are often characterized by change over time or a wide range of meaning, and their referents frequently include a diversity of subjects or objects."[18] To complicate matters further, the secondary materials required for much humanistic research are written in a wide range of languages.

Ranking algorithms are a particularly fashionable topic in research related to Web search engines, and we can hope that developments in that field will flow on to scholarly full-text databases. Having said that, it is clear that there are enormous differences between a Web search engine and a scholarly full-text database. Shaw studied a group of graduate literature students who were searching bibliographical databases and found that they tended to search broadly and "expected to wade through many citations to find the ones needed."[19] One student commented: "I'd rather be a little sloppy and allow for serendipity in things that I don't expect. And frankly…the system is only as good as the people who do the cataloguing. I don't really have that much trust that the descriptors are that good that I can rely on them."[20]

Clearly these are people who are likely to be skeptical about automatic relevance ranking of results. Humanists traditionally spend large amounts of time searching through extensive collections of primary and secondary materials in the hope of finding a few items that will be of use to them. The idea of delegating this process to a machine will be anathema to many humanists, but as we reach a situation in which vast amounts of textual information are stored digitally, traditional research methods may need to be reviewed. It seems wanton to ignore the search capacities that computers offer, but humanists will need to be convinced that the search engines can be trusted at least to reduce the volume of material to be scanned.

It seems wanton to ignore the search capacities that computers offer, but humanists will need to be convinced that the search engines can be trusted at least to reduce the volume of material to be scanned.

Buchanan and his colleagues made an interesting observation in their study of humanities researchers using Google's Web search engine. They found that "where naïve strategies were applied," researchers were often more satisfied with Google than specialist databases.[21] The authors attributed this to the way in which Google weighted retrieved Web pages, thus producing better results for "poorly focussed searches."[22]

Experience from the present study suggests that relevance ranking is particularly problematic with full-text book databases. In such databases, the documents are very large, much larger than the average journal article or Web page. How does one rank the relevance of a whole book to a specific search statement? Is the overall number of matches the significant factor, or is it more important to look for significant clusters of the search terms? In addition, should the whole book be regarded as a single document, or is it more productive to treat each chapter as a single document? These are issues

that require much closer study if the ranking systems of full-text book databases are to be improved.

Improving Retrieval via End-User Training

A survey of academic researchers conducted in 2002 in the United Kingdom found that 22 percent of arts and humanities respondents believed that they needed "a lot more training" in the use of electronic information sources.[23] Clearly, there is at least some recognition that searching databases is a skill that needs to be learned, although there are probably many researchers who have not yet realized this.

Librarians who help students and researchers to find information in databases have long realized that most database users are extremely unsophisticated in their use of the search options available. Most university libraries run courses to improve the searching skills of their patrons, but are these courses effective? And how advanced are the skills that they teach? Some librarians suggest that even Boolean operators are beyond the grasp of many users, and Marcia Bates came to the same conclusion after working with a group of senior humanities scholars.[24]

It seems clear that full-text databases present particular challenges for the researcher. Diane DiMartino and Lucinda Zoe surveyed a group of graduate students searching a full-text business database and found that 55 percent of those students were dissatisfied with their searches.[25] With so many large full-text collections becoming available, identifying search techniques that are simple to understand and effective in improving retrieval in full-text databases is clearly a priority now. What techniques can we recommend to researchers in the humanities?

Of course, we will advise them to choose search terms that are both uncommon and specific to the area that they are studying. For example, standardized abbreviations for scripture references, manuscripts, inscriptions, laws, and so on may prove to be productive search terms. The use of phrase searching is a simple but very effective technique. Because phrase searching uses natural language, searchers can draw upon their knowledge of the terminology of the discipline when choosing the phrases to search. These are thus informed search strategies that are likely to produce useful results. A recent study of the searching behaviors of a small sample of humanities researchers in New Zealand observed that "'discipline terms' (usually phrases representing a concept) were widely used. General keyword searches using such terms produced high numbers of results. If they were searched as bound phrases, retrieval improved, but less sophisticated users did not understand how to search them as bound phrases."[26]

Cited reference searching is a technique that many researchers have at least some familiarity, thanks to the long-established *Arts and Humanities Citation Index*, and this technique can be used when searching large collections of digitized secondary materials. The title of a known and relevant reference can be searched as a bound phrase (possibly in combination with the author's family name) to find later works that have cited that reference.

How useful are proximity operators? As E. Michael Keen has commented, proximity searching "is a very difficult method and difficult to get right: if pursued too narrowly Recall will suffer, and if too broadly no improvement in Precision will be experienced."[27]

How many researchers understand how to construct a search with proximity operators? And when constructing the search, how does the researcher decide the maximum number of words separating the search terms? Does the user have any clear concept of the significance of the terms appearing within five words of each other as opposed to 20 or 50 words?

We should view with caution any search facility that requires the searcher to assign numerical values to express desired proximity of search terms. Unless the user has conducted detailed linguistic analysis of previously identified documents, it will be impossible to assign the numerical value on any rational basis. The numbers chosen will be pure guesswork and the results unpredictable.

The same comments could be made with reference to boosting the weighting of individual search terms. The numerical value chosen will be largely arbitrary because the searcher can have no real concept of the difference between a boost factor of two and a boost factor of five. Even if searchers can be trained to construct searches using term weighting, it is questionable whether the results will be useful.

Conclusion

We are currently experiencing a "first phase" of large-scale digitization. Technology has advanced to a stage in which it is fairly inexpensive to digitize whole volumes and make them available for searching on the Web. This has opened up a dazzling prospect of large libraries available at our fingertips—fully searchable and fully viewable with only a few easy keystrokes. Journalists have spoken breathlessly of the "Herculean effort to digitize millions of books and make every sentence searchable" and of "providing researchers and students with an unprecedented tool for finding information."[28]

The excitement generated by this prospect has blinded many commentators to the distinct limitations of the digitization projects that are currently under way. The aim of the present paper has been to identify the difficulties associated with subject searching in digitized collections of secondary materials in the humanities. Although it is clear that scholars will welcome the easy availability of these materials on the Web, it seems only too likely that they will use these digital collections to look for documents that they have identified from other sources rather than performing subject searches to discover further resources.

It is mainly the large Web search companies and publishers who are driving the massive digitization projects currently in progress. As librarians, we have no input to this process, and our role can only be to observe, to evaluate, to develop techniques for using these resources as effectively as possible, and to teach those techniques to our clients. When the dust settles and euphoria turns to frustration, we may see a "second phase" of digitization, in which the "quick and dirty" digital libraries of today will be enriched and enhanced to become resources that can effectively meet the information needs of scholars.

John W. East is liaison librarian, Social Sciences and Humanities Library, University of Queensland, Brisbane, Australia; he may be contacted via e-mail at: john.east@uq.edu.au.

Notes

1. Carole L. Palmer, "Scholarly Work and the Shaping of Digital Access," *Journal of the American Society for Information Science and Technology* 56, 11 (2005): 1144.
2. Jeffrey Garrett, "KWIC and Dirty? Human Cognition and the Claims of Full-Text Searching," *Journal of Electronic Publishing* 9, 1 (2006), http://hdl.handle.net/2027/spo.3336451.0009.106 (accessed January 14, 2007); Alastair G. Smith, "Search Features of Digital Libraries," *Information Research* 5, 3 (April 2000), http://informationr.net/ir/5-3/paper73.html (accessed January 26, 2007).
3. Helen R. Tibbo, "Indexing for the Humanities," *Journal of the American Society for Information Science* 45, 8 (1994): 607.
4. David C. Blair, "STAIRS Redux: Thoughts on the STAIRS Evaluation, Ten Years After," *Journal of the American Society of Information Science* 47, 1 (1996): 4–22.
5. Peter Jackson and Isabelle Moulinier, *Natural Language Processing for Online Applications: Text Retrieval, Extraction and Categorization* (Amsterdam: John Benjamins, 2002), 211.
6. Tibbo, *Abstracting, Information Retrieval and the Humanities: Providing Access to Historical Literature*, ACRL Publications in Librarianship, no. 48 (Chicago: American Library Association, 1993), 5.
7. George Buchanan et al., "Information Seeking by Humanities Scholars," in *Research and Advanced Technology for Digital Libraries: 9th European Conference, Vienna, September 2005: Proceedings*, ed. Andreas Rauber, Stavros Christodoulakis, and A. Min Tjoa, Lecture Notes in Computer Science 3652 (Berlin: Springer, 2005), 227.
8. Debora Shaw, "Bibliographic Database Searching by Graduate Students in Language and Literature: Search Strategies, System Interfaces, and Relevance Judgments," *Library & Information Science Research* 17, 4 (1995): 340.
9. Ibid.
10. Blackwell Publishing, "Blackwell Synergy: Browsing and Searching," Blackwell Publishing, http://www.blackwell-synergy.com/page/browsing_searching (accessed January 29, 2007).
11. Tibbo, *Abstracting*, 95.
12. Sara D. Knapp, Laura B. Cohen, and Donald R. Juedes, "A Natural Language Thesaurus for the Humanities: The Need for a Database Search Aid," *Library Quarterly* 68, 4 (1998): 406–30.
13. Tibbo, "Indexing," 509.
14. Udo Hahn and Inderjeet Mani, "The Challenges of Automatic Summarization," *Computer* 33, 11 (2000): 29–36.
15. Ellen M. Voorhees and Donna K. Harman, "The Text REtrieval Conference," in *TREC: Experiment and Evaluation in Information Retrieval*, ed. Ellen M. Voorhees and Donna K. Harman (Cambridge, MA: MIT Press, 2005), 3–19.
16. Jackson and Moulinier.
17. Robert Krovetz, "Viewing Morphology as an Inference Process," *Artificial Intelligence* 118, 1/2 (2000): 277–94.
18. Stephen E. Wiberley, Jr., "Subject Access in the Humanities and the Precision of the Humanist's Vocabulary," *Library Quarterly* 53, 4 (1983): 430.
19. Shaw, 332.
20. Ibid.
21. Buchanan et al., 227.
22. Ibid.
23. Education for Change Ltd., SIRU (University of Brighton), and The Research Partnership, "Researchers' Use of Libraries and Other Information Sources: Current Patterns and Future Trends: Final Report," http://www.rslg.ac.uk/research/libuse/LUrep1.pdf (accessed January 26, 2007), 75.

24. Marcia J. Bates, "The Getty End-User Online Searching Project in the Humanities: Report No. 6, Overview and Conclusions," *College & Research Libraries* 57, 6 (1996): 520.
25. Diane DiMartino and Lucinda R. Zoe, "End-User Full-Text Searching: Access or Excess?" *Library and Information Science Research* 18, 2 (1996): 133–49.
26. Buchanan et al., 226.
27. E. Michael Keen, "Some Aspects of Proximity Searching in Text Retrieval Systems," *Journal of Information Science* 18, 2 (1992): 90.
28. Scott Carlson and Jeffrey R. Young, "Google Will Digitize and Search Millions of Books from 5 Leading Research Libraries," *Chronicle of Higher Education*, January 7, 2005, A37.