Distributing Digital Libraries on the Web, CD-ROMs, and Intranets: Same information, same look-and-feel, different media

Ian Witten, Sally Jo Cunningham, Bill Rogers, Rodger McNab,Stefan Boddie
Department of Computer Science
University of Waikato
Hamilton, New Zealand

Abstract: The Greenstone system from the New Zealand Digital Library provides a new way of making collections of information available in the same form over the World-Wide Web, on CD-ROM, or on local Intranets. Exactly the same information is available in each case, and exactly the same interface is used to access it. The New Zealand Digital Library is accessible over the Web and offers a wide variety of information collections. Sub-collections can be written to a CD-ROM, which can be used on a standalone PC by a single user. A local Web browser suffices to access the information on the disk just as though the PC were connected to the Internet. Simultaneously, if there is a network connection, the same disk acts as a network server to make exactly the same information available to others who need only use their standard Internet browser software. This technology has great appeal for many users, particularly those in developing nations where non-local Internet access can be precarious or prohibitively expensive.

1. Introduction

The emerging digital library movement is a child of the Internet and the World-Wide Web. Spurred on by visions of an "information superhighway," current digital library projects invariably concentrate on providing access to document collections over the Internet, where documents, users, and catalog may all be distributed widely. Often the search interface is WWW-based, in contrast to the telnet or phone-in access required by library OPACS and earlier commercial "online" bibliographic databases such as Dialog. Web-based digital libraries have significant advantages over their online predecessors. Users need not obtain and install search software on their own sites. In many areas Internet access incurs minimal charges, or at any rate is significantly cheaper than a direct telephone connection with the retrieval system. Finally, Web browsers provide a simple, standard means of access to a variety of digital library systems.

However, practical experience in digital library development indicates that in many situations, universal access via the Internet is neither possible nor desirable. A business, for example, might desire a digital library to make its proprietary documents available to its employees, but only if the company's security could be ensured by restricting access with an intranet. CD-ROM has been identified as the implementation platform of choice for collections targeted at large portions of the Third World; for many developing countries, particularly in Sub-Saharan Africa, Internet connections are still either non-existent, undependable, or prohibitively expensive to use. Despite its lowly status, the CD-ROM has many advantages. Relatively durable in the face of harsh environmental conditions, it incurs known, fixed costs for purchase and supporting hardware (White, 1992). It makes information accessible on a tangible medium that is under the user's control and is not subject to capricious decisions by others. A CD-ROM based digital library carries the further advantage of providing full document contents—a significant drawback to bibliographic systems being that their users in developing countries could locate descriptions of relevant documents, but were then often unable to obtain the documents themselves (El-Hadidy, 1994; Chowdhury, 1996). Finally, while a CD-ROM holds a reasonable amount of material in textual form, digital videodisk technology is already available which can store 12 Gb on a single disk—far larger than most extant textual digital libraries.

For this reason the Greenstone digital library software developed by the New Zealand Digital Library project allows a collection developer to create a digital library that is WWW-based, intranet-based, or available on a standalone or networked CD-ROM. All platforms support exactly the same interface, and the same search and retrieval methods. This standardization reduces the system learning curve for intranet or CD-ROM users who have previous experience with WWW browsers, and conversely allows those users currently without Internet access to more easily progress to Web searching and browsing when it becomes available to them.

An earlier version of this software has been used in a university-level distance learning course on computer literacy, where selected portions of various WWW sites were stored on CD-ROM for students to surf (Holmes and Rogers, 1997). Here, the primary advantages of avoiding an Internet connection were to smooth out variable page retrieval times, to avoid problems with off-site servers going down or being temporarily unavailable, and to eliminate communication costs. In secondary or primary school settings, this technique for capturing known portions of the WWW can be used to prevent students wasting lab time exploring sites that irrelevant to the task at hand, or that are inappropriate for their age groups.

The digital library collection described in this paper is comprised of a set of documents provided by the United Nations University, focusing primarily on food and nutrition. The goal of the United Nations University Press is to disseminate knowledge in the field of the global problems of human survival, development and welfare, in order to increase dynamic interaction in the world-wide community of learning and research. By making their documents available in a variety of formats—print, CD-ROM, WWW pages—this research and human development information can be distributed more widely, and in a form appropriate to the conditions required by information users.

Section 2 describes the software architecture. Multimedia collections are supported, and a single collection may include text, images, audio, and even video clips. Compression technology is used to ensure that the greatest possible volume of information is packed into a limited storage space. The interface software combines easy-to-use browsing with powerful search facilities. As discussed in Section 3, several ways are provided to find information in a collection; a user can conduct keyword searches, access known documents by title, or browse subject “bookshelves”.

2. System architecture

A great advantage of the WWW as a means of presenting and using information is that very little direct user interface programming is required. A system can generate simple text documents in HTML notation, and leave the task of display, printing, screen navigation, and so forth to a Web browser. As a result, the browser writer takes most of the burden of system dependence away from the application programmer. The CD version of the Greenstone library follows this structure: our software takes the form of a WWW server, communicating with an unmodified browser using IP networking software. While the primary goal is to have a system running on a stand-alone machine, the use of IP networking does also mean that the software will function as a WWW server over an external network. Figure 1 shows the general software organization. The gray box encloses the software components running on one machine.

Ideally, the WWW server would be a standard piece of software, and a digital library would take exactly the same form on a single machine as it does on our larger WWW serving equipment. This did not prove possible for a number of reasons—most significant of which was the amount of memory expected to be available on our target machines, which for this project include the older and smaller workstations commonly

in use in the Third World. The full digital library system on our WWW servers does make use of standard Internet server software. In the WWW version of our digital library architecture, pre and post processing of queries on the library are handled in tasks run via the CGI mechanism, and communicate via request queues with tasks running the MG document indexing and compression software (Witten et al, 1994). Much of the 'glue' software is written in Perl (Wall et al, 1996) and requires the large Perl interpreter and software library to be in memory.

In contrast, the CD-ROM version of the software is a single integrated piece of software incorporating the Web server, digital library pre/post processing, and MG. Only a single index need be in memory at any one time, as a CD-ROM usually only holds a single collection. All of the software is coded in C and C++ to avoid the significant overhead involved in using a Perl interpreter. The result is a system which will work satisfactorily on a workstation with 8 or 16 MB of main memory (depending on the memory requirements of the workstation's operating system).

A browser is directed to access the server in one of two ways. The simplest is to use the URL http://127.0.0.1 (127.0.0.1 means ' local machine'). Once the first page is loaded, further pages are referenced relative to the starting page, and so are also obtained from the server. This is convenient in that it requires no set-up on the browser. The alternative is to set the browser to use 127.0.0.1 as its 'proxy'. This means that all page requests are routed to the server. It functions like a fixed cache, satisfying requests when it can and passing demands that it cannot handle on to an external network (if available).

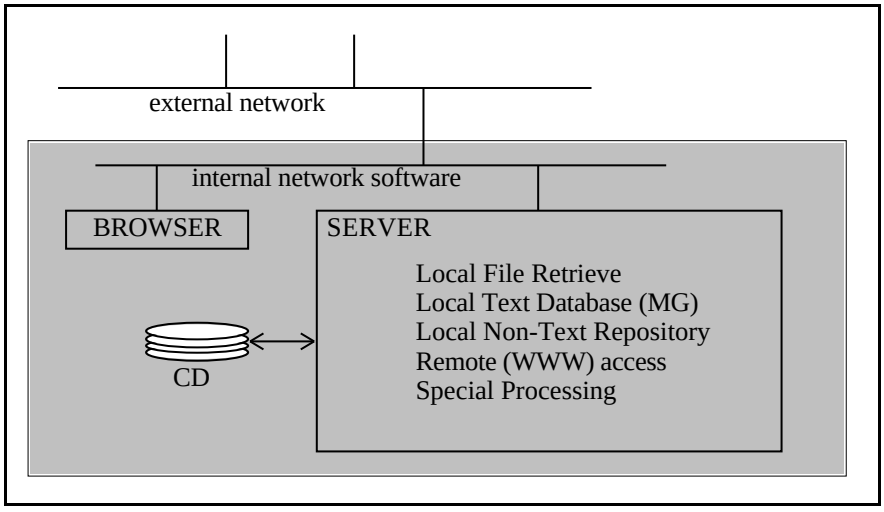

Figure 1: Browser-Server Interface

The server handles incoming page/file retrieval requests according to the requested item's availability and form of storage. If a page is not available locally, the request may be passed on to an external network. If each page or document in a collection is stored in a separate file, then a local file request can access the item on the CD-ROM. However, in general we avoid storing a collection's documents in separate files, because large numbers of files use CD-ROM space inefficiently. Instead, document files containing text are stored (and the extracted text is indexed) in an MG database, and non-text files are stored in a special repository file. The server has an index of the documents held in the MG database and the file repository. Incoming requests are checked against this index and may be retrieved from MG or the repository as appropriate. Major savings in collection storage requirements are possible by taking advantage of MG for text storage: typically text compresses to 25% of its original size, and the compressed index occupies around 7% of the size of the original text. This leads to a total storage requirement for the indexed collection of approximately one-third of the size of the original text alone. The system can also support a variety of types of

non-text items in the collection—audio, images, video clips—simply by including appropriate viewing utilities on the CD-ROM. For searching, the non-text items are represented by textual descriptions in the MG index.

A request which requires some computation on the server, such as the submission of a query from a user, would normally be handled with CGI requests. On our system, such requests are invoked by URL's starting http://127.0.0.1/server/. These are internally routed to handler routines within the server itself – particularly to MG components.

The major implementation difficulty experienced was with the IP network software, on machines which did not have network cards or modem software. To avoid installation complexity we chose to implement our own network layer to be used on such machines. In the absence of networking software the server loads our internal network software and communicates using that.

3. Searching and navigating a collection

The primary access method for documents in the United Nations University collection is keyword search (Figure 2a). The system supports searching over the *full* text of the document—not merely a document surrogate as is common in many commercial retrieval systems. While other collections we have built support a syntax for full Boolean searching, early user feedback from a similar document set (the Humanitarian Development collection, put together by the Global Help Project) indicated that Boolean searching was more confusing than helpful for the targeted users. Previous research suggests that difficulties with Boolean syntax and semantics are common, and are observed in diverse user groups (Borgman, 1996; Greene et al, 1990). Transaction log analysis over a number of library retrieval systems indicates that the most popular Boolean operator by far is the AND, with the Boolean OR and NOT rarely present in queries (Peters, 1993); we have confirmed this result in another New Zealand Digital Library collection (Jones et al, 1998). For all these reasons, the United Nations University interface default is ranked retrieval. However, to enable users to construct high-precision Boolean AND searches where necessary, selecting "search…for ALL the words" in the querying string produces the syntax-free equivalent of an AND query.

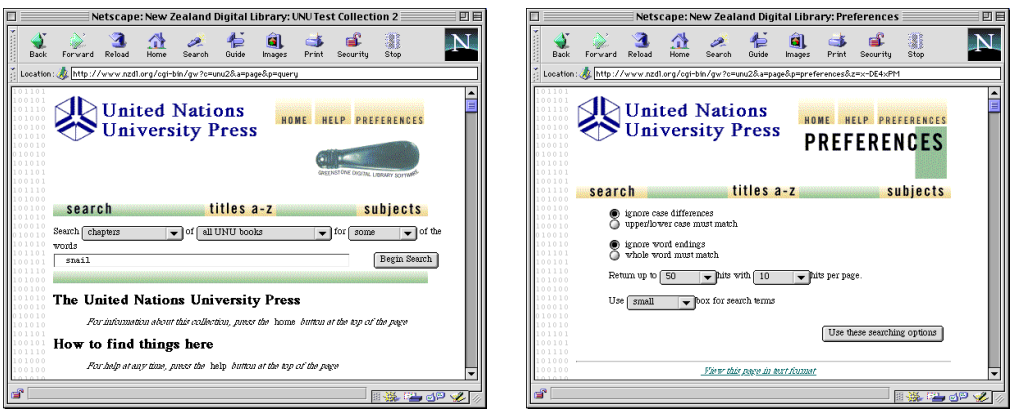

Figure 2: (a) Initial search screen for the UNU collection and (b) search preferences page

By default, search terms are stemmed and case differences are ignored. Most transaction log analysis from library online catalogs, digital libraries, and WWW search engines indicates that users tend to submit extremely brief queries. For example, the average query length for the New Zealand Digital Library's *Computer Science Technical Report* collection is only 2.5 words (Jones et al, 1998), a typical figure mirrored in retrieval studies conducted over two decades (Sandore, 1993). With such

brief queries the major difficulty encountered with search results is low search recall—hence the system automatically expands the query through stemming and case folding. These defaults can be modified by

The initial search screen (Figure 2a) also permits users to specify the "granularity" at which their search is done (that is, the size of the text against which the query is matched). Choices include *title*, *paragraph*, *same chapter or section*, and *book*. By selecting the smaller passage sizes, users can achieve a greater search precision, while selecting the larger ones tends to give a higher recall. Regardless of granularity, the results are always displayed in terms of a complete book, opened at the appropriate place.

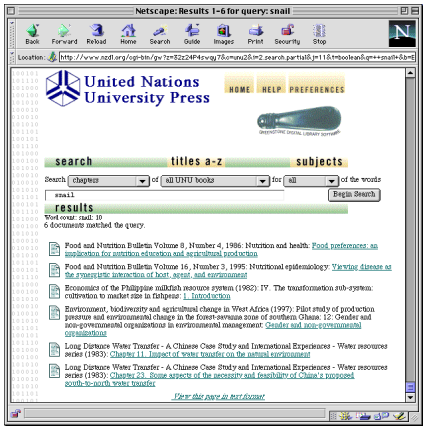

Figure 3: Query results page

We support browsing by taking advantage of the fact that the hierarchical structure of United Nations University Press documents is marked up in the document files. When an item in the "query results" list is selected (Figure 3), the user is presented with a photograph of the document's front cover and a table of contents with an arrow marking the item's position in the contents (Figure 4). Folders can be clicked open or closed, allowing the user to travel up and down the document's structure (in Figure 5, moving from a report up to the section headings for that issue of the bulletin). Clicking on "expand contents" will expand out the whole table of contents so that the user can browse the titles of all chapters and subsections to get a detailed view of the entire contents. "Expand text" displays the whole text of the current section or book, which is particularly useful when printing a complete work.

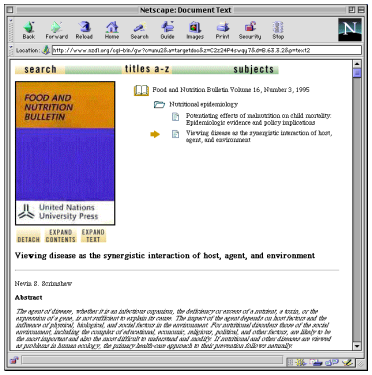

Figure 4: Viewing a selected item in the query results list

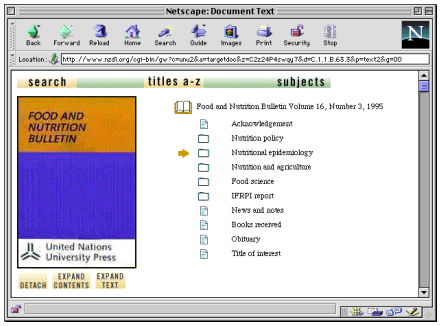

Figure 5: Moving up the document structure hierarchy

Browsing or searching by subject is supported by clicking the "subjects" button on the menu options bar of any search or results page . This brings up a list of subjects, represented by bookshelves (Figure 6). Users can click on any bookshelf to look at books on that subject, and click on a book to read it. Similarly, clicking on the "titles" button allows the user to browse through an alphabetized list of titles. If the user is currently viewing a document when the "subjects" or "titles" button is clicked, s/he will be taken to the place in the subjects or titles list that corresponds to that book. This supports the user in browsing for books on the same subject, or for books with similar titles.

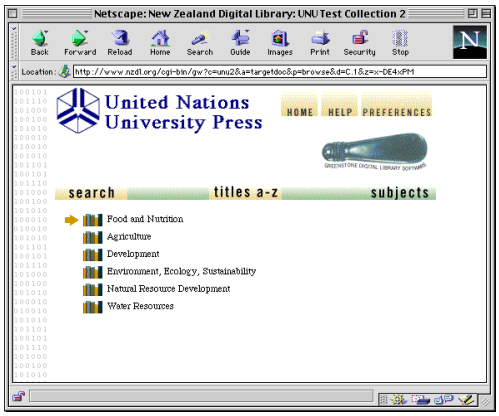

Figure 6: Browsing by subject

4. Conclusions

Despite near-universal current practice, the World-Wide Web is by no means the only way to deliver digital library services. Local networks and CD-ROM disks can be a viable alternative—and a necessary one in many operating environments. The humble CD-ROM can hold a lot of text, and DVD disks will enable easy distribution of very substantial collections

The challenge is to produce a scheme which can be used for distribution over each of these media, and look just the same to the user. The Greenstone software allows information to be made available in precisely the same form, using precisely the same interface, on a single-user (PC) computer, a local intranet, or the World-Wide Web. One reason for developing this technology was to permit access to important information in the Third World, which runs the risk of falling further behind because of inadequate network access. However, all who find the Internet capricious in terms of remote site availability, and suffer from highly variable and unpredictable network delays, will appreciate the advantages of having digital library information on site—whether in single-user or shared mode.

The United Nations University collection that we have described and illustrated is designed not, as most digital libraries seem to be, for technophiles, but for ordinary people with little or no computer experience. We have again run counter to common practice here to make the interface plain and easy to use. In a quest to improve usability for the ordinary person we have sacrificed features—actually deleted them from our software—that, although powerful, we have observed to be rarely employed by real users answering their real information needs.

References

Borgman, C.L. (1996) Why are online catalogs still hard to use*? Journal of the American Society for Information Science* 47(7), pp. 493-503.

Chowdhury, G.G. (1996) Developing modern information systems and services: Africa's challenges for the future, *Online & CDROM Review* 20(3), pp. 145-146.

El-Hadidy, B. (1994) The breakeven point for using CD-ROM versus online: a case study for database access in a developing country, *Journal of the American Society for Information Science* 45(4), pp. 273-283.

Greene, S.L., Devlin, S.J., (1990) Cannata, P.E., and Gomez, L.M. No Ifs, ANDs or Ors: a study of database querying, *International Journal of Man-Machine Studies* 32(3), pp. 303-326.

Holmes, G., and Rogers, W.J. (197) Gathering and indexing rich fragments of the World-Wide Web, *Proceedings of the International Conference on Computers in Education 1997* (Sarawak, Malaysia, Dec. 2-6), pp. 554-562.

Jones, S., Cunningham, S.J., and McNab, R. (1998) An analysis of usage of a digital library, *Working Paper 98/13*, Department of Computer Science, University of Waikato (Hamilton, New Zealand.

Peters, T. (1993) The history and development of transaction log analysis, *Library Hi-Tech* 11(2), pp. 41-66.

Sandore, B. (1993) Applying the results of transaction log analysis, *Library Hi-Tech* 11(2), pp. 87-97.

Wall, L., Christiansen, T., and Schwartz, R.L. (1996) *Programming Perl.* O'Reilly, Sebastopol (CA, USA).

White, W.D. (1992) CD-ROM in developing countries, *CD-ROM Professional* (May), pp. 32-35.

Witten, I.H., Moffat, A., and Bell, T.C. (1994) *Managing Gigabytes*. Van Nostrand Reinhold, New York, New York.