Browsing around a digital library: Today and tomorrow

Ian H. Witten

Department of Computer Science, University of Waikato,
Hamilton, New Zealand
ihw@cs.waikato.ac.nz

Abstract. What will it be like to work in tomorrow's digital library? We begin by browsing around an experimental digital library of the present, glancing at some collections and showing how they are organized. Then we look to the future. Although present digital libraries are quite like conventional libraries, we argue that future ones will feel qualitatively different. Readers—and writers—will work in the library using a kind of context-directed browsing. This will be supported by structures derived from automatic analysis of the contents of the library—not just the catalog, or abstracts, but the full text of the books and journals—using new techniques of text mining.

1 Introduction

Over sixty years ago, science fiction writer H.G. Wells was promoting the concept of a "world brain" based on a permanent world encyclopedia which "would be the mental background of every intelligent [person] in the world. It would be alive and growing and changing continually under revision, extension and replacement from the original thinkers in the world everywhere. ... even journalists would deign to use it" [14]. Eight years later, Vannevar Bush, the highest-ranking scientific administrator in the U.S. war effort, invited us to "consider a future device for individual use, which is a sort of mechanized private file and library ... a device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility" [2]. Fifteen years later, J.C.R. Licklider, head of the U.S. Department of Defense's Information Processing Techniques Office, envisioned that human brains and computing machines would be coupled together very tightly, and imagined this to be supported a "network of 'thinking centers' that will incorporate the functions of present-day libraries together with anticipated advances in information storage and retrieval" [8]. Thirty-five years later we became accustomed to hearing similar pronouncements from the U.S. Presidential office.

Digital libraries, conceived by visionary thinkers and fertilized with resources by today's politicians, are undergoing a protracted labor and birth. Libraries are society's repositories for knowledge, and digital libraries are of the utmost strategic importance in a knowledge-based economy. Not surprisingly, many countries

have initiated large-scale digital library projects. Some years ago the DLI initiative was set up in the U.S. (and has now entered a second phase); in the U.K. the Elib program was set up at about the same time; other countries in Europe and the Pacific Rim have followed suit. Digital libraries will likely figure amongst the most important and influential institutions of the 21st Century.

But what is a digital library? Ten definitions of the term have been culled from the literature by Fox [4], and their spirit is captured in the following brief characterization [1]:

> *a focused collection of digital objects, including text, video, and audio, along with methods for access and retrieval, and for selection, organization, and maintenance of the collection.*

This definition gives equal weight to user (access and retrieval) and librarian (selection, organization and maintenance). Other definitions in the literature, emanating mostly from technologists, omit—or at best downplay—the librarian's role, which is unfortunate because it is the selection, organization, and maintenance that will distinguish digital libraries from the anarchic mess that we call the World Wide Web. However, digital libraries tend to blur what used to be a sharp distinction between user and librarian—because the ease of augmenting, editing, annotating and re-organizing electronic collections means that they will support the development of new knowledge *in situ*.

What's it like to work in a digital library? Will it feel like a conventional library, but more computerized, more networked, more international, more all-encompassing, more convenient? I believe the answer is no: it will feel qualitatively different. Not only will it be with you on your desktop (or at the beach, or in the plane), but information workers will work "inside" the library in a way that is quite unlike how they operate at present. It's not just that knowledge and reference services will be fully portable, operating round the world, around the clock, throughout the year, freeing library patrons from geographic and temporal constraints—important and liberating as these are. It's that when new knowledge is created it will be fully contextualized and both sited within and cited by existing literature right from its conception.

In this paper, we browse around a digital library, looking at tools and techniques under development. "Browse" is used in a dual sense. First we begin by browsing a particular collection, and then look briefly at some others. Second, we examine the digital library's ability to support novel browsing techniques. These situate browsing within the reader's current context and unobtrusively guide them in ways that are relevant to what they are doing. Context-directed browsing is supported by structures derived from automatic analysis of the library's contents—not just the catalog, or abstracts, but the *full text* of the documents—using techniques that are being called "text mining." Of course, other ways of finding information are important too—user searching, librarian recommendations, automatic notification, group collaboration—but here we focus on browsing. The work described was undertaken by members of the New Zealand Digital Library project.

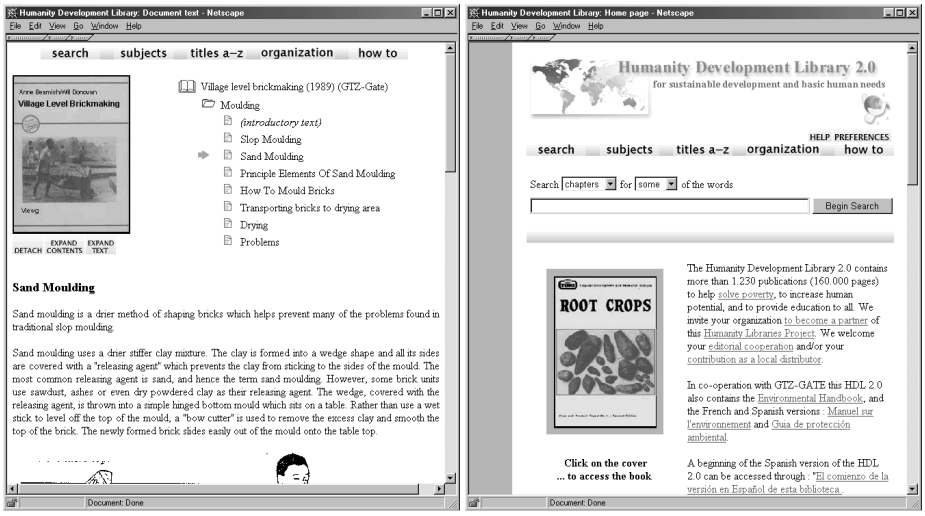

Fig. 1. (a) *Village Level Brickmaking*, (b) The collection's home page

2 The Humanity Development Library

Figure 1a shows a book in the *Humanity Development Library*, a collection of humanitarian information put together by the Global Help Project to address the needs of workers in developing countries (`www.nzdl.org/hdl`). This book might have been reached by a directed full-text search, or by browsing one of a number of access structures, or by clicking on one of a gallery of images. On opening the book, which is entitled *Village Level Brickmaking*, a picture of its cover appears at the top, beside a hierarchical table of contents. In the figure, the reader has drilled down into a chapter on *moulding* and a subsection on *sand moulding*, whose text appears below. Readers can expand the table of contents from the section to the whole book; and expand the text likewise (which is very useful for printing). The ever-present picture of the book's cover gives a feeling of physical presence and a constant reminder of the context.

Readers can browse the collection in several different ways, as determined by the editor who created it. Figure 1b shows the collection's home page, at the top of which (underneath the logo) is a bar of five buttons that open up different access mechanisms. A subject hierarchy provides a tree-structured classification scheme for the books. Book titles appear in an alphabetical index. A separate list gives participating organizations and the material that they contributed. A "how-to" list of helpful hints, created by the collection's editor, allows a particular book to be accessed from brief phrases that describe the problems the book addresses. However a book is reached, it appears in the standard form illustrated in Figure 1a, along with the cover picture to give a sense of presence. The different access mechanisms help solve the librarian's dilemma of where to

shelve books [9]: each one appears on many different virtual shelves, shelves that are organized in different ways.

Full-text search of titles and entire documents provide important additional access mechanisms. The search engine that we use, MG [15], supports searching over the full text of the document—not merely a document surrogate as in conventional digital library retrieval systems. User feedback from an earlier version of this collection indicated that Boolean searching was more confusing than helpful for the targeted users. Previous research suggests that difficulties with Boolean syntax and semantics are widespread, and transaction log analysis of several library retrieval systems indicates that by far the most popular Boolean operator is AND; the others are rarely used. For all these reasons, the interface default for this collection is ranked queries. However, to enable users to construct high-precision conjunctive searches where necessary, selecting "search ... for *all* the words" in the query dialog produces the syntax-free equivalent of a conjunctive query.

Just as libraries display new acquisitions or special collections in the foyer to pique the reader's interest, this collection's home page (Figure 1b) highlights a particular book that changes every few seconds: it can be opened by clicking on the image. This simple display is extraordinarily compelling. And just as libraries may display a special book in a glass case, open at a different page each day, a "gallery" screen can show an ever-changing mosaic of images from pages of the books, remarkably informative images that, when clicked, open the book to that page. Or a scrolling "Times Square" display of randomly selected phrases that, when clicked, take you to the appropriate book. The possibilities are endless.

The *Humanity Development Library* is a focused collection of 1250 books—miniscule by library standards, but nevertheless comprehensive within the targeted domain. It contains 53,000 chapters, 62 million words, and 32,000 pictures. Although the text occupies 390 MB, it compresses to 102 MB and the two indexes—for titles and chapters respectively—compress to less than 80 MB. The images (mostly in PNG format) occupy 290 MB. Associated files bring the total size of the collection to 505 MB. Even if there were twice as much text, and the same images, it would still fit comfortably on a CD-ROM, along with all the necessary software. A single DVD-ROM would hold a collection twenty times the size—still small by library standards, but immense for a fully portable collection.

3 An experimental testbed: The NZDL

The *Humanity Development Library* is just one of two dozen publicly-available collections produced by the New Zealand Digital Library (NZDL) project and listed on the project's home page (`www.nzdl.org`), part of which is shown in Figure 2a—this illustrates the the wide range of collections. This project aims to develop the underlying infrastructure for digital libraries and provide example collections that demonstrate how it can be used. The library is international,

Fig. 2. (a) Some of the NZDL collections, (b) Reading a Chinese book

and the Unicode character set is used throughout: there are interfaces in English, Maori, French, German, Arabic, and Chinese, and collections have been produced in all these languages. Digital libraries are particularly empowering for the disabled, and there is a text-only version of the interface intended for visually impaired users.

The editors of the *Humanity Development Library* have gone to great lengths to provide a rich set of access structures. However, this is a demanding, labor-intensive task, and most collections are not so well organized. The basic access tool in the NZDL is full-text searching, which is available for all collections and is provided completely automatically when a collection is built. Some collections allow, in addition, traditional catalog searching based on author, title, and keywords, and full-text search within abstracts. Our experience is that while the user interface is considerably enhanced when traditional library cataloging information is available, it is often prohibitively expensive to create formal cataloging information for electronically-gathered collections. With appropriate indexes, full-text retrieval can be used to approximate the services provided by a formal catalog.

3.1 Collections

The core of any library is the collections it contains. A few examples will illustrate the variety and scope of the services provided.

The historically first collection, *Computer Science Technical Reports*, now contains 46,000 reports—1.3 million pages, half a billion words—extracted automatically from 34 GB of raw PostScript. There is no bibliographic or "metadata" information: we have only the contents of the reports (and the names of the FTP

sites from which they were gathered). Many are Ph.D. theses which would otherwise be effectively lost except to a miniscule community of cognoscenti: full-text search reaches right inside the documents and makes them accessible to anyone looking for information on that topic.

As well as the simplified searching interface for the *Humanity Development Library* described above, users can choose a more comprehensive query interface (via a *Preferences* page). Case-folding and stemming can be independently enabled or disabled, and full Boolean query syntax is supported as well as ranked queries. Moreover, in the *Computer Science Technical Reports* searches can be restricted to the first page of reports, which approximates an author/title search in the absence of specific bibliographic details of the documents. Although this is a practical solution, the collection nevertheless presents a raw, unpolished appearance compared with the *Humanity Development Library*, reflecting the difference between a carefully-edited set of documents, including hand-prepared classification indexes and other metadata, and a collection of information pulled mechanically off the Web and organized without any human intervention at all.

There are several collections of books, including, for example, the English books entered by the Gutenberg project. Figure 2b shows a book in a collection of classical Chinese literature. The full text was available on the Web; we automatically extracted the section headings to provide the table of contents visible at the upper right, and scanned the book's cover to generate the cover image. One can perform full-text search on the complete contents or on section headings alone, using the Chinese language (of course your Web browser must be set up correctly to work in Chinese). There is also a browsable list of book titles.

An expressly bilingual collection of *Historic New Zealand Newspapers* contains issues of forty newspapers published between 1842 and 1933 for a Maori audience. Collected on microfiche, these constitute 12,000 page images. Although they represent a significant resource for historians, linguists and social scientists, their riches remain largely untapped because of the difficulty of accessing, searching and browsing material in unindexed microfiche form. Figure 3 shows the parallel English-Maori text retrieved from the newspaper *Te Waka Maori* of August 1878 in response to the query *Rotorua*, a small town in New Zealand. Searching is carried out on electronic text produced using OCR; once the target is identified, the corresponding page image can be displayed.

3.2 The Greenstone software

All these collections are created using the Greenstone software developed by the NZDL project [17]. Information collections built by Greenstone combine extensive full-text search facilities with browsing indexes based on different metadata types. There are several ways for users to find information, although they differ between collections depending on the metadata available and the collection design. You can *search for particular words* that appear in the text, or within a section of a document, or within a title or section heading. You can *browse documents by title*: just click on the displayed book icon to read it. You can *browse documents by subject*. Subjects are represented by bookshelves: just click

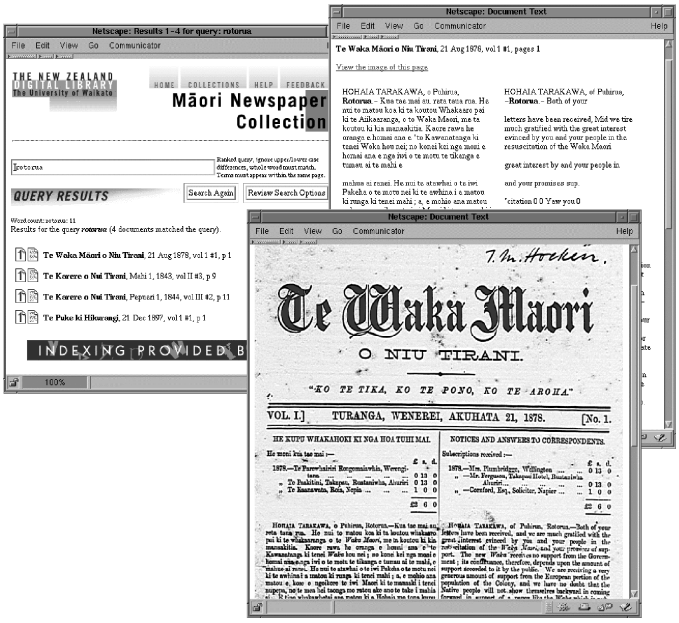

Fig. 3. Searching the *Historic New Zealand newspapers* collection

on a shelf to see the books. Where appropriate, documents come complete with a table of contents (constructed automatically): you can click on a chapter or subsection to open it, expand the full table of contents, or expand the full document.

A distinction is made between *searching* and *browsing*. Searching is full-text, and—depending on the collection's design—the user can choose between indexes built from different parts of the documents, or from different metadata. Some collections have an index of full documents, an index of sections, an index of paragraphs, an index of titles, and an index of section headings, each of which can be searched for particular words or phrases. Browsing involves data structures created from metadata that the user can examine: lists of authors, lists of titles, lists of dates, hierarchical classification structures, and so on. Data structures for both browsing and searching are built according to instructions in a configuration file, which controls both building and serving the collection.

Rich browsing facilities can be provided by manually linking parts of documents together and building explicit indexes and tables of contents. However, manually-created linking becomes difficult to maintain, and often falls into disrepair when a collection expands. The Greenstone software takes a different tack: it facilitates *maintainability* by creating all searching and browsing structures automatically from the documents themselves. No links are inserted by hand. This means that when new documents in the same format become available, they can be added automatically. Indeed, for some collections this is done by processes

that wake up regularly, scout for new material, and rebuild the indexes—all without manual intervention.

Collections comprise many documents: thousands, tens of thousands, or even millions. Each document may be hierarchically organized into *sections* (subsections, sub-subsections, and so on). Each section comprises one or more *paragraphs*. Metadata such as author, title, date, keywords, and so on, may be associated with documents, or with individual sections of documents. This is the raw material for indexes. It must either be provided explicitly for each document and section (for example, in an accompanying spreadsheet) or be derivable automatically from the source documents. Metadata is converted to Dublin Core and stored with the document for internal use.

In order to accommodate different kinds of source documents, the software is organized so that "plugins" can be written for new document types. Plugins exist for plain text documents, HTML documents, email documents, and bibliographic formats. Word documents are handled by saving them as HTML; PostScript ones by applying a preprocessor [12]. Specially written plugins also exist for proprietary formats such as that used by the BBC archives department. A collection may have source documents in different forms: it is just a matter of specifying all the necessary plugins. In order to build browsing indexes from metadata, an analogous scheme of "classifiers" is used: classifiers create indexes of various kinds based on metadata. Source documents are brought into the Greenstone system through a process called *importing*, which uses the plugins and classifiers specified in the collection configuration file.

The system includes an "administrative" function whereby specified users can examine the composition of all collections, protect documents so that they can only be accessed by registered users on presentation of a password, and so on. Logs of user activity are kept that record all queries made to every Greenstone collection (though this facility can be disabled).

Although primarily designed for Internet access over the World-Wide Web, collections can be made available, in precisely the same form, on CD-ROM. In either case they are accessed through any Web browser. Greenstone CD-ROMs operate on a standalone PC under Windows 3.X, 95, 98, and NT, and the interaction is identical to accessing the collection on the Web—except that response is faster and more predictable. The requirement to operate on early Windows systems is a significant practical impediment to the software design, but is crucial for many users—particularly those in underdeveloped countries seeking access to humanitarian aid collections. If the PC is connected to a network (intranet or Internet), a custom-built Web server provided on each CD makes exactly the same information available to others through their standard Web browser. The use of compression ensures that the greatest possible volume of information can be packed on to a CD-ROM.

The collection-serving software operates under Unix and Windows NT, and works with standard Web servers. A flexible process structure allows different collections to be served by different computers, yet be presented to the user in the same way, on the same Web page, as part of the same digital library [10].

Existing collections can be updated and new ones brought on-line at any time, without bringing the system down; the process responsible for the user interface will notice (through periodic polling) when new collections appear and add them to the list presented to the user.

4 Browsing in the digital library of the future

Current digital library systems often contain handcrafted indexes and links to provide different entry points into the information, and to bind it together into a coherent whole. This can produce high-quality, focused collections—but it is basically unscalable. Excellent new material will, of course, continue to be produced using manual techniques, but it is infeasible to suppose that the mass of existing, archival material will be manually "converted" into high-quality digital collections. The only scalable solution that is used currently for amorphous information collections is the ubiquitous search engine—but browsing is poorly supported by standard search engines. They operate at the wrong level, indexing words whereas people think in terms of topics, and returning individual documents whereas people often seek a more global view.

Suppose you are browsing a large collection of information such as a digital library—or a large Web site. Searching is easy, if you know what you are looking for—and can express it as a query at the lexical level. But current search mechanisms are not much use if you are not looking for a specific piece of information, but are generally exploring the collection. Studies of browsing have shown that it is a rich and fundamental human information behavior, a multifaceted and multidimensional human activity [3]. But it is not well-supported for large digital collections.

We look at three browsing interfaces that capitalize on automatically-generated phrases and keyphrases for a document collection. The first of these is a phrase-based browser that is specifically designed to support subject-index-style browsing of large information collections. The second uses more specific phrases and concentrates on making it convenient to browse closely-related documents. The third is a workbench that facilitates skimming, reading, and writing documents within a digital library—a qualitatively different experience from working in a library today. All three are based on phrases and keyphrases extracted automatically from the document text itself.

4.1 Emulating subject indexes: phrase browsing

Phrases extracted automatically from a large information collection form an excellent basis for browsing and accessing it. We have developed a phrase-based browser that acts an interactive interface to a phrase hierarchy that has been extracted automatically from the full text of a document collection. It is designed to resemble a paper-based subject index or thesaurus.

We illustrate the application of this scheme to a large Web site: that of the United Nations *Food and Agriculture Organization* (`www.fao.org`), an international organization whose mandate is to raise levels of nutrition and standards

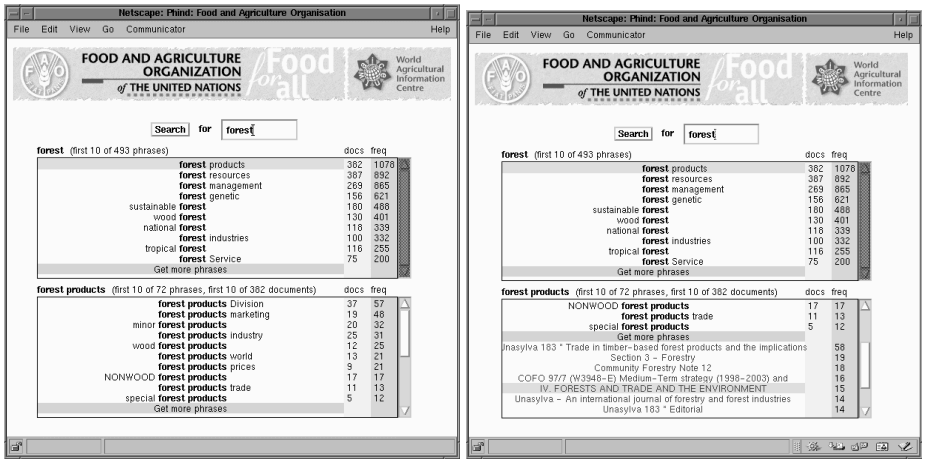

Fig. 4. (a) Browsing for information about *forest*, (b) Expanding on *forest products*

of living, to improve agricultural productivity, and to better the condition of rural populations. The site contains 21,700 Web pages, as well as around 13,700 associated files (image files, PDFs, etc). This corresponds to a medium-sized collection of approximately 140 million words of text. It exhibits many problems common to large, public Web sites. It has existed for some time, is large and continues to grow rapidly. Despite strenuous efforts to organize it, it is becoming increasingly hard to find information. A search mechanism is in place, but while this allows some specific questions to be answered it does not really address the needs of the user who wishes to browse in a less directed manner.

Figure 4a shows the phrase browsing interface in use. The user enters an initial word in the search box at the top. On pressing the *Search* button the upper panel appears. This shows the phrases at the top level in the hierarchy that contain the search word-in this case the word *forest*. The list is sorted by phrase frequency; on the right is the number of times the phrase appears, and to the left of that is the number of documents in which it appears.

Only the first ten phrases are shown, because it is impractical with a Web interface to download a large number of phrases, and many of these phrase lists are very large. At the end of the list is an item that reads *Get more phrases* (displayed in a distinctive color); clicking this will download another ten phrases, and so on. A scroll bar appears to the right for use when more than ten phrases are displayed. The number of phrases appears above the list: in this case there are 493 top-level phrases that contain the term *forest*.

So far we have only described the upper of the two panels in Figure 4a. The lower one appears as soon as the user clicks one of the phrases in the upper list. In this case the user has clicked *forest products* (that is why that line is highlighted in the upper panel) and the lower panel, which shows phrases containing the text *forest products*, has appeared.

If one continues to descend through the phrase hierarchy, eventually the leaves will be reached. A leaf corresponds to a phrase that occurs in only one document of the collection (though the phrase may appear several times in that document). In this case, the text above the lower panel shows that the phrase *forest products* appears in 72 phrases (the first ten are shown), and, in addition, appears in a unique context in 382 documents. The first ten of these are available too, though the list must be scrolled down to make them appear in the visible part of the panel. Figure 4b shows this. In effect, the panel shows a phrase list followed by a document list. Either of these lists may be null (in fact the document list is null in the upper panel, because the word *forest* appears only in other phrases or in individual unique contexts). The document list displays the titles of the documents.

It is possible, in both panels of Figures 4a and b, to click *Get more phrases* to increase the number of phrases that are shown in the list of phrases. It is also possible, in the lower panels, to click *Get more documents* (again it is displayed at the end of the list in a distinctive color, but to see that entry it is necessary to scroll the panel down a little more) which increases the number of documents that are shown in the list of documents.

Clicking on a phrase will expand it. The page holds only two panels, and if a phrase in the lower panel is clicked the contents of that panel move up into the top one to make space for the phrase's expansion. Alternatively, clicking on a document will open that document in a new window. In fact, the user in Figure 4b has clicked on *IV FORESTS AND TRADE AND THE ENVIRONMENT*, and this brings up a Web page with that title. As Figure 4b indicates, that page contains 15 occurrences of the phrase *forest products*.

We have experimented with several different ways of creating a phrase hierarchy from a document collection. An algorithm called SEQUITUR builds a hierarchical structure containing every single phrase that occurs more than once in the document collection [11]. We have also worked on a scheme called KEA which extracts keyphrases from scientific papers. This produces a far smaller, controllable, number of phrases per document [5]. The scheme that we use for the interface in Figure 4 is an amalgam of the two techniques [13].

The phrases extracted represent the topics present in the *Food and Agriculture Organization* site, as described in the terminology of the document authors. We have investigated how well this set of phrases matches the standard terminology of the discipline by comparing the extracted phrases with phrases used by the AGROVOC agricultural thesaurus. There is a substantial degree of overlap between the two sets of phrases, which provides some confirmation of the quality of the extracted phrases as subject descriptors.

4.2 Improved browsing using keyphrase indexes

Another a new kind of search interface that is explicitly designed to support browsing is based on keyphrases automatically extracted from the documents [6] using the KEA system [5]. A far smaller number of phrases are selected than in

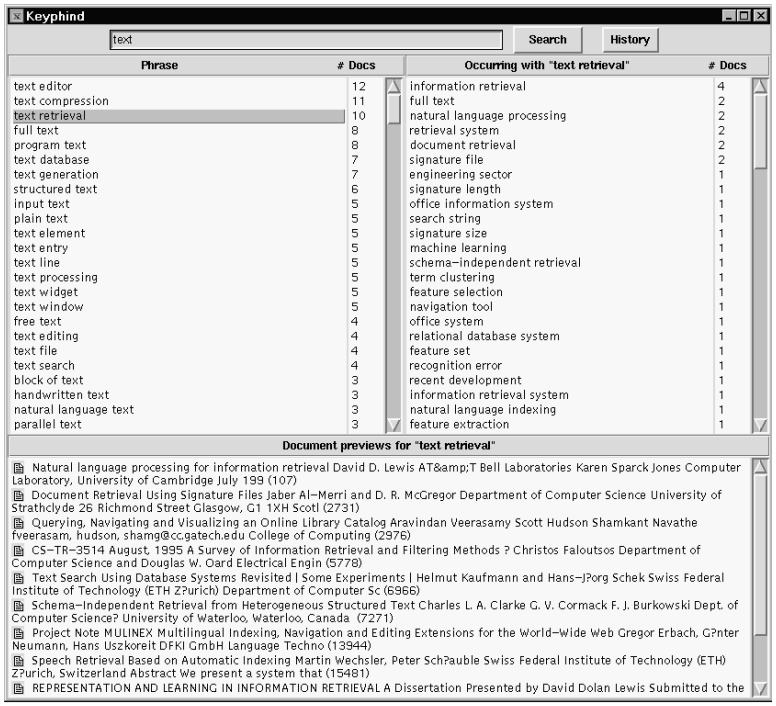

Fig. 5. Browsing a keyphrase index to find out about topics involving *text*

the system described above—only four or five per document. The automatically-extracted keyphrases form the basic unit of both indexing and presentation, allowing users to interact with the collection at the level of topics and subjects rather than words and documents. The system displays the topics in the collection, indicates coverage in each area, and shows all ways a query can be extended and still match documents.

The interface is shown in Figure 5. A user initiates a query by typing words or phrases and pressing the *Search* button, just as with other search engines. However, what is returned is not a list of documents, but a list of keyphrases containing the query terms. Since all phrases in the database are extracted from the source documents, every returned phrase represents one or more documents in the collection. Searching on the word *text*, for example, returns a list of phrases including *text editor* (a keyphrase for twelve documents), *text compression* (eleven documents), and *text retrieval* (ten documents), as shown in Figure 5. The phrase list provides a high-level view of the topics represented in the collection, and indicates, by the number of documents, the coverage of each topic.

Following the initial query, a user may choose to refine the search using one of the phrases in the list, or examine a topic more closely. Since they are derived from the collection itself, any further search with these phrases is guaranteed to

produce results—and furthermore, the user knows exactly how many documents to expect. To examine the documents associated with a phrase, the user selects it from the list, and previews of documents for which it is a keyphrase are displayed in the lower panel of the interface. Selecting any preview shows the document's full text.

Experiments with users show that this interface is superior to a traditional search system for answering particular kinds of questions: evaluating collections ("what's in this collection"), exploring areas ("what subtopics are available in area X"), and general information about queries ("what kind of queries will succeed in area X", "how can I specialize or generalize my query"). Note that many of these questions are as relevant to librarians as they are to library users. However, this mechanism is not intended to replace conventional search systems for specific queries about specific documents.

4.3 Reading and writing in a digital library

A third prototype system, developed by Jones [7], shows how phrases can assist with skimming, reading, and writing documents in the digital library. It uses the keyphrases extracted from a document collection as link anchors to point to other documents. When reading a document, the keyphrases in it are highlighted. When writing one, phrases are dynamically linked, and highlighted, as you type.

Figure 6 shows the interface. To the left is the document being examined (read or authored); in the center is the keyphrase pane; and to the right is the library access pane. Keyphrases that appear in documents in the collection are highlighted; this facilitates rapid skimming of the content because the darker text points out items that users often highlight manually with a marker pen. Different gray levels reflect the "relevance" of the keyphrase to the document, and the user can control the intensity to match how they skim. Each phrase is hyperlinked, using multiple-destination links, to other documents for which it is a keyphrase (the anchor is the small spot that follows the phrase). The center panel shows all the keyphrases that appear in this document, with their frequency and the number of documents in the library for which they are keyphrases. Controls are available to sort the list in various different ways. Some of these phrases have been selected by the user, and on the right is a ranked list of items in the library that contain them as keyphrases—ranked according to a special metric designed for use with keyphrases.

With this interface, hurried readers can skim a document by looking at the highlighted phrases. In-depth readers can instantly access other relevant documents (including dictionaries or encyclopaedias). They can select a subset of relevant phrases and instantly have the library searched on that set. Writers—as they type—can immediately gain access to documents that are relevant to what they are writing.

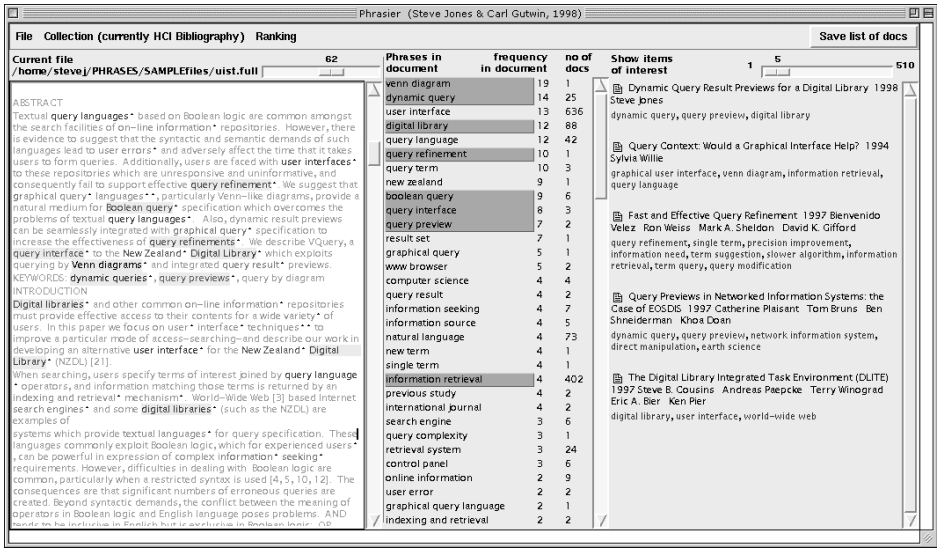

Fig. 6. Working on a paper inside the digital library

5 Conclusion

Digital libraries have finally arrived. They are different from the World Wide Web: libraries are focused collections, and it is the act of selection that gives them focus. For many practical reasons (including copyright, and the physical difficulty of digitization), digital libraries will not vie with archival national collections, not in the foreseeable future. Their role is in specialist, targeted collections of information.

Established libraries of printed material have sophisticated and well-developed human and computer-based interfaces to support their use. But they are not well integrated for working with computer tools: a bridging process is required. Information workers can immerse themselves physically in the library, but they cannot take with them their tasks, tools, and desktop workspaces. The digital library will be different: we will work "inside" it in a sense that it totally new.

But even for a focused collection, creating a high-quality digital library is a highly labor-intensive process. To provide the richness of access and interconnection that makes a digital library comfortable requires enormous editorial effort. And when the collection changes, maintenance becomes an overriding issue. Fortunately, techniques of text mining are emerging that offer the possibility of automatic identification of semantic items from plain text. Carefully-constructed user interfaces can take advantage of the information that they generate to provide a library experience that is qualitatively different from a physical library—not just in access and convenience, but in terms of the quality of browsing and information accessibility. Tomorrow, digital libraries will put the right information at your fingertips.

Acknowledgments

Many thanks to members of the New Zealand Digital Library project, whose work is described in this paper, particularly David Bainbridge, George Buchanan, Stefan Boddie, Rodger McNab and Bill Rogers who built the Greenstone system; Eibe Frank and Craig Nevill-Manning who worked on phrase extraction; Carl Gutwin, Steve Jones and Gordon Paynter who created phrase-based interfaces; and Mark Apperley, Sally Jo Cunningham, Te Taka Keegan and Malika Mahoui for their help and support. Rob Akscyn, Michel Loots and Harold Thimbleby also made valued contributions.

References

1. Akscyn, R.M. and Witten, I.H. (1998) "Report on First Summit on International Cooperation on Digital Libraries." `ks.com/idla-wp-oct98`.
2. Bush, V. (1947) "As we may think." *The Atlantic Monthly*, Vol. 176, No. 1, pp. 101–108.
3. Chang, S.J. and Rice, R.E. (1993) "Browsing: a multidimensional framework." *Annual Review of Information Science and Technology*, Vol. 28, pp. 231–276.
4. Fox, E. (1998) "Digital library definitions." `ei.cs.vt.edu/ fox/dlib/def.html`.
5. Frank, E., Paynter, G.W., Witten, I.H., Gutwin, C. and Nevill-Manning, C. (1999) "Domain-specific keyphrase extraction." *Proc Int Joint Conference on Artificial Intelligence*, Stockholm, Sweden, pp. 668–673.
6. Gutwin, C., Paynter, G., Witten, I.H., Nevill-Manning, C., and Frank, E. (in press) "Improving browsing in digital libraries with keyphrase indexing." *Decision Support Systems.*
7. Jones, S. (1999) "Phrasier: an interactive system for linking and browsing within document collections using keyphrases." *Proc Interact 99: Seventh IFIP Conference On Human-Computer Interaction*, Edinburgh, Scotland, pp. 483–490.
8. Licklider, J.C.R. (1960) "Man-computer symbiosis." *IRE Trans HFE-1*, pp. 4–11.
9. Mann, T. (1993) *Library research models.* Oxford University Press, NY.
10. McNab, R.J., Witten, I.H. and Boddie, S.J. (1998) "A distributed digital library architecture incorporating different index styles." *Proc IEEE Advances in Digital Libraries*, Santa Barbara, CA, pp. 36–45.
11. Nevill-Manning, C.G., Witten, I.H. and Paynter, G.W. (1997) "Browsing in digital libraries." *Proc ACM Digital Libraries 97*, pp. 230–236.
12. Nevill-Manning, C.G., Reed, T., and Witten, I.H. (1998) "Extracting text from PostScript." *Software—Practice and Experience*, Vol. 28, No. 5, pp. 481–491.
13. Paynter, G.W., Witten, I.H., Cunningham, S.J. and Buchanan, G. (2000) "Scalable browsing for large collections: a case study." *Proc Digital Libraries 2000*, Austin, TX.
14. Wells, H.G. (1938) *World Brain.* Doubleday, NY.
15. Witten, I.H., Moffat, A., and Bell, T.C. (1999) *Managing Gigabytes: Compressing and indexing documents and images.* Morgan Kaufmann, San Francisco.
16. Witten, I.H., Bray, Z., Mahoui, M. and Teahan, W. (1999) "Text mining: a new frontier for lossless compression." *Proc Data Compression Conference*, pp. 198–207.
17. Witten, I.H., McNab, R.J., Boddie, S.J. and Bainbridge, D. (2000) "Greenstone: a comprehensive open-source digital library software system." *Proc ACM Digital Libraries*, San Antonio, TX.