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

Ian Witten, Rodger McNab, Sally Jo Cunningham, Bill Rogers, 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 World-Wide Web and offers a wide variety of information collections. Subcollections can be written to a Greenstone 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. Alternatively, the same CD-ROM acts as a network server to make exactly the same information available, over any network that supports the http protocol, to others who need only use their standard Internet browser software. This technology has great appeal for users in developing nations, where Internet access to a non-local document collection can be precarious or prohibitively expensive.

1. Introduction

The emerging digital library field is very much an offspring of the Internet and World Wide Web—current digital library efforts concentrate primarily 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. These Web-based digital libraries share significant advantages over their online predecessors: users are not required to 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; and Web browsers provide a simple, standard means of access to a variety of digital library systems.

The Greenstone digital library software, developed by the New Zealand Digital Library project, is unique in that it 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 sort of 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 alternatively allows those users currently without Internet access to more easily step up to Web searching and browsing as it becomes available to them.

The necessity for supporting this range of access methods was dictated by practical experience in digital library development. As potential user groups were identified and collections tailored for them, it was noted that universal access via the Internet was neither possible nor desirable for several systems. 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 was 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. Additionally, a CD-ROM is relatively durable in the face of harsh environmental conditions, and incurs known, fixed costs for purchase of it and supporting hardware (White, 1992). A CD-ROM based digital library carries the further advantage of providing the documents themselves—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).

An earlier version of Greenstone has been used in a university-level distance learning course on computer literacy, where lecturer-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 even 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 from wasting lab time exploring sites that irrelevant to the task at hand, or that are inappropriate for the students' age groups.

The Greenstone-based digital library described in this paper is comprised of a set of documents provided by the United Nations University Press, 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 the information users.

Section 2 describes the Greenstone architecture. Multimedia collections can be developed under Greenstone, with a single collection text, images, audio, and even video clips. Compression technology is used to ensure that the greatest possible volume of information is packed on to a CD-ROM. The interface software combines easy-to-use browsing with powerful search facilities. As discussed in Section 3, there are several ways to find information in a Greenstone collection; a user can conduct keyword searches, access known documents by title, or browse subject "bookshelves. Issues involved in setting up new collections under the Greenstone software are discussed in Section 4.

2. System architecture

Cobble some stuff from earlier papers?

3. Searching and navigating a collection

The primary access method for documents in the UNU collection is the keyword search (Figure n). 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 Greenstone collections support the full syntax for Boolean searching, early user feedback from a similar document set (the Humanitiarian Development collection, put together by Greenstone and the Global Help Project) indicated that Boolean searching was more confusing than helpful for the targeted users. Hence, the UNO interface default is ranked retrieval. 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 Greenstone collection (Jones et al, 1998). To finesse this problem while still permitting users to construct high-precision Boolean AND searches, selecting "search…for ALL the words" in the querying string produces the syntax-free equivalent of an AND query.

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 Greenstone's Computer Science Technical Report Collection is only 2.5 words (Jones et al, 1998), a typical result mirrored in retrieval studies conducted over two decades (Sandore, 1993). With such brief queries the major difficulty encountered with search results is a low search recall—hence the system automatically expands the query through stemming and case folding.

The advanced search page 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, or book. By selecting the smaller passage sizes, users can achieve a greater search precision, while selecting the larger passages for matching 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.

We support browsing by taking advantage of the fact that the hierarchical structure of UNU Press documents is marked-up in the document files. When an item in the query results list is selected, 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 n). Folders can be clicked open or closed, allowing the user to travel up and down the document's structure (in Figure n, 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. "Expand text" will display the whole text of the current section or book.

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 n). 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 alphabatetized 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.

4. Building a collection

In general, how to set up a new collection, thoughts on how the interface might differ, …

5. Conclusions

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.

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