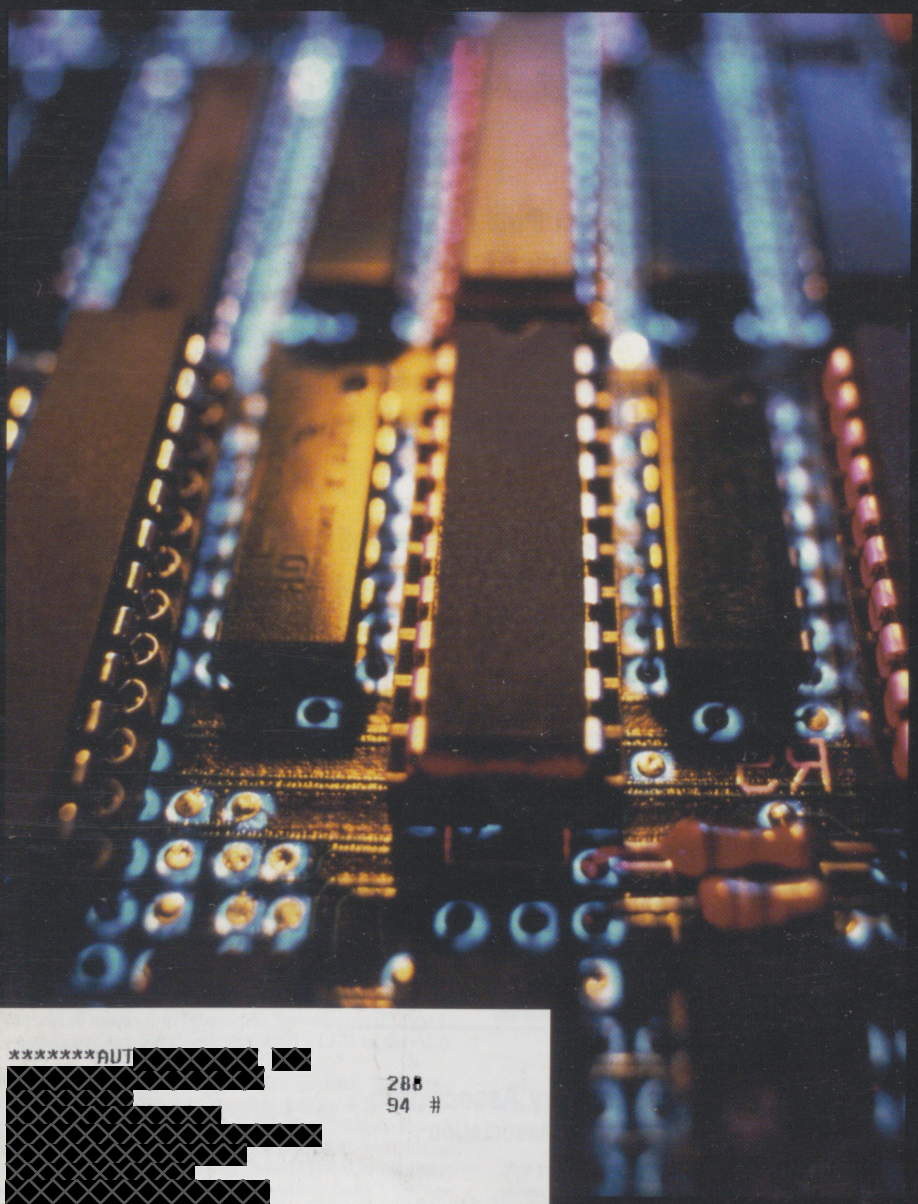
Information Technology AND Libraries

ISSN 0730-9295

Volume 23 Number 1

MARCH 2004



IN THIS ISSUE

- The Internet Archive and Its Role in Preserving Digital Media
- A Decentralized Image Retrieval System for the STEM Digital Library
- How Scholars Approach an Academic Library Web Site

*******AUT
288
94 #

LITA NATIONAL FORUM

OCTOBER 7–10, 2004
St. Louis, Missouri

10 Years of Connectivity
Libraries, the World Wide Web, and the Next Decade

Plan now to join us in St. Louis at the Sheraton West Port Chalet Hotel for the Forum, a three-day educational event that includes preconferences, keynote sessions, and more than thirty concurrent sessions.

First held in 1998, the LITA National Forum is a highly regarded annual event for those whose work involves new and leading edge technologies in the library and information technology field.

Registration is limited in order to preserve the advantages of a small conference. Take advantage of the informal Friday evening reception and networking dinners to get to know LITA leaders, Forum speakers, and peers.

At the Forum, LITA will host a Sponsor Showcase. As many as twelve sponsors will be on hand to discuss their latest products and services. Forum attendees will have ample time to visit the Sponsor Showcase.

Early registration begins in June 2004. For more information, watch www.lita.org/forum04. Updates will be posted regularly.

---

a tradition of the
Library and Information Technology Association
A division of the American Library Association

Dan Marmion, Editor
University of Notre Dame, Notre Dame, Indiana (dmarmion@nd.edu)
Marc Truitt, Managing Editor
University of Houston (mtruitt@uh.edu)
Tom Zillner, Book Review Editor
Wisconsin Library Services, Madison, Wisconsin (tzillner@wils.wisc.edu)
Andy Boze, Software Review Editor
University of Notre Dame, Notre Dame, Indiana (Boze.1@nd.edu)

EDITORIAL BOARD: **Judith Carter; Judith Gelernter,** The Union Club, New York; **Eddy Hogan,** California State University, Monterey Bay; **Alan Manifold,** Purdue University; **Scott Nicholson,** Syracuse University; **Paul Soderdahl,** University of Iowa; **Joan Frye Williams,** Information Technology Planning, Management; **Melinda Stowe,** University of Rochester, ex officio.

Information Technology and Libraries (ISSN 0730-9295) is published quarterly in March, June, September, and December by the American Library Association, 50 E. Huron St., Chicago, IL 60611. It is the official publication of the Library and Information Technology Association, a division of the American Library Association. SUBSCRIPTIONS: Subscription price, $30, is included in membership dues. Nonmembers may subscribe for $55 per year in the U.S.; $60 in Canada, Mexico, Spain, and other PUAS countries; $65 in other foreign countries. Send subscription requests to *Information Technology and Libraries,* Subscription Department, American Library Association, 50 E. Huron St., Chicago, IL 60611; 800-545-2433; fax: (312) 944-2641; subscription@ala.org. Single copies, $20. Periodical-class postage paid at Chicago, Illinois, and at additional mailing offices. POSTMASTER: Send address changes to *Information Technology and Libraries,* 50 E. Huron St., Chicago, IL 60611.

Information Technology and Libraries publishes material related to all aspects of libraries and information technology, including digital libraries, metadata, authorization and authentication, electronic journals and electronic publishing, telecommunications, distributed systems and networks, computer security and intellectual property rights, technical standards, geographic information systems, desktop applications, online catalogs and bibliographic systems, optical information systems, software engineering, universal access to technology, futuristic forecasting, library consortia, vendor relations, and technology and the arts. *ITAL* welcomes unsolicited manuscripts. Submissions should follow the guidelines stated under "Instructions to Authors" on page 96 of the June 2003 issue or on the LITA Web site at www.lita.org/ital. Manuscripts of feature articles, communications, and tutorials should be addressed to Dan Marmion, 221 Hesburgh Library, The University of Notre Dame, Notre Dame, IN 46556-5629. Books for review should be addressed to Tom Zillner, WILS, 464 Memorial Library, 728 State St., Madison, WI 53706. Software for review should be addressed to Andy Boze, Hesburgh Library, The University of Notre Dame, Notre Dame, IN 46556-5629.

ADVERTISING: The Goldman Group, 14497 N. Dale Mabry Hwy., Ste. 205N, Tampa, FL 33618; telephone (813) 264-2772; fax (813) 264-2343.

As a matter of policy, *ITAL,* as the scholarly organ of LITA, does not review LITA publications.

PRODUCTION: ALA Production Services (Troy D. Linker; Ellie Barta-Moran, Angela Hanshaw, Kristen McKulski, and Karen Sheets), American Library Association, 50 E. Huron St., Chicago, IL 60611.

Publication of material in *Information Technology and Libraries* does not constitute official endorsement by LITA or the ALA.

Abstracted in *Computer & Information Systems, Computing Reviews, Information Science Abstracts, Library & Information Science Abstracts, Referativnyi Zhurnal, Nauchnaya i Tekhnicheskaya Informatsiya, Otdyelnyi Vypusk,* and *Science Abstracts Publications.* Indexed in *CompuMath Citation Index, Computer Contents, Computer Literature Index, Current Contents/Health Services Administration, Current Contents/Social Behavioral Sciences, Current Index to Journals in Education, Education, Library Literature, Magazine Index, NewSearch,* and *Social Sciences Citation Index.* Microfilm copies available to subscribers from Proquest, Ann Arbor, Michigan.

The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences—Permanence of Paper for Printed Library Materials, ANSI Z39.48-1992.∞

Information Technology AND Libraries

Volume 23, Number 1 March 2004 ISSN 0730-9295

Abstracts and contents pages of recent issues of *ITAL* can be found on the LITA Web site at www.lita.org/ital.

3 Ephemeral to Enduring: The Internet Archive and Its Role in Preserving Digital Media
ELI EDWARDS

9 DIRECT: A Decentralized Image Retrieval System for the National STEM Digital Library
JINSHAN TANG, SRIDHAR R. AVULA, AND SCOTT T. ACTON

16 Designing for Experts: How Scholars Approach an Academic Library Web Site
THURA MACK, MARIBETH MANOFF, TAMARA J. MILLER, AND ANTHONY D. SMITH

COMMUNICATIONS

23 Content Management Strategy for a College Library Web Site
MARK DAHL

30 Thermal-Transfer Printing: A Better Way to Print Library Labels
CHERYL D. WALTERS

37 Book Reviews

40 Index to Advertisers

cover design Kevin Heubusch ● **interior design** Dianne M. Rooney

Copyright ©2004 American Library Association.

All material in this journal subject to copyright by ALA may be photocopied for the noncommercial purpose of scientific or educational advancement granted by Sections 107 and 108 of the Copyright Revision Act of 1976. For other reprinting, photocopying, or translating, address requests to the ALA Office of Rights and Permissions.



I count on LITA's programs, publications, and working groups to give me a measure of what's important to library automation.
–Karen Coyle, University of California Digital Library

LITA is the association that **educates, serves, & reaches out** to the library and information technology community.

Find out more about the benefits of LITA membership today at
www.lita.org!

Library & Information Technology Association **LITA**

lita@ala.org • 800.545.2433, ext. 4270

a division of the American Library Association

Ephemeral to Enduring: The Internet Archive and Its Role in Preserving Digital Media

Eli Edwards

The Internet was first created as a communications network, but it has produced myriad effects in many sectors of technology, business, and society. Among these important changes are the growth of electronic publishing and the increasing use of the Internet as a medium for disseminating diverse documents, literature, art, journalism, commentary, and miscellaneous expressions of information and thought. Digital media, and the rise of Internet distribution of it, has caused profound changes in librarianship as well.

In spring 2001, it was estimated that more than 93 percent of all new information is "born digital"—originating in digital form.[1] But just as the Internet, particularly the World Wide Web, is noted for its growing capacity and diversity of material, it is also notorious for its impermanence. The average life span of a Web site is approximately six weeks.[2] Lyman cites that 44 percent of Web sites online in 1998 were not found the following year.[3]

How to preserve digital media has been a growing concern of librarians, archivists, computer scientists, and scholars of various backgrounds. Digital libraries have been formed—often as cooperative efforts between similar brick-and-mortar libraries—and libraries have developed protocols to add digital media and bibliographic data to their current collections. According to Feldman, "these archiving activities—important and revolutionary as they are to the future of information service—still focus [on] reproducing print products in a digital venue."[4]

In addition, there is increasing pressure to produce and distribute information in solely digital formats in many sectors of publishing. However, there are grave concerns that governmental and commercial interest in preserving digital media has been less than enthusiastic, and thus the results have been less than rigorous or impressive.

Archiving the Internet

There still remains the challenge of archiving the Internet itself. There is a significant growth in Internet-only material, including multimedia (images, audio, and video), journalistic accounts, fiction, personal journals, academic and business documents (papers, presentations, and proceedings), and government documents. Scholars view the Internet as an important artifact that warrants archiving and preservation, but efforts to archive the Internet have been piecemeal. Right now, there is no international standard, no agreed-upon protocol, and no single consortium that has developed a way to archive the Internet.

In the absence of a centralized effort, a number of Internet-archival projects have arisen. Most of these projects are in the purview of national libraries, such as the British Library's new project to archive Internet Web sites, focusing on sites of British interest.[5] As in the case of the U.K. effort, most government and national libraries have strategized to preserve national material—both official government Web sites as well as Web pages—that is registered under each nation's domain.

Among these projects is one that is easily set apart from the others. The most well-known of the Internet-archiving projects and, in some ways, the most controversial, it is also one of the earliest projects of its kind and it has the broadest mission. The Internet Archive (IA) has proven to be both an anomaly and a vanguard in its attempt to archive as much of the Internet as it can via its technological processes. What IA collects, how it collects, and how it gives access to its collections are important considerations in defining and evaluating the role of Internet-archival projects, particularly as a major activity of digital librarianship.

IA was founded as a nonprofit organization in 1996 by Brewster Kahle, founder and CEO of Alexa Internet. IA is the noncommercial sister organization of Alexa Internet (now a wholly-owned subsidiary of Amazon.com), and the two companies share technology. Alexa's Web crawlers search the Internet and accumulate Web pages and related information to those pages. After a six-month period, the pages themselves go to IA, while the information is used by Alexa to add such functionalities as listings of related sites based on users' patterns and such bibliographic information as the site's domain and what organization is running it, how long the site has been running, user ratings, and basic subject cataloging for the Web site.

What Is in the IA?

As of January 2004, IA is in excess of 300 terabytes of data, which includes more than 1,200 short films in MPEG-2 and MPEG-4 formats and more than 30 billion Web pages. Right now, IA has the following collections:

Web Sites

IA has several collections of Web pages, including the following:

Eli Edwards (eedwards@slis.sjsu.edu) is an MLIS student at the School of Library and Information Science, San Jose State University.

- a general collection of Web sites (now accessible via the Wayback Machine interface);
- two collections of Web sites from the 1996 and 2000 presidential elections;
- Web pages containing content related to the terrorist attacks on the World Trade Center and the Pentagon on September 11, 2001 (news, reactions, commentary); and
- a Web Pioneers collection highlighting early commercial and organizational sites on the Web.

Multimedia

IA's multimedia collections have historical and such current material as live and animated images. The collections include the following:

- a movie archive of more than 1,200 ephemeral short films (which have limited copyright or are part of the public domain) donated by the Prelinger Archives; and
- a new sister site (www.televisionarchive.org) that has worldwide broadcast news from September 11, 2001.

Digitized Content

IA also offers access to digitized text archives, including the following:

- The ARPANET collection, commonly referred to as the precursor of the Internet, includes "memoranda, interview notes, periodicals, papers, and other materials documenting the development of the Advanced Research Projects Agency Network (ARPANET) of the U.S. Department of Defense"[7]
- The Million Book Project
- Project Gutenberg
- Children's Library, the entire collection of the International Children's Digital Library (see the following paragraphs for more information on this project)

While IA has always proclaimed free access to researchers and the public for its collections, the methods for accessing the growing collection of Web pages was a barrier to most casual users of IA. Researchers interested in accessing any of the Internet collections had to submit proposals and were provided with Unix shell accounts on the IA's servers; downloading content required some knowledge of Unix programming as well as account maintenance.

This changed in October 2001, when IA introduced the Wayback Machine, a custom-search engine that allowed users to enter a URL and receive—from that URL—links to pages that have been archived, whether the Web site is currently active or dead. One feature of the Wayback Machine is the ability to search for video, audio, and images in addition to text, binary, and PDF files.[8] The interface is designed solely to work with the collection of Web sites; access to FTP and Usenet sites is still limited to the former procedure.

IA's one hundred terabytes of data are stored on digital linear tape (DLT) and Integrated Drive Electronics (IDE) hard drives in archive format—100-megabyte ARC files consisting of multiple individual files.[9] To aid in long-term preservation, the collections are copied on multiple sites, and IA plans to migrate the data from current to new DLT tape at least every ten years; it also is collecting software and emulators for current data formats that may become obsolete in the future.[10]

IA has taken as its mission to preserve and to offer free unlimited access to its collections, with the implicit philosophy that the Internet is not only a medium of cultural artifacts, it is itself an artifact that warrants the same consideration and efforts toward preservation and access that books, newspapers, periodicals, films, and paintings currently have.

Gaps in IA

Despite the technological efforts, IA has major gaps in its basic collection. The Web became part of the infrastructure of the Internet in 1992; IA is missing material from the first four years of the Web. In addition, while there is a small amount of non-Web material from the Internet (such as the aforementioned Usenet and FTP archives), IA is lacking in content from other prominent pre-Web protocols, especially Gopher.

In addition, the technology is unable to retrieve or archive restricted-access Internet material (whether via password protection or Internet protocol [IP] recognition), as well as Web sites on secure servers. This excludes material that is part of fee-based subscriber services (online databases and journals), as well as free content that requires registration for access. There is also the problem of *orphan pages*, Web sites that exist in isolation with no links to or from their locations to other sites. Since Alexa's/IA's technology works by following links from one site to another, a site with no links to the rest of the Web is inaccessible and will go unarchived (although Alexa does have a feature for Web-site owners to submit their sites for archiving). The Web crawlers are limited by how many times they may visit a domain to collect new or modified pages; while the intervals of new archival material for a site have been drastically reduced, there are still sizeable gaps in content that is lost between each archived page.

Another technology hurdle IA faces is a problem with certain types of dynamic content on Web pages. Cascading style sheets (CSS), PHP programming, and other XML protocols that render Web pages in standard, localized HTML

can be archived and displayed as intended from IA. Other types of dynamic content, such as JavaScript, will not work as originally programmed if the site is archived. For any dynamic content that requires access to the originating host (not the server that the page itself is stored on), the links to the originating host are broken in the automated archiving process. Thus, the archival copies of the site will have reduced functionality in comparison to the original site or may not function at all.

Moreover, there are ways for content providers to, intentionally or unintentionally, exclude their material from IA. IA abides by the standard for robot exclusion—its Web crawlers will not archive any site that has a robots.txt file in the site's server directories.[11] It also maintains the policy of removing material from IA at the request of the Web-site owner. Since September 11, 2001, a significant number of federal agencies have removed or revised their Web sites due to concerns regarding sensitive data being too accessible to potential terrorists. The U.S. Nuclear Regulatory Commission (NRC) asked IA to remove copies of the NRC Web site after the agency deleted information about the 103 nuclear reactors currently operating in the United States. IA complied with the request, and the original pages are no longer available.[12]

Copyright Issues

IA has been conceived as not only a repository of Internet material, but also as an Internet library, specially formed to preserve, store, and provide access to artifacts in digital format. The directors of IA consider its activities, and its very mission, as a function of scholarship and preservation of historical and cultural material; thus, its archival activities are a fair-use exception to laws regarding access of copyrighted material.[13] Some content providers disagree.

In a letter to the *Chronicle of Higher Education,* Professor Stephen R. Brown of American University complained that "[t]he Internet Archive is nothing more than an enormous copyright violation disguised as a library."[14] Newspapers, especially those that have used their digital and microfilm archives as profitable ancillary revenue streams, have also complained that IA is undermining their archival collections by offering the same material for free, without the express permission of the papers. However, since IA does not provide daily or weekly archival copies of newspaper or journal Web sites and is willing to remove material on request, the complaints have been somewhat muted.

How IA handles such copyright issues has significant affect on its collections. As a private nonprofit, it relies on philanthropic contributions not only for its funds but also for its collections. Its technical and operational infrastructure is based on automated processes; it does not seek permissions or special access to any Internet material. Therefore, IA is limited to the public, or visible, Web. However, it cannot penetrate, and thus cannot preserve, the invisible Web: the sites (newspapers, databases, subscription-based journals, fee-based sites) and information sources that comprise a large part of the Web but are not accessible to every user.

The IA Is Not a Library

However, even though IA bills itself as a library of the Internet, it lacks many basic functions and controls exercised by libraries. Currently, IA does not index or catalog its collections; classifications are based solely on the Web site's URL (the Wayback Machine does have an advanced feature that recognizes multiple aliases for a single URL). There is only one access point for searching IA—the URL. Currently, neither IA nor its interface has any keyword search capability and cannot provide bibliographic descriptions of the material.

And unlike libraries, there is a very limited notion of collection development for IA. Its goal is to archive as much of the Internet as its technology will allow. There is no judgment on the veracity, quality, or appropriateness of the content being archived. And there are no filters to prevent age-inappropriate material or content censored in another jurisdiction from being accessed by anyone who knows the URL for such sites. IA's general collection includes such controversial Web sites as www.rotten.com (a site that collects highly graphic and explicit photos from autopsies, accidents, and crime scenes), neo-Nazi and racial-supremacist sites, and anti-abortion sites that have been accused of provoking violence against women's clinics and medical personnel (such as www.christiangallery.com/atrocity or the "Nuremberg Files" Web site).

IA is interested not only in preserving "born-digital" items; it sees a role for itself in carrying out digital preservation of print material. To accomplish this, IA is expanding its partnerships with other organizations, including traditional libraries, to provide digitalization and electronic dissemination of bibliographic resources. It has announced two high-profile projects in this regard:

- With the University of Maryland's Human-Computer Interaction Lab, IA will establish and maintain the International Children's Digital Library, "a large-scale digital archive of literature for children ages three to thirteen."[15] The project, with assistance from the Library of Congress, will begin with a collection of 225 digitized children's books and will eventually grow to ten thousand titles from more than 100 countries.
- IA has donated a copy of its entire collection to the new Bibliotheca Alexandrina (Alexandria Library), as well as two thousand hours of Egyptian and American television broadcasts. IA also donated a

book-scanning facility to digitize books in Arabic that would be available via interlibrary loan. The digitized holdings will also be available via IA.

According to Kahle, the Internet is a medium for artifacts that are considered to be ephemeral, both bibliographic and nonbibliographic.[16] IA wants to make those artifacts enduring for current and future scholars and the public. It provides resources for cultural and historical scholarship, the study of technology's effects on society (cultural, legal, economic), and "snapshots in time" of historical events. IA also extends the functionality of the Internet. Even very popular Web sites will only exist as long as the Web-site owner is willing and able to continue and maintain the site. The current recession and its extremely destabilizing effect on high-tech companies have especially affected new media and Internet-only content providers. This has led to sudden removals and disappearances of popular niche and mainstream Web sites. However, archival copies of these dead sites allow for continued access to their content.

Library-Based Internet Archiving Projects

Web and Internet archiving projects that are developed and implemented by libraries tend to use a different model and have a different focus. Such projects include:

- PANDORA (Preserving and Accessing Networked Documentary Resources of Australia)—National Library of Australia
- AOLA (Austrian On-Line Archive)—a joint project between the Austrian National Library and the Vienna University of Technology
- Our Digital Island—State Library of Tasmania
- NedLIB (Networked European Deposit Library)—an EU-funded consortium that includes libraries in the Netherlands, France, Norway, Finland, Germany, Portugal, Switzerland, and Italy
- Kulturarw³—The Swedish Archive[17]

All of these efforts, which include the Bibliothèque nationale de France and Die Deutsche Bibliothek (the German National Library), are based on a primary strategy to preserve national material on official government Web sites as well as Web pages that are registered under the nation's domain.[18] In addition, the vast majority of these projects are based in part or in whole on the manual selection, collection, and metadata treatment of Web sites to meet the criteria of national or cultural interest, with limited use of automatic *harvesting*—that is, the use of robot or spider software to search the Web for eligible material and capture it for archiving. The Swedish Archive is the only other archive besides Kahle's to draw all of its material via harvesting; however, what is harvested is still determined by the criterion of Swedish interest.[19]

There is another key distinction between the emerging library model of Internet archiving and the methods used by IA—permissions. As mentioned above, IA does not seek the permissions of domain owners or copyright holders when amassing content for preservation, though it will remove any page or site on the basis of a complaint by the content owner. The scope of IA makes any attempt to license proprietary information or otherwise negotiate permissions for desired content a technical and logistical impossibility at this time. The philosophical underpinnings of IA, as articulated by its founder and chief spokesperson, buttress this: Kahle perceives the Internet as a unique historical artifact that will be of invaluable use to current and future scholars. Thus, IA's attempt to store and provide access to digital content, whether proprietary or in the public domain, is inherently covered by the fair-use exception to copyright regulations.

The archival models used by national libraries have taken a much different approach. Most of the projects have made the negotiation of licensing and permission to access, preserve, and display proprietary digital content an initial and central task. Die Deutsche Bibliotek, for instance, is working with German publishers to create and maintain a digital deposit for their online journals, periodicals, and other Web-based material.[20]

There are weaknesses to this library model, one of which is the probability of underestimating the historical or scholarly value of material, particularly by favoring commercial, well-designed content over amateur content (such as personal Web pages). Rauber, Aschenbrenner, and Schmidt of AOLA mention a couple of assumptions that are used to argue against the archiving of personal or trivial content: "Who is interested in some fellow's homepage?" and "Important information is published in 'real' media anyway." They respond to these assumptions by arguing that "Letters [are] more interesting than books," and that "ads, posters, and snippets tell more about a society than 'high-quality information sources.'"[21] They then ask rhetorically, "What if only codices had been preserved [from previous eras]?" What may appear to be detritus or vanity projects today might provide important contextual clues to future historians and other social scientists. In this way, IA's holdings and overall collection policy may significantly counteract the contextual gaps in other archives of the Internet.

Advice from the Founder

However, IA is intent on the goal of being an Internet library, not simply an auxiliary repository of digital doc-

uments. Founding director Kahle regards electronic disseminators of information as digital librarians and has suggested a code of ethics for digital librarianship. The suggested code includes the following tenets:

- Don't give away user logs except for scholarly use.
- Take the job of information serving seriously.
- Count on wide use of the information served, for good uses and bad, so be proud of the information and the collection.
- Users learn as much from a question that has no answer as from the ones with answers. This requires a complete and up-to-date collection.
- Assume that the patron will not know your affiliations . . . do not tempt patrons to use a service they would regret if they knew more about you.
- Respect your patrons.[22]

Libraries and IA

Lyman asks the following question: "Which profession should develop digital archives—librarians or computer scientists?"[23] His answer is that both professions offer overlapping strategies that include unique functionalities based on the philosophies of each discipline. Moreover, according to Lyman, IA illustrates the computer scientists' model of Internet archiving, one that is meant to preserve not only the content and information on the Internet but also the structure and interconnections of the Internet as a technical object.[24]

The potential for IA's successes and failures in preserving our past and present is already apparent to many in the technology realm. Scholars and technology are increasingly concerned with the possibility of a Digital Dark Age, a period in the not-so-far future when the manuscripts and ephemera used by historians, social scientists, and others to examine the past and present will not exist in a significant accumulation to yield useful historical or social context. At a conference of special librarians in 2003, historian David McCullough and futurist Stewart Brand both addressed the growing disparity of digital content to analog or physical media and the dire consequences of not being able to understand our history. Brand referred to the Digital Dark Age directly, while McCollough lamented that future historians will not have the manuscripts and diaries that have been so instrumental to his own work. Both noted the importance of ephemera to provide context to events, locations, and even structures.

In the meantime, IA's potential to assist in staving off this Dark Age is already being recognized. California has a large, well-organized state library, which maintains a catalog of material that the state has published, with information on how it can be accessed by other agencies and the public. A recent report on current efforts of state governments to provide permanent public access (PPA) to electronic information found that the state of California has made only inconsistent efforts to preserve access to digital information, despite the attempts of the state library: "Due to limited staff and other resources, it is often difficult for the state library to capture many electronic government documents."[25]

In a May 2003 catalog of California state publications, there was an entry for an e-commerce report published and placed on the state government Web site in PDF format in 2000. According to the catalog entry, the publication is out of print, but the PDF file was still accessible on the Web via IA. However, the PDF has since been removed from IA's server.

In effect, the state of California is relying on a private nonprofit to provide public access to public information.

The issues that IA has with digital media are similar or identical to the concerns and issues of physical and digital libraries. Defining fair use for digital-based information is a major area of concern. Libraries and IA share issues of digital rights management, storage, preservation, and access. While IA's mode of operation does not follow the traditional library model of archiving and preservation, its intent in archiving digital media is a familiar one within librarianship: to preserve and provide current and future access to scholars and the general public. Many information scientists, archivists, and others concerned with the future of preserving and accessing digital artifacts agree that the IA model and the emerging library model of Internet archiving are complementary efforts that have already shown some measure of success and have the potential to develop significant collections of digital material.

IA is just one project in a growing, collaborative effort to preserve an immense information system. Unlike most of those projects, IA is not funded directly by tax dollars nor is it under the administrative restrictions of a government agency. IA has been able to develop and implement procedures for Web archiving that have saved billions of bytes of data from obsolescence and neglect. Unfortunately, it will not be able to archive the greater part of the Internet, such as the orphaned sites, dynamic Web pages, or sites that restrict access based on passwords, IP recognition, or other protocols, which increasingly comprise what is termed the Deep Web. However, IA is playing a key role in the preservation of digital online media. In order to extend the life of information on the Internet, libraries, archivists, and computer scientists will need to examine IA and its successes and failures and draw lessons from what has been done by IA and other projects, as well as what still needs to be done, to preserve our growing yet precarious digital heritage.

Author's note: Eli Edwards is in no way affiliated with, or, in any way, connected to IA, Alexa Internet, or any of its principals, staff, board members, or donors.

References and Notes

1. Richard Wiggins, "Digital Preservation: Paradox and Promise," *Library Journal netConnect* (spring 2001): 12–15.

2. Jim McCue, "Can You Archive the Net?" *The Times* (London), Apr. 29, 2002.

3. Peter Lyman, "Archiving the World Wide Web," *Building a National Strategy for Preservation: Issues in Digital Media Archiving* (Washington, D.C.: Council on Library and Information Resources, 2002). Accessed May 5, 2002, www.clir.org/pubs/reports/pub106/web.html.

4. Susan E. Feldman, "'It Was Here a Minute Ago!': Archiving on the Net," *Searcher* 5, no. 9 (Oct. 1997): 52.

5. McCue, "Can You Archive the Net?"

6. Internet Archive, "Internet Archive: Archive Collections," The Internet Archive, Apr. 26, 2002. Accessed Apr. 13, 2002, www.archive.org/index.php.

7. Internet Archive, "Arpanet," The Internet Archive, Apr. 26, 2002. Accessed Apr. 13, 2002, www.archive.org/texts/arpanet.php.

8. Greg R. Notess, "The Wayback Machine: The Web's Archive," On the Net, *Online* 26, no. 2 (Mar./Apr. 2002): 61.

9. Internet Archive, "Storage and Preservation," About the Internet Archive. Accessed Apr. 13, 2002, www.archive.org/about/about.php#storage.

10. Ibid, "Preservation."

11. Jeffery Selingo, "In Attempting to Archive the Entire Internet, a Scientist Develops a New Way to Search It," *Chronicle of Higher Education* 44, no. 26 (Mar. 6, 1998): A27–A28.

12. "Now You See It," Net Effects, *Foreign Policy* no. 128 (Jan./Feb. 2002). Accessed Mar. 12, 2002, www.foreignpolicy.com/issue_janfeb_2002/net_effect.html.

13. Internet Archive, "Future Libraries: How People Envision Using Internet Libraries," About the Internet Archive, Oct. 4, 2002. Accessed Apr. 13, 2002, www.archive.org/about/about.php#future.

14. Stephen R. Brown, "Is Online Archive Fair Use?" Letters to the Editor, *Chronicle of Higher Education* 44, no. 34 (May 1, 1998): B9.

15. "Executive Summary," International Children's Digital Library—Info for Grown-Ups. Accessed Apr. 13, 2002, www.icdlbooks.org/adults/exec.html.

16. Internet Archive, "Future Libraries."

17. PANDORA Archive. Accessed Aug. 18, 2002, http://pandora.nla.gov.au/index.html.

18. AOLA. Accessed Aug. 18, 2002, www.ifs.tuwien.ac.at/~aola.

19. Our Digital Island. Accessed Aug. 18, 2002, http://odi.statelibrary.tas.gov.au.

20. NedLIB. Accessed Aug. 18, 2002, www.kb.nl/coop/nedlib.

21. Kulturarw^3. Accessed June 24, 2003, www.kb.se/kw3/ENG/Default.htm.

22. Andreas Aschenbrenner, "Related Work," Long-Term Preservation of Digital Material—Building an Archive to Preserve Digital Cultural Heritage from the Internet, Dec. 2001. Accessed Aug. 19, 2002, www.ifs.tuwien.ac.at/~aola/publications/thesis-ando/Related_Work.html.

23. Andreas Aschenbrenner, "Kulturarw^3—The Swedish Archive," Long-Term Preservation of Digital Material—Building an Archive to Preserve Digital Cultural Heritage from the Internet, Dec. 2001. Accessd Aug. 19, 2002, www.ifs.tuwien.ac.at/~aola/publications/thesis-ando/Kulturarw^3.html.

24. Andreas Aschenbrenner, "Die Deutsche Bibliotek," Long-Term Preservation, Dec. 2001. Accessed Aug. 19, 2002, www.ifs.tuwien.ac.at/~aola/publications/thesis-ando/Die_Deutsche.html.

25. Andreas Rauber, Andreas Aschenbrenner, Alfred Schmidt, "Motivation: The Invention of the Press: Archiving the Internet: Challenges, Projects, and the Austrian Perspective," Slide 4, talk at INST conference, "Knowledge Networking in Cultural Studies," in Reichenau, Austria, May 26, 2001. Accessed Aug. 19, 2002, www.ifs.tuwien.ac.at/~aola/publications/slides_reichenau_01.pdf (English version).

26. Brewster Kahle, "Ethics of Digital Librarianship," Feb. 1992. Accessed Apr. 26, 2002, www.archive.org/about/ethics_BK.php.

27. Lyman, "Archiving the World Wide Web."

28. Ibid.

29. Joan Allen-Hart, "California State Report," *State-by-State Report on Permanent Public Access to Electronic Government Information* (June 2003): 46. Accessed Aug. 17, 2003, www.ll.georgetown.edu/aallwash/State_report.pdf.

DIRECT: A Decentralized Image Retrieval System for the National STEM Digital Library

Jinshan Tang, Sridhar R. Avula, and Scott T. Acton

This paper describes the Decentralized Image Retrieval for Education (DIRECT) service for the National Science, Technology, Engineering, and Mathematics (STEM) Digital Library (NSDL). DIRECT augments NSDL by providing content-based image-retrieval (CBIR) functionality. CBIR allows the user to designate a query image so that the service can search the library for images of similar content. DIRECT matches images not by text metadata but by the color or texture of the image objects; the matching process does not depend on a match between the cataloger description and the user description. DIRECT is a peer-to-peer service built for decentralized digital libraries. The content-based image-retrieval service is available to all collections in NSDL without imposing new standards or protocols. With DIRECT, NSDL can support images that have yet to be cataloged or have incomplete metadata without adding overhead (expenditure of additional resources) to the collections.

In order to support education in science, technology, engineering, and mathematics (STEM), the National Science Foundation (NSF) has established the National Science, Technology, Engineering, and Mathematics Education Digital Library (NSDL) program. The aim of the program is to create a digital library that ultimately meets the "needs of students and teachers at all levels—kindergarten to undergraduate, graduate, and lifelong learning—in both individual and collaborative settings, as well as formal and informal modes."[1] The resultant program is an integrated and distributed information environment for accessing quality-assured digital resources from many sources. NSDL tasks include the following: (1) supply educators with easy access to high-quality, organized educational resources and the tools, interfaces, and other services to help them effectively use these resources; (2) support students with straightforward access to scientific data that encourages learning by direct experience; and (3) support a new integrated community by addressing the full spectrum of science, mathematics, engineering, and technology education.[2]

One challenging research problem that needs to be addressed in the NSDL research community is mining images in the digital library. The most common method of searching for images involves matching a text query to text metadata that have been entered manually. Text-based searches often prove ineffectual for the following reasons: (1) the image does not have associated metadata; (2) the metadata are incomplete; (3) the metadata include technical terms unknown to the user; and (4) the subjectivity of the metadata leads to a mismatch in terms used by the cataloger and user. To assist users in finding desired images from the expected tens of millions of images in NSDL, innovative services must be pursued to increase the usefulness of image retrieval for both teachers and students. Content-based image retrieval (CBIR) techniques can be designed to meet this aim.

The advantage of CBIR over exhaustive manual text annotation is obvious. As thousands of images are added to digital libraries each day (in such application areas as earth science and telemedicine), the ability to effectively catalog and manually describe each image is precluded. Since images constitute a vital vehicle for learning (from preschool to higher education), the effective maintenance of rapidly expanding image libraries is a critical goal. CBIR allows users to find images that have not been cataloged or have been partially cataloged. Of course, CBIR will not replace text-based searches entirely where text annotation exists, but CBIR will augment the searching capability with image-related information. Adding CBIR to digital libraries provides users with a new mechanism for accessing data that does not require knowledge of technical terms and jargon. Metadata often use technical terms to describe the content of the image; a text-based search would require the user to imitate the metadata vocabulary. CBIR also provides a mechanism for accessing images without requiring strong English skills and is particularly usable by young children and nonnative English speakers.

CBIR techniques have obtained considerable attention during the past decade. Examples include MIT's Photobook system, Columbia's VisualSEEK system, UCSD's Virage, and IBM's QBIC system.[3] In each case, the search and retrieval of images is based on a central database. These systems place the responsibility of extracting and indexing features on the collections and digital librarians. Centralized image-retrieval systems cannot meet the needs of NSDL as it is a heterogeneous collection of distributed resources.

In this paper, the authors describe the development of a decentralized image-retrieval system for education (DIRECT). With DIRECT, CBIR is a service that runs autonomously; the collections do not need to be modified

Jinshan Tang (jt6w@virginia.edu) is a Research Scientist, **Sridhar R. Avula** (sar4v@virginia.edu) is a Research Assistant, and **Scott T. Acton** (acton@virginia.edu) is the Walter N. Munster Associate Professor of Electrical and Computer Engineering, Department of Electrical and Computer Engineering, University of Virginia, Charlottesville.

to accommodate DIRECT. DIRECT applies a consistent CBIR mechanism to the entire system and gives CBIR functionality to collections without effort by the collection providers. Collections that do not have their own CBIR system or are not able to invest in a CBIR system (such as small collections or collections that do not contain a large proportion of images) gain greater distribution of their holdings via DIRECT's CBIR functionality. The DIRECT service finds collections in the library and requests the images using the same method that the portal uses to search and request images. DIRECT then performs feature extraction on the images and stores the features in its own database. A portal or client application can query DIRECT, and it uses the features of the desired image to search the library for images with the same features. DIRECT automatically monitors collections in order to keep the DIRECT index up to date.

Another advantage of DIRECT is that DIRECT supports not only global feature-based image retrieval, but local feature image retrieval. Local feature image retrieval is a segment-based image retrieval approach. In this approach, the critical process is *image segmentation*—the task of locating the constituent homogenous regions of the image. In past CBIR systems, segmentation has been performed manually or by computer-assisted methods. We are describing a fully automated method based on *k*-means classification. In our approach, the given image is segmented into *dominant components* and the properties of such components as color and texture are used to retrieve similar images containing similar segments.

Background and Related Work

NSDL

Launched in December 2002, NSDL is a comprehensive source and public network that provides educational materials and a lifelong learning environment for science, technology, engineering, and mathematics education at all levels. NSDL is attaining immediate utility at the pre-college, undergraduate, and graduate levels, as well as providing a mechanism for continuing education.

The NSDL vision is to create a "digital learning place."[4] Realizing this vision entails an environment that is more than just a repository of digital resources, but rather an environment that "encourages and supports users in their efforts to create, discover, explore, and interact with digital resources."[5] NSDL is "a *center of innovation* in digital libraries as applied to education, and a *community center* for groups focused on digital library-enabled science education."[6]

This paper focuses on the DIRECT system for CBIR. Currently, DIRECT has been implemented as a stand-alone prototype using a subset of the available NSDL collections. DIRECT is expected to be fully integrated into NSDL by late 2004.

Peer-to-Peer Network Systems

A peer-to-peer network system provides a new way of utilizing distributed resources. In a peer-to-peer system, all participants in the system, such as image collections, education collections, searching services, and image retrieval services, are regarded as peers. Control of the system is distributed throughout the peers. Participants are peers in a peer-to-peer system. One of the benefits of such a peer-to-peer system is that it utilizes the resources of all the peers involved. Unlike an Internet application that must invest more resources to support more and more users, a peer-to-peer application gains resources and capabilities with each participant, making the system scalable and inexpensive for thousands to hundreds of thousands of participants.

Although peer-to-peer systems were used in the earliest incarnations of the Internet, these systems have been overshadowed by the centralized Internet applications of the Web. Lately, personal computers have reached a level of ubiquity that promises significant potential for peer-to-peer systems.

Several existing systems use peer-to-peer services, such as Napster, Gnutella, and Freenet.[7] Napster transfers files in a peer-to-peer manner while maintaining a centralized index. When using the system, a client will query the central system and receive a list of peers that have the desired files. The client then will contact the peer to download the file. When peers join, they notify the central system, and the central system updates the index.[8] This type of system is very fast and efficacious and avoids the costs of a central library but incurs the cost of a centralized index and imposes updating the index on the peers.

Gnutella and Freenet are completely decentralized.[9] A client can literally query each peer in the system when searching for a file. This query is accomplished by connecting to several neighboring peers and sending the query to each. These peers will, in turn, relay the query to each of their neighboring peers, and the query will extend in a like manner until a timeout is encountered. If the file is found, the response is passed back through the network of neighbors, and the file is transferred from peer to peer. These services are completely decentralized and distribute the cost of the system among the peers, but the search mechanism is extremely slow and usually will not search the entire system before the user-invoked time limit has been reached.

Gnutella decreases the response time and increases scalability. This system introduces peers that query their neighbors and automatically create indexes. This indexing peer is a reflector node and contains the index for ten

or twenty regular nodes. These reflector nodes, when used as the backbone of the Gnutella network, can reduce the number of nodes that need to be searched and improve performance.[10] However, the Gnutella performance is inefficient in comparison to centralized systems.[11]

DIRECT

Image Retrieval Service in NSDL

As pointed out in the beginning of this article, centralized image-retrieval systems, which place the responsibility of extracting and indexing features on the collections and digital librarians, cannot meet the needs of NSDL as it is a heterogeneous collection of distributed resources. The query of distributed, heterogeneous collections of images requires a CBIR service. DIRECT is such a system being developed that satisfies this requirement.

In DIRECT, CBIR will be a service that will run autonomously; the collections will not need to be modified to accommodate DIRECT. DIRECT will apply a consistent CBIR mechanism to the entire system and gives CBIR functionality to collections without effort by the collection providers. Collections that do not have their own CBIR system or are not able to invest in a CBIR system (such as small collections or collections that do not contain a large proportion of images) will gain greater distribution of their holdings via DIRECT's CBIR functionality.

The DIRECT service will find collections in the library and request the images using the same method that the portal uses to search and request images. DIRECT will perform feature extraction on the images and store the features in its own database. A portal or client application can query DIRECT, and it will use the features of the desired image to search its database for images with the same features. DIRECT will automatically monitor collections in order to keep its database up-to-date. The structure of DIRECT service is shown in figure 1.

Image Features Used in DIRECT

DIRECT supports both global feature-based image retrieval and local feature-based image retrieval. For global feature-based image retrieval, global color histograms and texture measurements are used to conduct an image query within the digital libraries. For local feature-based image retrieval, there is a segment-specific approach to feature extraction and content description for digital images.[12] The image query and retrieval engine is based on object features extracted by segmentation, and image retrieval occurs by matching the constituent image segments.

In global feature-based image retrieval, global color histograms and global texture measures are used to perform image retrieval. For each image in the database, DIRECT records a quantized (sixteen-element) histogram for each color band. The histogram is a function that records the frequency of occurrence of each color. Although perceptually motivated color spaces exist, such as l*a*b*, DIRECT utilizes the RGB color space to achieve enhanced processing versatility. If DIRECT required an l*a*b* representation for color, the majority of images in the distributed database could not meet the specification, due to lack of information about the *true* color content. The color histograms used in this CBIR system are quantized using a preset table that exploits the color discrimination of the human visual system.[13] In order to expedite searching, DIRECT constructs a histogram feature library in the centralized database. As each band of each image is represented by only a sixteen-element histogram, for each image, only forty-eight integers are needed to be stored in the centralized library. After construction of the histogram feature database, a distance measure is used to compare the similarity between the histogram of the query image and the image in the database. In DIRECT, histogram intersection is used due to its simplicity and robustness.[14] Figures 2 (a) and (a*) show retrieved images when only color is employed.

In terms of texture extraction, DIRECT captures the granularity (periodicity) and the directionality of two-dimensional patterns in the digital library images so that users can retrieve images containing similar patterns (textures). The texture retrieval method in DIRECT is based on the wavelet.[15] Figures 2 (b) and (b*) show retrieved images using only global texture features in DIRECT.

Besides allowing only one aspect of the image (color or texture) to be utilized in image retrieval, DIRECT also provides image retrieval combining texture and color. In order to combine two features to perform image retrieval,

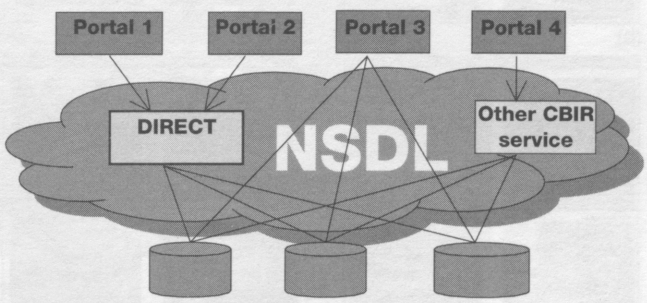


Figure 1. The DIRECT service is available to all portals. Here, portals 1 and 2 use the service. Portal 3 does not and searches the collections directly. Portal 4 has implemented its own CBIR service. Both CBIR services search all collections.

a matching functional is required. The matching functional determines the relevance of match between the query image specified by the student or teacher and an image in the distributed digital library. The first step in constructing the matching functional is normalization of the color and texture features for each image. Given a query image Q and a database image D, the matching functional is computed using the following:

$$D(Q, D) = w_1\overline{D}_H(Q, D) + w_2\overline{D}_T(Q, D)$$

$\overline{D}_H(Q, D)$ represents the dissimilarity in color histograms. $\overline{D}_T(Q, D)$ represents the normalized dissimilarity of texture.

The parameters w_1, w_2 determine the relative weighting for the image features. In the initial DIRECT system, we allow only discrete weights of 0, 1, 2, and 3. A weight of zero would exclude a particular feature (for example, texture) from the query. A weighting using 1, 2, and 3 would assign a prioritization to features in the search process. When all weights are equal (=1), the characteristics of color and texture are considered equally in the image search. Figures 2 (c) and (c*) are the retrieved images when color and texture are combined.

For local feature-based image retrieval, the image is segmented into different regions and the image retrieval is based on the segments rather than global features. Segmentation is the process of partitioning the image into the constituent homogeneous regions, and a segment is one of the homogeneous regions. In DIRECT, a *k*-means image segmentation is used.[16] Figure 3 shows an example of DIRECT's image segmentation algorithm. After the segmentation, DIRECT retains the five largest segments (called the *dominant segments*) and uses them to perform image retrieval based on the individual color and texture features of the segments.

Local feature image retrieval using color and texture features is similar to image retrieval based on global color histograms and global texture measures, except that features are extracted from segments, not globally from the images. Color and texture databases are generated using the local color and texture features. The color data are extracted as follows: (1) for each image, DIRECT obtains five dominant segments; (2) for a given dominant segment, the color content is extracted in the form of quantized RGB histograms, as with global image feature extraction; (3) the construction of the local texture feature database is also produced in a manner similar to the construction of the global texture feature database: to obtain the texture feature, a square is inscribed in the segment,

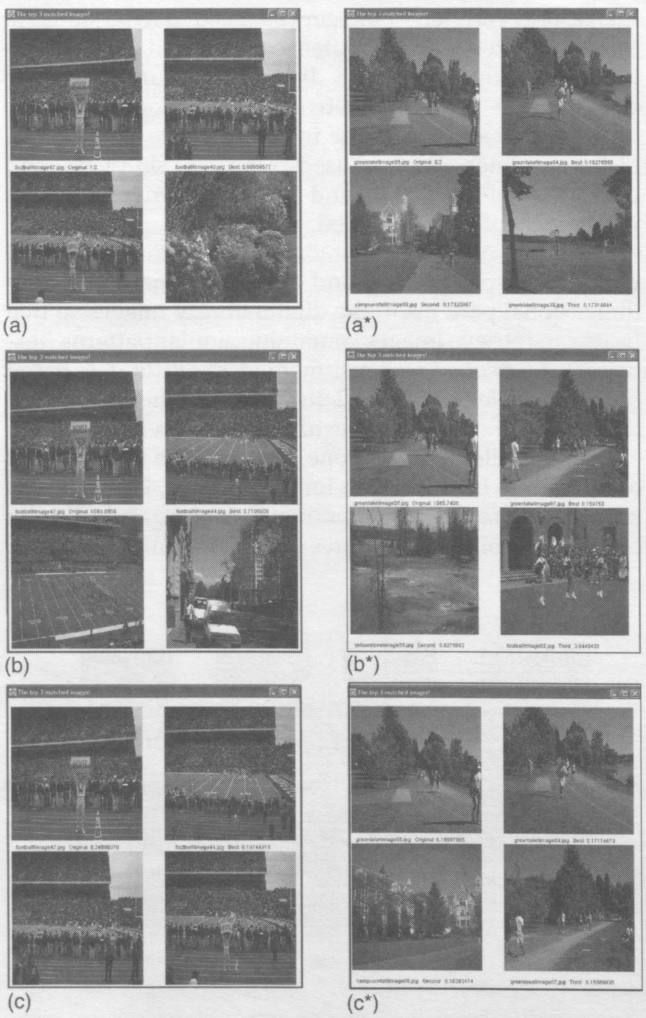


Figure 2. Image retrieval using different global features: images retrieved using only color, (a) and (a*); images retrieved using only texture, (b) and (b*); images retrieved using color and texture features simultaneously, (c) and (c*).

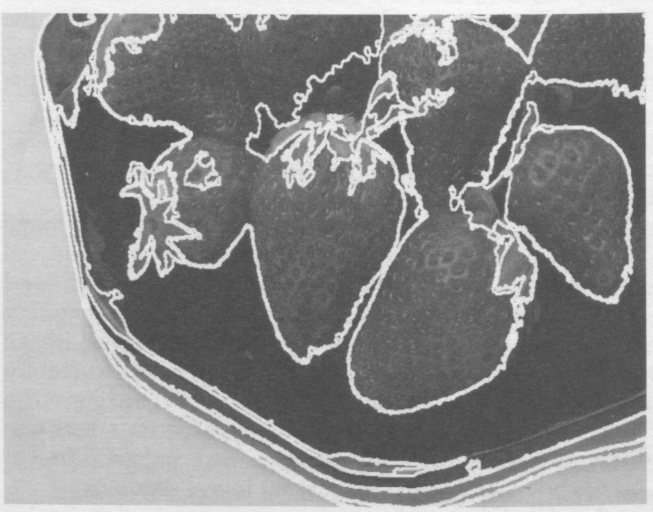


Figure 3. Image segmentation algorithm in DIRECT.

and an image block is extracted; and (4) the texture feature for the image block obtained by a wavelet method is used as the texture feature of the segment.[17]

When performing retrieval, the query image first is segmented and the dominant components found are highlighted so that the user can select one component. Once a dominant component is chosen, it is treated as an image and the retrieval mechanism based on global features can be used to perform the retrieval. The major difference is that the feature match is evaluated between two segments, not two images.

Figure 4 shows an example of image retrieval based on segments. Here, in the query image, the component representing the body of water was used to retrieve the images from the database. Figure 5 shows the retrieval results using the global features. In both cases, the same weights (=1) were used for texture and color.

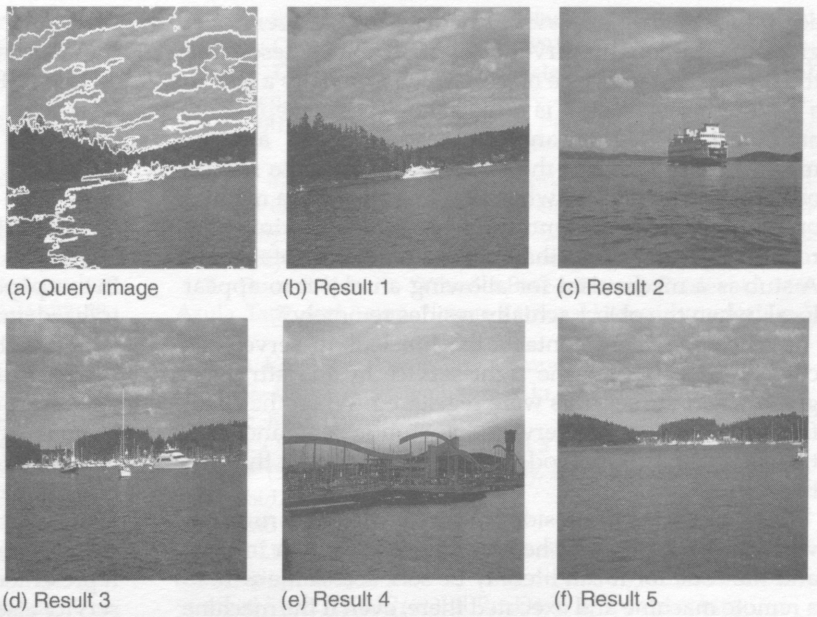


Figure 4. Images retrieved using only the water segment as a query image (a).

Implementation of Decentralized Peer-to-Peer File Retrieval in DIRECT

A major component of the DIRECT service is supporting decentralized peer-to-peer file retrieval. DIRECT will work as a service in the peer-to-peer environment of the NSDL collections to retrieve images from image collections. All participants in the system such as image collections and the DIRECT searching service will be regarded as peers, and control of the system will be distributed throughout these peers.

DIRECT creates a network of services by creating a brokered peer-to-peer system. A broker is an entity that connects service providers with service users or clients in a way that prevents either party from needing to know about the other *a priori*.[18] The service provider informs the broker that it provides a certain type of service. The client queries the broker for a certain type of service and, if there is a match, the broker connects the client with the service provider. In the DIRECT approach, the portals are clients; the collections are service providers, and such entities as the index and the content-based image-retrieval portion of DIRECT are both clients and service providers.

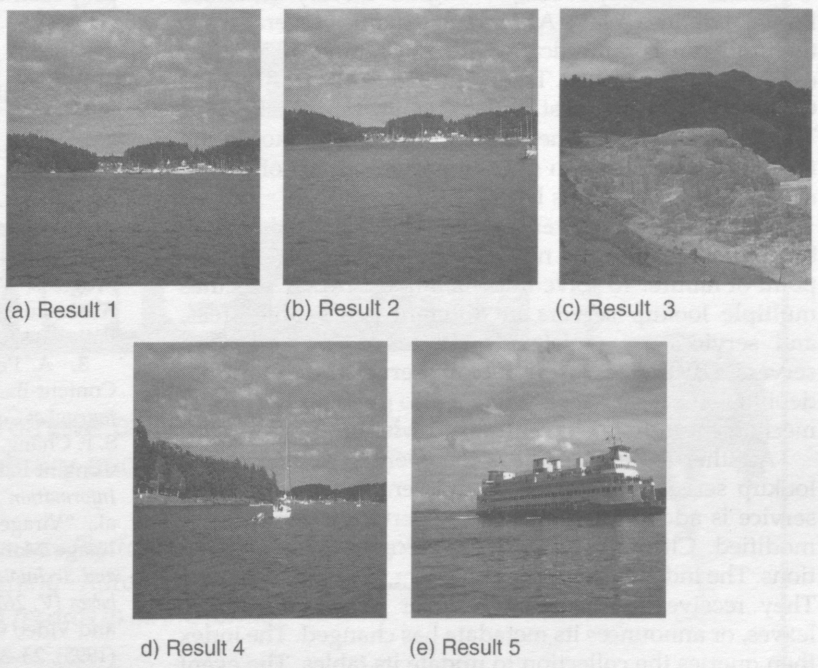


Figure 5. Images retrieved using the global texture and color features.

Currently, DIRECT adopts Java's Jini to implement a broker in a peer-to-peer system.[19] Jini provides an application known as a *lookup server* that acts as a broker for the system. The lookup server comes with a predefined application program interface (API) for finding and

accessing it. When a service provider registers a service, it gives the Jini lookup server a set of attributes describing the service and a service object, which is usually a *stub* to a remote object. A stub is a special object that represents an object located in another process or on another machine. The stub has the same methods as the remote object it represents and when the program calls a method on the stub, the stub communicates across the wire to the remote object and calls that method on the remote object. A stub is a mechanism for allowing an object to appear local, when the object actually resides remotely.[20]

When the client contacts the Jini lookup server, the client can search for the right service by the attributes given when the services were registered. When the client finds the appropriate service, the stub is sent, and then the client can call methods on the remote object through the stub.

The power of Jini resides in the "write once, run anywhere" idea of Java.[21] The stub object is written in Java, and the code for it can literally be sent across the wire to a remote machine and executed there, even if the machine has never seen the code previously. Therefore, a portal does not need to know how a collection implements a search protocol to query it. For instance, University of California–Berkeley's Simple Digital Library Interface Protocol defines a Java API but also defines several ways the stub can communicate with the remote object (for example, Hypertext Transfer Protocol [HTTP] or Common Object Request Broker Architecture [CORBA]).[22] This ability to allow the service provider to choose the form of communication is an important aspect of such a decentralized service as DIRECT.

The Jini lookup servers enable a decentralized system, but the broker itself is not only centralized but a single point of failure. To solve this dilemma, DIRECT executes multiple lookup servers in different geographic areas, and services are registered with each of the lookup servers. DIRECT allows a lookup server to fail and not debilitate the entire library, and it also prevents the broker mechanism from being centralized with one system.

Another feature of Jini is the event system. The Jini lookup server will broadcast an event whenever a new service is added, or if an existing service is removed or modified. Clients can register to receive these notifications. The indexes use this mechanism to keep up-to-date. They receive notification whenever a collection joins, leaves, or announces its metadata has changed. The index then queries the collection to update its tables. The event mechanism allows a client to react to a service provider without having the service provider know about the client.

Jini provides the functionality to implement DIRECT as a decentralized peer-to-peer service. Jini supplies a broker so that DIRECT can find collections dynamically and allows mobile software that can communicate with each collection using the collection's desired communication mechanism. Last, Jini provides an event mechanism so that DIRECT can be notified when it needs to update its index.

Conclusion

In this paper the authors have described DIRECT, a decentralized image-retrieval system for the NSDL. DIRECT uses distributed database technology to exchange digital images in a peer-to-peer environment. With a flexible, sustainable, decentralized approach, the existing NSDL collections and services do not have to be significantly modified to utilize DIRECT. Most importantly, DIRECT gives the ability to search NSDL images by content. Such a content-based image-retrieval tool will allow access to the constituent collections of the library by such underrepresented groups as nonnative English speakers. The service also will allow search and retrieval for images that have not been annotated manually. As the number of digital images in the NSDL rapidly increases, automatic segmentation and description of the digital image content, as proposed with DIRECT, becomes a critical need.

References

1. L. Zia, National Science, Technology, Engineering, and Mathematics Education Digital Library (NSDL): Program Solicitation, NSF-02-054, 2002. Accessed Dec. 1, 2002, www.nsf.gov/pubs/2002/nsf02054/nsf02054.html#TOC.
2. C. Manduca, F. McMatin, and D. Mogk, Pathways to Progress: Vision and Plans for Developing the NSDL, 2001. Accessed Jan., 21, 2004, http://doclib.comm.nsdlib.org/PathwaysToProgress.pdf.
3. A. Pentland, R. W. Picard, and S. Sclaroff, "Photobook: Content-Based Manipulation of Image Databases," *International Journal of Computer Vision* 18, no. 3 (1996): 233–54; J. R. Smith and S. F. Chang, "Querying by Color Regions Using the VisualSEEK Content-based Visual Query System," *Intelligent Multimedia Information Retrieval*, ed. Mark T. Maybury (1996); J. R. Bach et al., "Virage Image Search Engine: An Open Framework for Image Management," *Symposium on Electronic Imaging: Science and Technology—Storage and Retrieval for Image and Video Databases IV*, 2670 (1996), 76–87; M. Flickner et al.,"Query by Image and Video Content: The QBIC system," *IEEE Computer* 28, no. 9 (1995), 23–33.
4. Manduca, McMatin, and Mogk, Pathways to Progress.
5. Ibid.
6. National Science Foundation, "NSDL: The National Science Digital Library" (initial version). Accessed Dec. 5, 2002, http://nsdl.org/render.userLayoutRootNode.uP.
7. Napster, How Napster Works. Accessed Dec. 10, 2002, http://wiki.cs.uiuc.edu/cs427/Napster:+How+it+works; DSS Group, Gnutella: To the Bandwidth Barrier and Beyond, 2000.

Accessed June 10, 2002, http://dss.clip2.com; Freenet, The Free Network Project, 2001. Accessed May 7, 2002, http://freenetproject.org.

8. Napster, How Napster Works.

9. DSS Group, Gnutella: To the Bandwidth Barrier and Beyond; Freenet, The Free Network Project.

10. DSS Group, Gnutella: To the Bandwidth Barrier and Beyond.

11. M. Ripeanu, "Peer-to-Peer Architecture Case Study: Gnutella Network," *Proceedings of the First (2001) International Conference on Peer-to-Peer Computing*, eds. Ross Lee Graham and Nahid Shahmehri, Linkoping University, Sweden, Aug. 27–28, 2001 (Piscataway, N.J. : IEEE, 2002), 99–100.

12. Sridhar Avula, Jinshan Tang, and Scott Acton, "Image Retrieval Using Segmentation," *Proceedings of the 2003 IEEE Systems and Information Engineering Design Symposium*, eds. Matthew H, Jones, Barbara E. Tawney, K. Preston White Jr., University of Virginia, Charlottesville, Apr. 24–25, 2003 (Charlottesvile, Va.: Department of Systems and Information Engineering, Univ. of Virginia, 2003), 289–94.

13. Y. Ma and B. S. Manjunath, "NETRA: A Toolbox for Navigating Large Image Databases," *Proceedings of the International Conference on Image Processing*, Santa Barbara, California, Oct. 26–29, 1997, vol. 1 (New York: Institute of Electrical and Electronics Engineers, 1997), 26–29.

14. M. Swain and D. Ballard, "Colour Indexing," *International Journal of Computer Vision* 7, no. 1 (1991): 11–32.

15. M. Do and M. Vetterli, "Wavelet-Based Texture Retrieval Using Generalized Gaussian Density and Kullback-Leibler Distance," *IEEE Transactions on Image Processing* 11, no. 2 (2002): 146–58.

16. J. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1 (Berkeley: Univ. of California Pr., 1967): 281–97.

17. Do and Vetterli, "Wavelet-Based Texture Retrieval," 146; Avula, Tang, and Acton, "Image Retrieval Using Segmentation," 289.

18. B. Knighten, Peer-to-Peer Computing, 2002. Accessed Dec. 10, 2002, http://cnscenter.future.co.kr/resource/hot-topic/p2p/PtP_IDF_Rev1.11-web.pdf.

19. S. Oaks and H. Wong, *Jini in a Nutshell: A Desktop Quick Reference* (Sebastopol, Calif.: O'Reilly, 2000).

20. Ibid., 102.

21. Ibid., 111.

22. SDLIP: The Simple Digital Library Interoperability Protocol, 1999. Accessed Dec. 1, 2002, www-diglib.stanford.edu/~testbed/doc2/SDLIP/SDLIPProjDesc/sdlipProject.htm.

Looking for **Marc**?

BookWhere™ *Online*



Simultaneously search and retrieve bibliographic records online from thousands of library and media databases.

No software to install - No firewall issues - Platform independent - Simple navigation
Powerful, simultaneous searching - Automatically updates databases
Easily export records in MARC format

Toll Free: 800.661.7274 **FREE Product Trial** www.webclarity.info

WebClarity SOFTWARE INC

Looking Beyond . . .

Designing for Experts: How Scholars Approach an Academic Library Web Site

Thura Mack, Maribeth Manoff, Tamara J. Miller, and Anthony D. Smith

This study examines the use of an academic library Web site by experienced researchers and active scholars. It is part of a larger effort to understand how experienced users approach online information resources and how fully the library's Web site meets their needs. Subjects were asked to complete eight online tasks, beginning each task at the library's home page. Data were gathered by means of screen- and audio-capture software, and human observers as study participants worked through sets of tasks. Results were analyzed in terms of the experience and expertise of the participants, success rate, and the first click indicating the chosen path to the information requested. Subjects had high success rates for most tasks. Searching for information about journals and locating journal articles proved to be the most difficult tasks to successfully complete. Analysis of session recordings revealed some traits of expert users that can be used to improve Web site design, and indicated a correlation between success in searching and the double-expertise of subject knowledge combined with frequent use of the library's Web site.

This research stems from an interest in the online information-seeking behaviors of experienced researchers and the extent to which a customizable portal might improve their ability to locate and manage information sources. By analyzing the information-seeking habits of active scholars, the authors hope to identify online tools that would improve the efficiency of scholars' information retrieval. A broad range of solutions are under consideration, from the design elements in a traditional Web site to the feasibility of employing advanced technologies to create a personalized information environment. One example of such an advanced technology would be an *expert system* powered by artificial intelligence and responsive to natural language, that could assist researchers in discovering and using library resources. The ultimate aim is to explore the usefulness of a scholar's portal that not only mirrors a research environment, but also establishes a customizable virtual workspace allowing the user to organize both resources and ideas.

The authors chose to begin their study with a formal process of usability testing, employing the results in an ongoing Web-design process. In June 2001, a team of four librarians and a graduate assistant collaborated to develop a test of the University of Tennessee (UT) Libraries' Web site as it was being redesigned. One member of the team has ongoing responsibility for the library Web site design. The team's first objectives were to determine how to design and implement a usability test, to gather feedback on the new site, and to plan, create, and document a user-centered design process for all future Web sites.

The study comprised two parts: (1) a set of Web-usability tasks was developed to test the ease of information discovery and retrieval; and (2) a structured, open-ended interview was then conducted to determine user preferences, expectations, and current online searching practices.

This article will discuss the Web-usability portion of the study. The interview data will be reported separately.

Data Collection Methods

The study employed thirteen subjects who participated in individual Web-usability sessions lasting between forty and ninety minutes. Volunteers were recruited by contacting the graduate student association and library faculty representatives, giving an explanation of the study and what participation would involve. The research team offered two incentives: that the subjects would be assisting the library in better serving the needs of the research community and that participants would be entered in a drawing for a gift certificate for a dinner for two.

The team sought to attract a purposeful sample, rather than the randomly drawn samples typical of quantitative research. Purposeful sampling allows for selection of a limited number of information-rich cases for in-depth observation.[1] Within purposeful sampling, an *intensity* sample of cases was utilized that manifests several characteristics intensely, but not extremely.[2] The team wanted to observe and identify the Web-use efficiency and effectiveness of faculty and graduate students actively engaged in research. An attempt was made to recruit mainly from the science and humanities disciplines, given a working hypothesis that those two disciplines would have different information needs and use the Web site in substantially different ways. However, the team eventually decided that broad participation from all three of the traditional discipline areas would be beneficial. All but one of the participants were conducting a current research project (see table 1).

Thura Mack (tmack@utk.edu) is Reference and Training Librarian, **Maribeth Manoff** (mmanoff@utk.edu) is Coordinator for Networked Services Integration, and **Anthony D. Smith** (adsmith1@utk.edu) is Digital Initiatives Coordinator at the University of Tennessee Libraries, Knoxville. **Tamara Miller** (tamaramiller@montana.edu) is Associate Dean of Libraries at Montana State University, Bozeman.

Sample-size determination for qualitative research is not driven by the statistical guidelines familiar in quantitative work. The size of the sample is less important than the rationale for including participants and the scope of the study.[3] Rather than trying to reduce statistical sampling error, the team made an effort to avoid *discovery failure,* or the failure to discover an attitude, perception, or behavior within the scope of this study. Usability sessions were continued until the information gathered became repetitive. Thirteen subjects appears to be an appropriate study size based on previous observations of usability-study effectiveness with similar subjects.[4]

All but one of the users in this study can be classified as expert in their respective discipline areas or knowledge domains. They are active researchers at an advanced level. More than three-fourths of the subjects reported experience with the older version of the library Web site, with 60 percent of subjects using the site for an average of three or more hours per week. Thus, a majority of subjects could be categorized as *double-experts* with both high levels of domain knowledge and substantial experience with Web-based library content and services (see table 2).

For this study, the usability sessions were scheduled to coincide with the launch of a new library Web site design. Although many of the participants indicated heavy use of the library Web site, none had experience with the new Web site design. The main page and several of the first-level pages were completely redesigned. Much of the original functionality and many lower-level pages remained unchanged after the launch of the new Web site. Frequent users of the library Web pages were likely to encounter familiar pages and resources within a few clicks from the front page.

Table 1. Participants

Discipline	**%**
Science	62
Humanities	23
Social science	15
Status	
Faculty	46
Graduate student	54
Current research	
Yes	92
No	8

All of the usability sessions were conducted in one of the library's bibliographic instruction rooms, normally used for classroom teaching. Instruction rooms functioned very well for the study as they both provided sanctuary from the busy areas of the library and housed the needed computer equipment. The rooms are equipped with an instructor's workstation and a projector that displays the instructor's monitor on a wall screen. Each subject was seated at the instructor's computer, allowing the research group to closely observe the participant's searches on the wall screen without the need to hover around the participant. It is believed that this contributed to a more relaxed environment for participants.

The workstations used for bibliographic instruction already had been configured to provide access to all of the library's information resources. This permitted the study team to organize the room and conduct the testing with minimal workstation preparation. It was necessary to add one piece of software and a microphone to capture each test session. In order to gather as much information as possible, a decision was made to record the screen activities and the comments of the participant and moderator. TechSmith's Camtasia software was used to record each session. The software proved to be economical and provided a permanent record of the screen activities, synchronized with the audio, for all of the sessions. Camtasia is relatively easy to install and requires a minimal hardware configuration to operate effectively. An Intel-based, Windows NT workstation was used to conduct the study, and this worked superbly. Figure 1 gives the recommended system requirements as well as the actual system used in this study. Each session was captured as a compressed Audio Video Interleave (AVI) file and then archived to CD-ROM.

Table 2. Use of the library Web site

Use library Web site	**%**
Yes	77
No	23
How many hours per week?	
3 or more hours	60
1–2 hours	30
Less than 1 hour	10
Favorite online uses	
Databases	50
Journal articles	30
Other	20

In addition to the participant, each session included a moderator and an observer. The moderator was charged with leading the participant through the required steps. The observer's role was to note incidental information about the performance of each task that later could be compared with the recording. The observer also contributed by asking questions and interjecting ideas during the discussion portion of the session.

Because the participants would be recorded, it was necessary to obtain written consent from each participant as part of the requirements for using human subjects at UT. The participants were also asked to complete a brief questionnaire designed to gather basic demographic data, level of online experience, and discipline areas of research activity. The authors recognized that the test environment might create anxiety among the participants, and therefore a conscious effort was made to make the subjects feel relaxed. The moderator and observer spent a few minutes at the beginning of each session introducing themselves and learning a little about the participants and their academic interests. Bottled water was provided for each participant. A written script guided the entire session to ensure consistency across sessions. The moderator's script included reassurance that the session was a test of the library's Web site and not a test of the participant. These extra efforts were intended to make the subjects feel more comfortable about the questioning and thereby provide a more realistic reflection of the information-seeking process.

The part of the study reported here consisted of eight tasks meant to discover the difficulty level of locating and retrieving information resources from the library's Web site. The tasks were designed to include the major information resources and services typically found in an academic library. Similar instruments used at Appalachian State University, Massachusetts Institute of Technology, the University of Arizona, and others were reviewed before developing the usability tasks.[5] An attempt was made to carefully construct tasks that would reflect a real-world situation in which the information seeker may not have a full citation. For example, the first question asks the participant if the library owns a copy of *The Color Purple* by Alice Walker. It does not state that the item is a monograph or that most reference librarians would first check the online catalog to answer this type of question. Constructing the tasks in this fashion allowed the authors to observe how scholars might approach a typical, modern academic library with its diverse set of information resources.

The tasks were pretested by members of the library staff and revised to improve the wording and remove ambiguity. The entire test session was then pretested with a graduate student to give the team experience with the script and the timing of a session, and to test the recording routine.

For each task, the moderator read the question aloud and then placed a printed copy of the question near the keyboard so the participant could refer to it while attempting the task. A time limit of three minutes was placed on each task as a way to measure the effective completion of that task. After three minutes, the moderator stopped the participant and moved on to the next task.

Analysis

The study team began the task analysis by collating data. Using the recordings, along with the notes taken by observers, the team calculated the success rate and the percentage of participants who were able to find the information within the three-minute time frame. The authors noted the path each user took in their search for the information requested in that particular task. The path information, a record of each page the participant visited, and the terms they typed into any search box were transferred to a worksheet, arranged by participant. Arranging the data in this way, by task and by participant, made it easier to detect patterns at a single glance.

To better assess information-seeking behavior, the team took a closer look at success rate and *first clicks* (the first page each participant visited in their information quest). The team previously had determined the best,

TechSmith system requirements	System used
Microsoft Windows 95 OSR2, 98, Me, NT 4.0, 2000, XP, or later version	Microsoft Windows NT 4.0
90 MHz processor (400 MHz recommended)	450 MHz processor
16 MB RAM (64 MB recommended)	384 MB RAM
Windows-compatible sound card and microphone	Windows-compatible sound card and microphone
12 MB of hard-disk space for program installation	30 GB hard disk space

Figure 1. Camtasia software information

Table 3. Summary of task analysis

Task	% Success rate	% First click on "best" route	% First click on any direct route
1. Is *The Color Purple* by Alice Walker available at the UT Libraries?	85	62	NA
2. Is *Understanding Physics* by Isaac Asimov available at another library in the region?	69	15	38
3. Where could you request an item on interlibrary loan?	92	62	15
4. In what formats does the UT Libraries receive the magazine *Scientific American*?	54	62	NA
5. Search for information on medieval universities or human cloning.	92	NA	NA
6. Find K. Wesley Berry's article "The Lay of the Land in Cormac McCarthy's "The Orchard Keeper and Child of God" or Peter M. Cox's article "Acceleration of Global Warming Due to Carbon-Cycle Feedbacks in a Coupled Climate Model."	62	38	NA
7. Where can you find a definition for the word "gaff"?	92	8	54
8. Where can you submit a reference question online?	92	8	54

most direct route for successful completion of each task, starting from the library's home page. For each actual path taken, the team asked the following questions (see table 3): Was the first page the user visited along this best route? If not, was the first click on any path that would lead directly to the information, or would the participants have to retrace their steps in order to get where they needed to go?

The first task was to discover the availability of a book in the library's collections. The most direct route to this information is from the home page to the catalog, where a number of different searches would yield the correct results. Eight of the thirteen participants went directly to the catalog and three more eventually got to the OPAC, for a success rate of 85 percent.

The second task asked the participant to determine whether a book was available in another library in the region. Again, there is a direct link on the library's home page to a union library catalog called Kudzu that contains the holdings of fifteen southeastern academic libraries. Only two of the participants selected the preferred route, directly linking to Kudzu from the home page. Five others first clicked on the Other Catalogs link. This page has a description of Kudzu along with a link so that participants were able to find the information without backtracking. In addition to these seven, two more subjects eventually arrived at the Kudzu catalog and found the book in question.

The third task was to locate the pages for submitting an interlibrary loan request. There is a direct link from the home page to the online interlibrary loan form, and eight subjects used this link to reach the form on their first click. Two more subjects linked to Forms, where they could choose Interlibrary Loan from the list. Two more participants were successful by a more indirect route, for an overall success rate of 92 percent.

The fourth task was to find the formats available for a specific journal title. The library catalog contains three records describing various formats of the journal. The best route is to choose the catalog, then conduct a browse-type search using the periodical title index. Eight users went directly to the catalog but chose keyword searches, which retrieved lengthy lists of volume-holdings statements. The results were confusing to users, as the information needed did not appear on the first screen. Only seven participants, or 54 percent, were able to finally locate the journal-title formats.

For the fifth task, participants were asked to find any information on a specific subject. There was no predetermined *best* path for this task. Interestingly, the Subject Guides link from the home page, which four people chose, was not a particularly fruitful first click, as the subject areas covered on these pages are too broad for the task given. Five others went first to Databases, two went to Library Catalogs, and two more went to Internet Search Engines. Only one participant became lost and was not able to find any information on one of the topics given.

The sixth task, which instructed the participants to locate a specific journal article, was one of the most complicated. Users were given the author and title of the article, but not the journal in which it appeared. This task

User Interface Consulting

- Library websites
- OPACS
- Children's Catalogs
- Staff Services

Can patrons navigate the wealth of resources your library offers?

For a **FREE** evaluation of your library home page, contact:

Ezra Schwartz
(773) 256-1418
ezra@artandtech.com
http://www.artandtech.com

required people to link to Databases, locate an appropriate database, and search within the database to locate the citation. Five participants went to Databases on their first click, and three more eventually discovered the correct path to an article, for a success rate of 62 percent.

The seventh task was to find a word definition online. In this case, the Subject Guides page was considered the best first click, as it leads to a list of links that includes a number of online dictionaries. Only one of the participants went first to Subject Guides, and seven went to Internet Search Engines. Several of the latter went to Google.com, where they either located an online dictionary or found the definition directly. The relatively high success rate of those who sought an answer from outside the library Web site and the inability of users to locate the library dictionaries from the home page indicate both a failure of the library Web design and a growing tendency to seek simple answers using Internet search engines.

The last question was where to submit a reference question online. A new service called AskUs.Now had been launched at about the same time as the new Web site and was available on the library's home page. Choosing AskUs.Now (which one user did on the first click) brings up a page offering chat reference during certain hours, along with links to the e-mail reference service, which was the target page. Several other links from the home page are less direct paths to the e-mail reference service, and seven users chose one of these alternate links as a first click. In the end, twelve of the thirteen participants found the online reference form.

Discussion

The participants had relatively high success rates regardless of discipline or status. The three participants who completed every task successfully within the three-minute time limit were all native English-speaking graduate students who were actively engaged in research projects and who reported making regular use of the library Web pages—qualifying them as double-experts in terms of both their subject area and online library use. Successful searching has been characterized as requiring this double-expertise in general Internet searching studies.[6]

There is further evidence of a strong correlation between frequent access to the library Web site and success in locating information and services online. Of the two participants who had the least success, one had never used the library online and the other reported usage at the lower end of the scale. Even so, some faculty members with equally low library Web page experience had higher success rates, perhaps because of their strong discipline knowledge. One computer science faculty member, who reported no regular use of the library Web page, completed all but one of the tasks successfully. His research area is indexing and information retrieval, which he felt allowed him to use unfamiliar online information systems with ease. The relationship between successful online searching and (1) online experience; (2) familiarity with resources within a discipline; or (3) deep knowledge within a domain warrants further examination (see table 4).

Although each task started at the library home page, participants were not required to stay within the library site as they searched. They sometimes chose to use Internet search engines, primarily for locating general subject information and to obtain the word definition. Seven participants went to Google to find a dictionary or a definition and one went directly to dictionary.com, which she said she used frequently. Four participants left the library Web site and used either Google or Yahoo! to gather subject information for task five. These participants indicated that they preferred to use an Internet search engine to gather broad information on a topic before beginning a more narrow and specific search for information, particularly when a topic was new to them.

One of the most challenging tasks given to the participants was locating a specific article or its citation with the author and title given, but without knowing the journal title. The key to success was selecting the correct database from more than one hundred offered on the Web site.

When participants could comfortably stay within their area of subject expertise, this task was accomplished with ease. However, when they were not familiar with the literature and tools in a field, there was little success in guessing at an appropriate database in which to search. The need to offer better subject assistance in searching the journal literature was evident here. One possibility for achieving this goal within a scholar's portal would be a federated searching function, which would allow the user to search across a number of databases within an unfamiliar subject area.

University faculty and graduate student users—with their attendant levels of experience and expertise—expect to find the information they are seeking. If the desired information, or a clear path to it, does not appear near the top of a display, they are unlikely to read the full page, much less scroll down through search results to find it. In several instances, participants reached the right screen, or made a proper search, but did not read far enough down to realize they had succeeded. Their assumption was that the needed information should appear near the top, unless they had made an error in search protocol or in database selection. Rather than scroll through a list, they would simply initiate a new search in hopes of more specific results. This evidence provides a challenge to further refinements in the design of the Web site. How much information, and how many paths to further information, can appear on one page and still be effective?

The authors discovered that most participants were not confused by terms used on the library Web pages, with a few notable exceptions. The cleverly named Kudzu, a virtual catalog of southeastern academic libraries, was not easily recognized at first glance. Users did not select it from the front page, but did select it from lower-level screens where the link appears under the description Other Catalogs. A similar pattern was apparent with AskUs.Now and other services or resources that had been given names or titles outside the usual library jargon. AskUs.Now was selected less often from the home page than from a lower-level page that contained the clues needed to allow the user to identify it as a reference service. It seems that regular users have learned to identify with catch phrases; however, new users will need to be provided with additional context.

Most participants found the redesigned library page an improvement over the older design. There are more direct links on the first page than previously. A common navigation bar appears at the bottom of all screens with quick links to the main university page, branch libraries, library hours, help information, a site-search link, and a link to a site index. Two participants attempted to use the site-search tool to locate journals, books, or articles. They each tried this strategy more than once, seeming to mistake this link for a general search engine of all library resources. It could be that a more hierarchical structure to the site, with a corresponding site map, would help to minimize the confusion of different search engines. No one used the site index provided for any of the tasks, perhaps because the title "A to Z" does not accurately describe this navigational tool.

The subjects in this study were persistent as they attempted to complete each task. The somewhat artificial

Table 4. Success rates and expertise

Status	Department	Use library Web pages?	Hours per week	Native English speaker	Success rate %
Graduate student	Sociology	Yes	10.0	Yes	100
Graduate student	English	Yes	5.0	Yes	100
Graduate student	Industrial engineering	Yes	5.0	Yes	100
Faculty	Music	Yes	5.0	Yes	87
Graduate student	Sociology	Yes	3.0	Yes	87
Faculty	Botany	Yes	1.0	No	87
Faculty	Computer science	No	0.0	Yes	87
Faculty	Environmental biology	Yes	3.0	No	75
Graduate student	Physics	Yes	2.0	No	75
Faculty	Geological science	Yes	0.25	Yes	75
Faculty	Industrial engineering	No	0.0	Yes	62
Graduate student	Information science	Yes	1.0	Yes	50
Graduate student	Mechanical engineering	No	0.0	No	50

nature of the test may have motivated them to continue searching longer than they normally might. The less-experienced searchers had a tendency to continue further down a wrong path, doggedly looking for clues. More-experienced searchers were likely to refine their searches, backtrack when lost, or try multiple searches to complete the task.

There were a number of ways in which the participants recovered when they had strayed from the best path or any direct path. Moving backward through the search with the browser Back button was a common way to make sure they had not missed a relevant selection, or to try an alternative choice. Some experienced users became aware of their mistake when they encountered a familiar lower-level page and could then reorient themselves within the Web site. Almost all the participants benefited from the presence of multiple links to commonly used tools and resources at different levels within the site. The provision of a variety of paths allowed the participants to successfully locate information and services even when they had overlooked a direct link from the main page.

Implications

The research team intends to conduct further testing of user-centered Web design applying lessons learned from this study. The screen- and audio-capture software provided an unobtrusive method for creating a permanent record of each session in a level of detail that the human observer could not accomplish. The development of the session script was critical to ensuring consistency across all the sessions. The team tried to create a comfortable setting for each participant. The screen- and audio-capture software allowed the team to observe the participants from a comfortable distance. The portion of the script that stressed that the test was of the Web site, not the participant's skill, set a nonjudgmental tone.

The study led the team to consider a number of design options to make the library's Web site a more useful information-seeking tool. For example, would a visual path and site map help in navigation? A visual path that is always visible would help orient the user within the Web site. The site map provides another method for orienting a user within the Web site. Are there ways to provide better context for terms such as Kudzu and AskUs.Now or to better describe the content of the Subject Guide pages? One solution for dealing with nonintuitive terminology for links and services would be to provide a description of the link that appears on MouseOver. Web designers must determine how to provide just enough information on a page, given the tendency not to scroll past the first screen.

Many users had difficulty locating journal articles in an unfamiliar subject area. In the current arrangement, abstracting and indexing databases are arranged as alphabetical lists by general subject area. The participants in the study were able to locate the list of databases that they needed for the journal article task, but then had difficulty with where to go from there. A portal product that incorporated a federated search feature would go a long way towards solving this problem. For example, in the Ex Libris portal product, MetaLib, users who choose a general subject area are presented with a list of databases. Instead of having to choose from among them, they can enter a term into a single search box that will be applied to all the databases.

The second part of the study consisting of structured interviews is revealing some fascinating information about how scholars might benefit from a customizable portal. That portion of the study was more open-ended, generating rich data requiring detailed analysis. The team will report additional findings in the near future.

References

1. Philip Schatz, Sampling in Research. Accessed Jan. 14, 2003, http://schatz.sju.edu/methods/sampling/sampling.html.
2. M. Patton, *Qualitative Evaluation and Research Methods*, 2d ed. (Newbury Park, Calif.: Sage, 1990).
3. Philip Schatz, Sampling in Research.
4. Jakob Nielsen, *Usability Engineering* (San Francisco, Calif.: Morgan Kaufmann Pub., 1994).
5. Lynn Lysiak, "A Web Usability Study of Freshmen at Appalachian State University, University of Arizona, and Georgia Southern University," presented at the University of Tennessee Libraries, Jan. 2001; MIT Libraries, Web Site Usability Test, Apr. 2, 1999. Accessed Jan. 7, 2002, http://macfadden.mit.edu:9500/webgroup/usability/results; Elaina Norlin, "Usability Tests and the User-Friendly Web Site: The University of Arizona Experience," The University of Arizona Library, 1998. Accessed Jan. 14, 2003, http://dizzy.library.arizona.edu/library/teams/access9798/asu/index.htm.
6. Christoph Hölscher and Gerhard Strube, "Web Search Behavior of Internet Experts and Newbies," Ninth International World Wide Web Conference, Amsterdam, May 15–19, 2000. Accessed Jan. 14 2003, www9.org/w9cdrom/81/81.html.

Content Management Strategy for a College Library Web Site

Mark Dahl

Watzek Library at Lewis and Clark College uses a dual strategy to manage the content of its Web site. Informational pages are created with a template system based on the following tools: Macromedia Dreamweaver and Contribute, PHP, server-side includes, and cascading style sheets (CSS). This system allows the site to be updated easily by several staff members and permits all pages to be presented with a graphical banner and sidebar or in a text-only fashion. Pages are presented that organize electronic and other research resources using the aforementioned system; a relational database is used to drive the content. This database is populated using locally developed PHP software that allows the building of pathfinders of resources organized by subject and category. Overall, this system provides for collaborative content upkeep, flexible presentation options, structured data, and reuse of data.

Lewis and Clark College's Aubrey R. Watzek Library implemented a new Web site in 2003. Beyond the design and architecture of the site, a major challenge was developing a framework that allows several different staff members to update content on the site with minimal knowledge of HTML or complex Web-editing tools. This article will discuss the concept of content management in an academic library setting and then elaborate on the system that was chosen at Lewis and Clark.[1]

"Content management" is a term that refers to the collection, management, and publishing of information online. The concept of content management comes into play when creating and maintaining a Web site. It involves collaboration of multiple specialists, such as designers, programmers, writers, and editors.

Most commercial content management systems are software packages designed for the collaborative construction and maintenance of large Web sites. Generally, they support a separation of content from presentation, thereby allowing the look and feel of a Web site to be controlled centrally while content-updating duties are distributed. They tend to feature defined document schemas to allow for easy repurposing of documents to fit various presentation styles and document formats. They also typically provide access and workflow controls to specify who can view, edit, and approve content that is added to a site.

In the last few years, commercial and homegrown content management systems have made their way into the higher education and academic library world. At the college or university level, they are often implemented as second- or third-generation Web sites in order to decentralize the upkeep of the site. Oftentimes, they are put in place to address frustrations that stem from centralized management of a Web site.[2]

College libraries have a great deal of content to present on their Web sites. Their catalogs, online databases, and digital repositories can easily contain hundreds of thousands of bibliographic records, electronic articles, images, and other files. This content is typically managed using an integrated library system, vendor-supplied databases, digital library system software and increasingly, federated searching systems.

When one thinks of a traditional library Web site, one thinks of the pages on the site outside of the specialized Web services previously mentioned. It includes those pages that provide information about the library and links to the resources that the library offers. Some academic libraries now build their Web sites as a part of a college- or university-wide content management system.[3]

But many academic libraries still find it most effective to create a Web site that is independent from their parent institution's Web site. Some large academic libraries have implemented commercial content management systems to manage their own sites.[4] For many academic libraries, this is not an affordable option. The solution used at Watzek Library, described in this article, demonstrates that many of the advantages of full-blown content management systems are available from a less expensive, smaller-scale solution. Watzek Library has created an inexpensive content management system that employs a mixture of commercial and homegrown tools.

Watzek Library's Web Site

In the meetings about Watzek Library's new Web site, most of the library's Web committee was concerned chiefly with the design and architecture of the site. Once a design for the home page and interior pages had been chosen, a smaller group began seriously thinking about a content management system.

The group decided that the system sought should:

- allow several library staff members to update pages on the site with little knowledge of HTML, Web-editing tools, or Web site file management;
- separate the content of the Web site from its presentation as much as possible so that the site's fonts, colors, and layout could be changed centrally and so that the content could be presented without graphical add-ons if desired;
- structure the data on the Web site as much as possible to promote

Mark Dahl (dahl@lclark.edu) is Library Technology Coordinator at Lewis and Clark College Library, Portland, Oregon.

consistency and ease migration to future platforms; and
- allow redundant data used across the Web site to be updated in one central location.

It was clear that for the portion of the Web site that supplied information about the library and its services, only the first two objectives applied, distributed updating and separation of content from presentation. Informational pages, such as those describing library hours, library staff, departments, and reference services, did not adhere to a particular schema. They also did not typically feature redundant information. Each page was a uniquely structured document containing information specific to that page.

To address the first two objectives for the informational pages, a system was constructed that uses a combination of HTML, CSS, Dreamweaver templates, server-side include files, and PHP.[5]

The portion of the Web site to which the third and fourth objectives (structured data and redundant use of data) did apply were those pages linking to the research resources, including research databases and Web sites. It was decided that the new site should have discipline-specific subject pathfinders featuring links to and descriptions of databases, Web sites, and even online catalog records of print materials, as well as alphabetical lists of databases and full-text databases. Many databases and Web sites were to be listed in several different places across these Web pages. To manage these pages, a database-driven system using PHP and PostgreSQL was built.[6]

Both solutions have taken advantage of the library's own Linux-based Web server and the staff's knowledge of HTML, CSS, and dynamic Web technologies such as PHP and PostgreSQL. They represent an important shift from static HTML pages to pervasive use of dynamically driven content.

Templates, Dreamweaver, and PHP-Include Files

The first concern was developing a distributed system for creating and maintaining the informational portion of the Watzek Web site. The librarians who had maintained the Web site in the past were used to using the Macromedia Dreamweaver Web-editing tool. To make a new page, they would copy an old one and work from there. To control the style and size of text on the page, they would allow Dreamweaver to insert font tags. This system was functional, but it did produce some pages with inconsistent fonts and styles. It also offered no means of globally changing fonts or the header and footer used across all of the Web pages.

Because of the staff's familiarity with the Dreamweaver environment and its position as the standard Web-editing software on campus, they decided to stick with that environment for the maintenance of the new Web site. Dreamweaver has a system that allows one to build a template for a page that has certain fixed, unchangeable regions, and other areas designed to be edited. The staff decided to take advantage of this template system to deploy the design for the interior pages.

Normally, when using a Dreamweaver template, one would code the basic outline of the page (for Watzek Library, a banner and sidebar) into the template and then leave the portion of the page designed to hold the changeable content empty and editable. Instead of directly coding the banner and sidebar into the interior page template, PHP-based references to server-side include files containing this code were used.[7]

Using server-side includes in the templates makes it possible to keep code for a regular banner and sidebar and a text-only banner and sidebar in a central directory. PHP *browser-sniffer* code checks to see what browser version the user has.[8] If it is compatible with Watzek's new sidebar, it then sends out the graphical version of the page requested. If a visitor to the Web site has an older browser or requests the text-only site, the PHP code will send out a text version of the page requested.

In the future, other options for output may be created such as a printer-friendly version of pages or a version of pages compatible with wireless devices.[9] In addition to keeping the code for the banner and sidebar in a central directory, a linked style sheet was used to control all of the fonts and colors used on the home page and the template pages. In order for these to have a global effect, it was decided to avoid using HTML tags that will void the effects of the CSS settings.

This system makes global changes to the interior page banner and sidebar easy. And by using PHP coding to create links, it is now only necessary to edit one central file to change the links used on the sidebar found on the home page and interior pages.

Fine-Tuning Content Management

After creating the interior page template, a training session was organized for the librarians and staff who would be updating and creating the new informational pages on the site. Many had used Dreamweaver in the past, but most were not particularly comfortable with it. The session provided a good introduction to using the new template system, but it took actual practice using the system for them to become proficient at it. Those who worked with the system for a while became comfortable with it and could quickly transfer information from the old Web pages to the new ones. Figure 1 shows a page being edited in the Dreamweaver template; Figure 2 shows the same page as displayed to the public.

Dreamweaver, however, can be a somewhat cumbersome tool for collaborative maintenance of a Web site. A Web-editing tool designed for professional Web editors, it has powerful features for coding raw HTML as well as a word processor-style editor that produces HTML. But its numerous toolbars and menus make it somewhat intimidating for someone simply updating the textual content of a Web page. In addition, its file management features are not especially intuitive. Frequently, there are instances where people will check out files that they are working on and forget to check them in, thereby locking the files and preventing others from working on them. Occasionally, users will accidentally overwrite a newer copy of a file on the server with an older copy. Sometimes they will use font tags to specify fonts on a page when staff are trying to control style using a central style sheet.

Macromedia Contribute is a recently released Windows program designed to perform only the content-contribution function of maintaining a Dreamweaver-created Web site. Fortunately, it resolves some of the above problems. It allows users to browse to a Web page and simply click Edit to edit the page in a word processor-style editor. The downloading of the most recent version of the page and the check out and check in are handled transparently. Contribute respects Dreamweaver page templates and automatically synchronizes the templates available with the latest on the server. Furthermore, Contribute users can be restricted from accessing certain files on a Web site and restricted from changing certain attributes of pages on a site, such as dynamic code and font tags.[10]

Contribute and Dreamweaver together make up a small-scale content management system. The Dreamweaver user designs and builds the Web site and associated templates while the Contribute users populate the site with content. In general, this system simplifies the Web-page updating process and enforces greater consistency. Because of these advantages, a copy of Contribute has now been purchased for everyone in the library who edits Web pages. Though some staff have been initially resistant to switch over to it, hopefully it will make maintaining site content simpler and less intimidating, while eliminating some of the problems experienced with Dreamweaver.

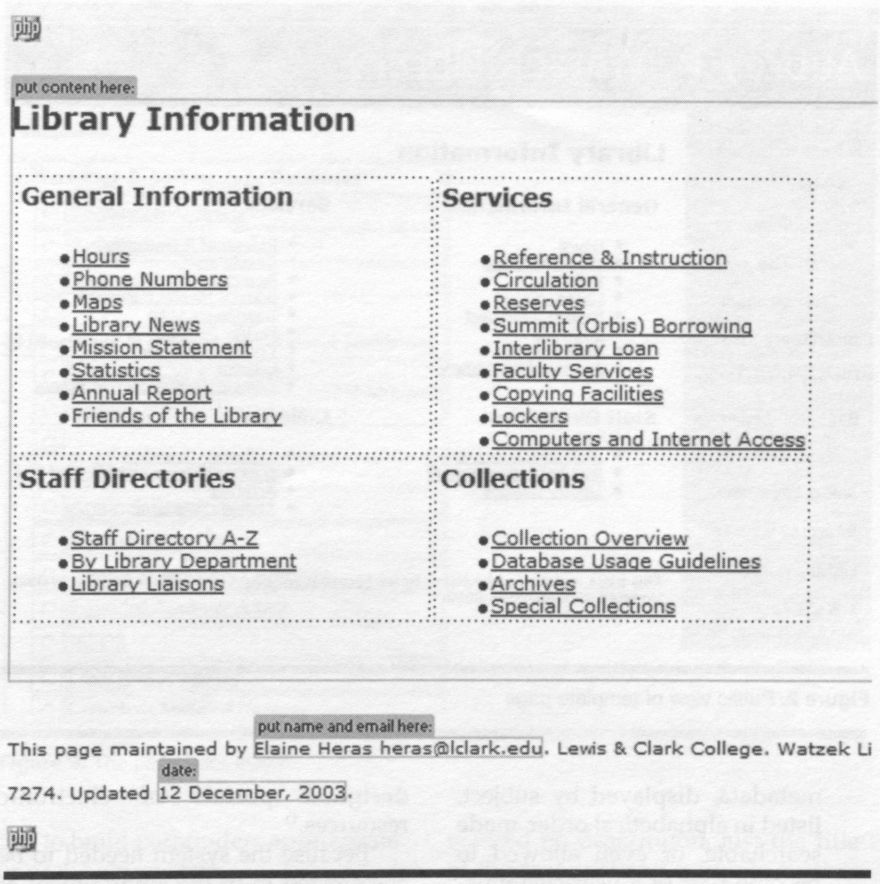


Figure 1. Dreamweaver template

Resources Editor

Though the above system works well for the informational aspects of Watzek Library's Web site, the library decided to expand upon it and implement a database-driven system for the portions of the site that display research resources. This decision was made for the following reasons:

- Each resource needed to be represented with a consistent set of metadata. These metadata included such things as: title, description, URL, and dates of coverage. A database would ensure compliance with this set of metadata.
- There would be less effort and more consistency if the data about a particular resource that might be displayed in several places were kept in one place, something a database would allow.
- A database would provide flexibility in how these resources were displayed. They could be listed with some or all of their

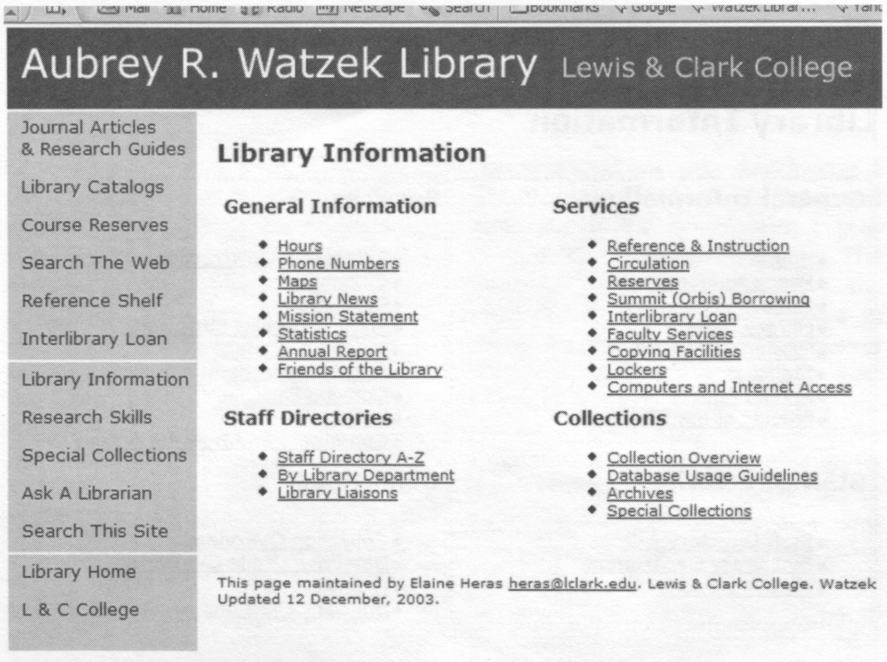


Figure 2. Public view of template page

metadata, displayed by subject, listed in alphabetical order, made searchable, or even allowed to become part of a user-customizable Web page.

- A database-driven system would allow librarians to make changes to the data associated with these resources with an easy-to-use Web interface. No Web editor such as Dreamweaver would be required for editing this portion of the Web site.

There is a strong precedent for the creation of database-driven Web pages, including subject bibliographies, in libraries.[11] Likewise, there are many open-source options out there for creating database-driven Web portals made up of hyperlinked electronic resources. The Internet Scout project, for example, offers a software suite called the Scout Portal Toolkit. The toolkit is a comprehensive package for building a Web portal of resources categorized by subject.[12] MyLibrary is another such application, with emphasis on user-designed portals of electronic resources.[13]

Because the system needed to be customized to fit the environment at Lewis and Clark, the staff decided to build such a system here. This would allow metadata specific to Watzek Library to be defined, a system of categorizing and organizing resources to be designed, and allow for determination of exactly how the resources were to be presented on the Web site. Both MyLibrary and the Scout Portal Toolkit could probably have been customized to meet the library's needs. But the library decided that programming from the ground up would be a simpler and more direct route to what was desired.

Building Databases of Resources and Pathfinders

To build the pathfinder system, the library technology coordinator gathered input from a core group of librarians on the features they desired. Using parts of PHP code from a similar application that he had developed at his previous position at Central Oregon Community College, he began constructing a PHP/PostgreSQL-based system. PHP was a particularly good language for this task because of its strong PostgreSQL connectivity features and the many convenient functions that it has for manipulating HTML. PostgreSQL was a good candidate for a database because it is robust, open source, and comes with Red Hat Linux, the library's Web-server operating system.

The system is based on *resources*. Each resource represents a Web site, database, online reference source, or link to a print book or journal via Watzek Library's Web-based catalog. Resources are kept in a Resources SQL table with several fields that are loosely based on the Dublin Core metadata standard, including publisher (vendor), title, description, and coverage. Subjects are not included in this descriptive data.[14] Using PHP, a Web-based interface was built that lets staff enter (or catalog) new resources into this table as well as search and edit existing ones.

In order to present these resources to users of the Web site, it was decided to organize them in pathfinders that would guide users to appropriate resources for a particular subject. Therefore, a second table was added to the database to hold the data needed to order resources in pathfinders. It was necessary for the pathfinders to have the following characteristics:

- They would have a unique name (the "subject")
- They would be organized with categories that would contain resources within them (for example, categories in a history pathfinder might be called History Databases, History Web sites, or History Primary Sources)
- Resources could be put in a certain order within those categories

A pathfinder SQL table holds the pathfinder data. It is made up of rows that represent the position of a resource in a pathfinder. Querying the pathfinder table for a particular pathfinder name and then linking the results to the resources table provides all the data that are needed to construct a subject-pathfinder Web page.[15]

In order to build pathfinders, the staff interface was expanded so that it could be used to add entries to the pathfinder SQL table. With much feedback and testing by a group of three librarians, a Web-based tool was built that allows librarians to pick resources from a pull-down menu or from a search and place them at the appropriate spot in a pathfinder. With this tool, they can add, remove, reposition, or recategorize resources in a pathfinder as desired. Figure 3 shows a pathfinder in the staff interface.

Once a system had been constructed to hold the data, the next step was to populate it with resources. This process involved training librarians on the system, then going through the existing databases-by-subject and Web sites-by-subject pages and creating entries for resources with those data. Interestingly, it was found that not all of the information about the library's research resources fit nicely into the metadata fields that had been defined for resources. But through a process of discussion, compromise, and adjustments to procedures, work-arounds were found that allowed the presentation of the data thought to be important to Watzek Library's patrons.[16]

As resources were entered, they were placed in the appropriate pathfinders. To some extent, general rules were devised as to how to structure pathfinders, such as where to put discipline-specific databases and where to put multidisciplinary databases. In most cases, these rules worked; other times they did not. But the flexibility of the system allowed staff to build pathfinders appropriate for each subject area.

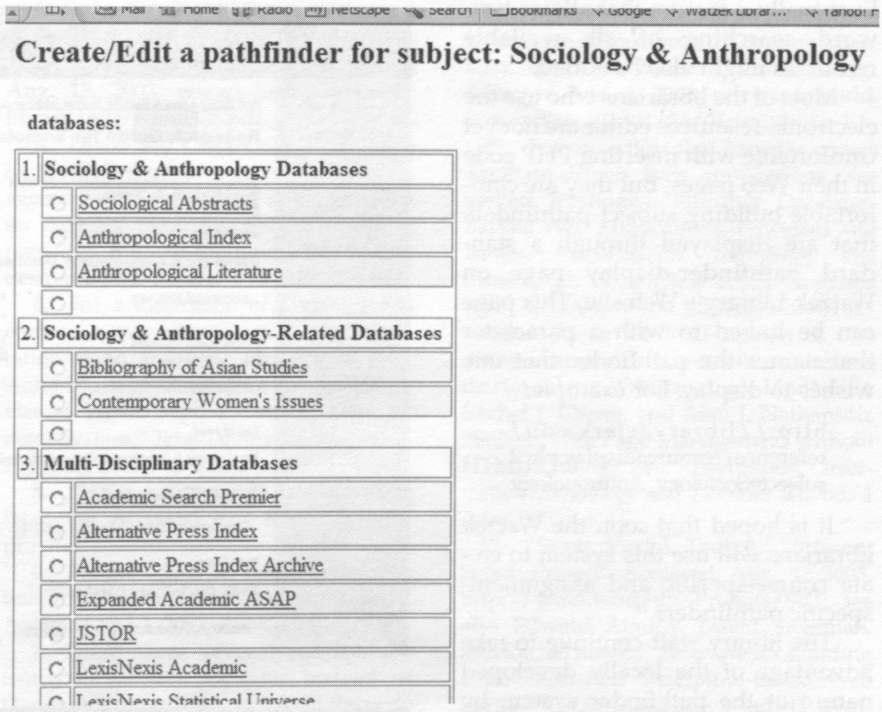


Figure 3. The pathfinder editor

Before this project, the library had separate sets of Web pages that listed databases-by-subject and Web sites-by-subject. Switching to the pathfinder system made it logical to combine databases and Web sites for a particular subject on a single pathfinder. This caused the need for the distinction between a Web site and a database to come into question, and eventually it was decided to maintain the distinction for the present.[17]

Displaying the Data

With the resources in the database and organized in pathfinders, a great deal of discussion continued around the question of how to present the pathfinders as Web pages that the library's patrons would view. Some librarians wanted complete descriptions of the resources on the pathfinder pages. Others wanted no description, just the title of the resource and a link to further information.

To build the pathfinder presentation system, PHP's object-oriented features were implemented so that resources could be displayed simply by creating an instance of a PHP class at a desired spot within an HTML document. This allows the display of resources from the database to be done within any PHP-enabled Web pages on the library's server, including pages built with the template system previously discussed. It allows an entire pathfinder or even just one category of a pathfinder to be displayed at any location within a Web page. Figure 4 shows the default method of displaying a pathfinder to the public.

In addition to displaying the pathfinders (which list resources by subject), the display system was also set up to provide alphabetical lists of all databases and all full-text databases so that users could browse to a particular resource by name.

Eventually, a feature that allows keyword searching of all available resources might also be added.

Most of the librarians who use the electronic-resources editor are not yet comfortable with inserting PHP code in their Web pages. But they are comfortable building subject pathfinders that are displayed through a standard pathfinder-display page on Watzek Library's Web site. This page can be linked to with a parameter that names the pathfinder that one wishes to display. For example:

> http://library.lclark.edu/reference/resourcedisplay.php?subject=Sociology_Anthropology

It is hoped that soon the Watzek librarians will use this system to create course-specific and assignment-specific pathfinders.[18]

The library staff continue to take advantage of the locally developed nature of the pathfinder system by further customizing it to meet specific library needs. Recently, a brief description field was added for each resource, which will display on the pathfinder pages instead of the longer, full descriptions. A field was also recently added for information on how to find book reviews in a particular resource, if applicable.

The addition of more interactive features for patrons is under consideration. For example, patrons might be able to add comments about the resources on the site, or to suggest or add resources to the database on their own. Furthermore, they may be allowed, MyLibrary-style, to build their own pathfinders or portal-style pages out of the library's basket of resources. The existing system could be modified to add such features or the data could be migrated to a different system, such as MyLibrary or The Scout Portal Toolkit.

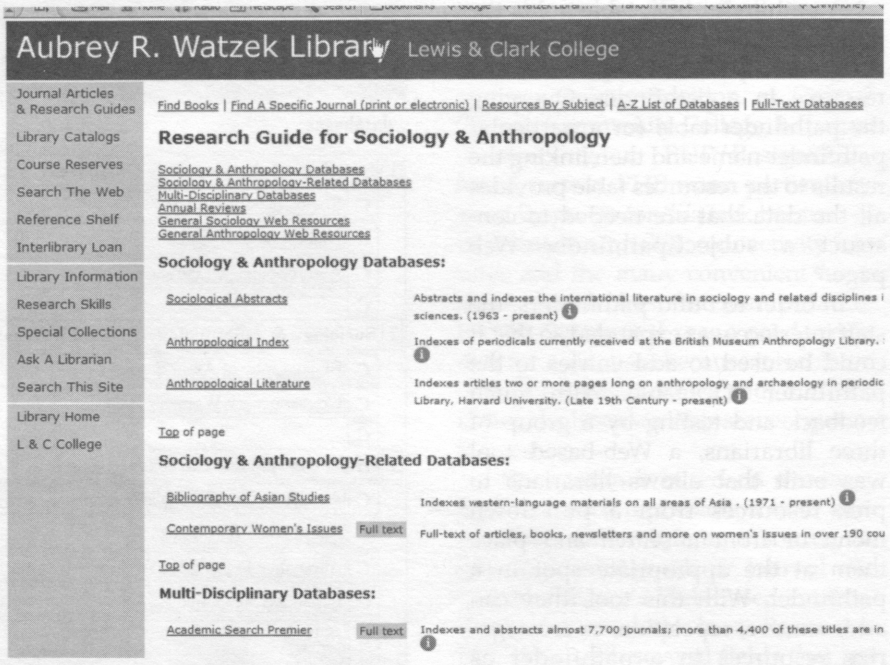


Figure 4. Public view of pathfinder page

Conclusion

Watzek Library has now moved from a Web site based on static HTML pages to a site that is heavily dependent on dynamically generated pages. Both the template system and the database-driven pathfinder pages rely on PHP scripting. In contrast to a traditional Web site configuration, where scripts are reserved for special purposes and kept in a special directory, scripting on the library Web site is pervasive.

This configuration provides numerous advantages; at the same time, it is also complex and creates challenges. The previous Web site was based on straight HTML files, and was therefore easy to mount and support. The coding of the site was simple enough for anyone familiar with HTML to understand, and the site could be mounted using an account on the campus Web server without any special configuration issues.

The new site depends on a PHP-configured Web server, many interdependent scripts, and a relational database. The development and implementation of the site took a core group of librarians and the library technology coordinator about six months. Of that time, the library technology coordinator spent approximately two months full-time programming and editing PHP code. This code, as well as the other technical systems that support the site, needs continual adjustment and fine-tuning. And as long as this system is used to maintain the site, the library will be committed to having expertise in the technologies used to create it.

From the perspective of those who update content on the site, the new site's content-management strategy has worked out well. A core group of staff members updates the informational pages, and they have been able to use the template system with limited technical assistance. Sometimes the intricacies of using Dreamweaver or even Contribute can frustrate staff and lead to formatting problems and inconsistencies on the

pages. But they are generally able to iron these out on a case-by-case basis.

An important strength of the system is that it allows staff to quickly make global changes to the look and feel of the site. As an example of this, nine months after implementing the new site, the front page illustration and the fonts used across the whole site were revised.

The pathfinder system has also worked out well. With the old site, only a select group of people created and maintained the database and Web sites-by-subject pages. Now, all of the Watzek librarians are able to actively participate in the description and organization of the resources offered on the site. This wider level of participation has helped provide richer descriptions for the resources offered by the library. It has also presented a challenge in that some librarians, particularly those not previously involved with the Web site, sometimes forget about their newfound Web-editing responsibilities. Another challenge is making sure that all contributors use the various metadata fields in the system in the same way.

Despite these drawbacks, the content management system has made Watzek Library's site easier to work with for staff and more appealing and relevant for library patrons. Furthermore, its flexibility will allow the site to evolve as the library launches new Web-based services in the future.

References and Notes

1. The URL of the Web site discussed in this article. Accessed Apr. 8, 2003, http://library.lclark.edu.

2. For good overviews of the functions of commercial content management systems and their implementation in the context of higher education, see Paula Browning and Mike Lowndes, "Content Management Systems: Who Needs Them?" *Ariadne* 30 (Dec. 2001). Accessed Apr. 4, 2003, www.ariadne.ad.uk/issue30/techwatch; see also Wayne Powel and Chris Gill, "Web Content Management Systems in Higher Education," *Educause Quarterly* 26, no. 2 (2003). Accessed Aug. 15, 2003, www.educause.edu/ir/library/pdf/eqm0325.pdf.

3. Lewis and Clark College utilizes a locally coded campuswide content management system for many of the pages on its Web site. Unfortunately, it does not provide the presentation flexibility desired by the library.

4. For a discussion of a commercial content management system implementation at an academic library, see Pixie Anne Mosley, "Recreating a Web of Service at TAMU with a Content Management System," *Texas Library Journal* 79, no. 1 (spring 2003): 16–19.

5. PHP is a server-side Web-scripting language. Accessed Apr. 8, 2003, www.php.net.

6. PostgreSQL is an open-source relational database management system. Accessed Apr. 3, 2003, www.postgresql.org.

7. There were several reasons for using server-side includes instead of Dreamweaver templates to hold the code for the banner and sidebar: First, Dreamweaver did not display the interior page correctly in its internal editor, so having the code there for it might confuse those using the template. Second, it was decided to offer a text-only version of the sidebar and banner because some older browsers were not supportive of the CSS-2 standards that were being used in the sidebar and banner. Third, Dreamweaver does not offer a good system for updating templates across all users' computers. If a staff member made a change to a template on the Web site, that individual would have to remind everyone to download the latest version of the template to their local computer.

8. Code of the browser sniffer. Accessed Apr. 8, 2003, www.eit.ihk-edu.dk/instruct/php-sniffer.php?e=0,7,22#0.

9. Watzek Library's site is currently coded in HTML 4.0 transitional, but could be easily converted to XHTML using Dreamweaver. As an XML-compatible document format, XHTML opens up the possibility of making Web pages compatible with wireless devices and other nontraditional Web-browsing systems. XHTML 1.0: The Extensible Hypertext Markup Language, 2d ed., World Wide Web Consortium. Accessed Apr. 9, 2003, www.w3.org/TR/xhtml1.

10. For a review of Macromedia Contribute, see Alyson Behr, "Updating Made Easy," *InfoWorld* 25, no. 4 (Jan. 24, 2003). Accessed Apr. 4, 2003, www.infoworld.com/article/03/01/24/04contribute_1.

11. For an article that broadly covers database-driven Web applications, see Steven Westman, "Building Database-Backed Web Applications: Processes and Issues," *Information Technology and Libraries* 21, no. 2 (June 2002). Accessed Apr. 4, 2003, www.lita.org/ital/2102_westman.html; for a discussion of two different approaches to creating database-driven subject Web pages, see Linda Bills, Rachel J. Cheng, and Alan J. Nathanson, "Subject Web Page Management without HTML Coding: Two Approaches," *Information Technology and Libraries* 22, no. 1 (Mar. 2003): 4–11.

12. "Scout Portal Toolkit," Internet Scout Project. Accessed Apr. 4, 2003, http://scout.wisc.edu/Projects/SPT; see also Edward Almasy, David Sleasman, and Rachel Bower, "Software for Building a Full-Featured Discipline-Based Web Portal," D-Lib Magazine 8, no. 11 (Nov. 2002). Accessed Apr. 4, 2003, www.ala.org/ala/lita/litapublications/ital/2102westman.htm.

13. Accessed Apr. 4, 2003, http://dewey.library.nd.edu/mylibrary.

14. Accessed Apr. 3, 2003, www.dublincore.org.

15. Each row in the pathfinder table has a field for the pathfinder name, the resource ID, the name of the category in which the resource is placed, the position of that category in the pathfinder, and the position of the resource within that category.

16. For example, sometimes the resources descriptions had hyperlinks in them. It was discovered that simply inserting the HTML needed to create the link in the resources editor would make the link appear.

17. At Watzek Library, the databases include traditional article indexes as well as other online reference sources subscribed to by the library. Web sites are basically everything else available on the Web.

18. Building the pathfinder display system so that URLs remained short and did not require special characters was found to be helpful. This facilitates remembering the URL and passing it between people on paper and electronically.

Thermal-Transfer Printing: A Better Way to Print Library Labels

Cheryl D. Walters

Thermal-transfer printing, a technology borrowed from the manufacturing sector, offers libraries a flexible method for printing durable, accurate, legible, and attractive labels that reliably adhere to most book surfaces. When guided by an electronic program customized to meet a library's particular needs, a thermal-transfer printing system offers virtually limitless variations in font, format, and functionality. It can print labels directly from the online catalog, thereby guaranteeing that call numbers on labels match what patrons see in the catalog. This article explains thermal-transfer printing and how it compares with other printing technologies, briefly explores applications in both the manufacturing and library environments, and describes in detail how Utah State University Libraries and a few other libraries use it to improve the accuracy, appearance, and durability of their spine labels.

It is no secret that libraries often make use of technology first developed and tested in the business and manufacturing sectors. A well-known example of this borrowing is bar-coding technology; a more recent but lesser-known example is the automated storage and retrieval system (robotic retrieval of densely stored materials) being incorporated in some of the newest library buildings. Now libraries have another opportunity to capitalize on a technology developed and tested by companies and businesses—thermal-transfer printing.

As described in this article, Utah State University Libraries (USUL) uses thermal-transfer printing to print better labels via a system combining thermal-transfer printing technology with an electronic program that formats and prints data directly from the library online catalog. Because a program instructs the printer how to print the label, the possibilities for variations in font, format, and functionality are virtually limitless. The program automatically sets font type and size according to data length, moves data around on the label, translates codes into text, prints batches of a single label with or without incrementing enumeration, and prints label sets, among other things.

USU's printing system was created in 2000–2001 in collaboration between USUL and a label vendor, Computype. Shortly afterward, the label vendor worked with the online catalog producer Dynix to make this printing technology available to other libraries. Several other libraries' printing systems also are briefly described in the following paragraphs. Because of the flexibility of these thermal-transfer printing systems, each library's setup and experience is different.

Background

Like many libraries, USUL has searched a long time for a way to efficiently print attractive, legible labels that endure rather than fade, smear, or simply fall off the book throughout its years of use. Providing such labels seems simple, and yet all of the different ways of creating library labels that were tried in the past proved unsatisfactory in one way or another.

In the early part of the twentieth century, USUL, like many others, manually wrote call numbers and other vital elements on book spines using a stylus pen and *library hand*, a standardized form of handwriting taught to aspiring librarians in library courses. While this method created durable text on the spine, its legibility depended on the vagaries of handwriting skills while its accuracy depended on the diligence of the writer. Superceding the stylus, typewritten labels improved legibility by adding the consistency of a typewritten font, but unfortunately suffered the same potential for errors as their predecessor along with introducing new flaws: a tendency for the text to fade and the label to peel over time. Another promising method yielded a very durable label. The Selin label used heat to *cook* the label onto the book but proved inconvenient, expensive, labor-intensive, and not very attractive or flexible.[1] Dot-matrix printed labels, while easy to create, suffered from various shortcomings associated with inked-ribbon technology (misalignment, dried out ribbons) as well as text that smeared or faded and label stock that was only as good as the adhesive used. While font adjustments were possible, they were not easy to effect, usually requiring multistep routines.[2] More recently laser printing—with its crisp letters, flexible fonts, and word-processing capabilities—showed great initial promise but then disappointed with text that smeared and label stock that eventually peeled and only came in inconvenient sheets of fifty-six labels.[3] All of the above printing methods shared a certain potential for errors in call numbers and other text because labels did not print directly from the online catalog's item records where users see the call numbers. Call numbers displayed in the online catalog were thus not guaranteed to match the call numbers printed on spine labels, a serious, persistent problem that resulted in books that were misshelved or missing.

During more than one hundred years of operation, the libraries at USU tried all of the aforementioned

Cheryl D. Walters (chewal@ngw.lib.usu.edu) is Head of Cataloging at Utah State University, Logan. Graphics for this article were created by Nathan Walker, Systems Administrator, Utah State University Libraries.

traditional label-printing methods with unsatisfactory results. Indeed, if one walks through Merrill Library's long stacks of books, label inadequacies are clearly evident: labels faded so badly they can hardly be read, labels peeling off the spines, and fonts that are difficult to read. For a number of years, typewritten, dot-matrix, or laser labels were covered with a clear layer of glue to help labels adhere to the book. But this was messy and time-consuming; while the label lasted longer, it often originated with print that was too light or contained typos, depending on the label creator and type of printer. More recently, label protectors replaced glue but also added another cost to the process.

Desired Specifications

In 2000, as Head of Cataloging at USUL, the author began working with the bar-coding and labeling company Computype to design a label-printing system driven by an electronic program based on the library's needs and specifications.[4] The following desired attributes were specified:

- the ability to print from any workstation;
- low-maintenance equipment;
- dark, crisp print;
- the ability to print either single or paired labels, without wasting labels;
- text that would not smear;
- a font that would clearly distinguish between all letters and numbers;
- a scalable font to fit the label, automatically increasing or diminishing according to the length or width of the call number;
- the ability to print directly from the item record in the online catalog to ensure that the label on the book always matched the call number in the online catalog;
- the ability to translate the library's various location and collection codes into understandable label headings;
- affordability; and
- label stock that adhered well to most book surfaces.

To meet these numerous criteria, a printing technology that manufacturers used extensively but which libraries had only recently discovered was selected: thermal-transfer printing. The proposed system would use a second small component called a LabelMorphor to integrate the thermal-transfer label printer with the library's online catalog.

What Is Thermal-Transfer Printing?

Thermal-transfer printing found wide acceptance in Japan about twenty years ago because of its flexibility in printing complex kanji characters.[5] About eight years ago, United States companies discovered its usefulness in printing compliance labels and bar codes, especially for situations where printed labels were exposed to harsh manufacturing or environmental conditions. Avon Corporation, the well-known cosmetic company, moved from inkjet printing to thermal-transfer printing to improve its inventory tracking. It integrated Zebra IOSS thermal printers into its warehouse-management system and production lines. An Avon packaging-quality engineer reported that the use of bar codes and the new printing system expedited the flow of goods shipped and reduced errors. "We know we have less chance for errors since the system requires less data input," he said.[6] Another company, Selectron Corporation, switched to a high-resolution thermal-transfer printer to print tiny circuit-board labels. The labels had to withstand high temperatures and a water bath as part of the manufacturing process. The flexibility of the new thermal-transfer printers with their varying font sizes eliminated the need to stock many different label sizes and reduced label wastage by 90 percent.

Thermal Transfer versus Direct Thermal

There are actually two kinds of thermal printing: thermal transfer and direct thermal. Both use a thermal printhead containing two hundred to six hundred heating elements (dots) per linear inch (dpi), but while direct thermal applies the heated print head directly to chemically treated paper, thermal-transfer printers apply the printhead to a ribbon that transfers the image or text to the label material (called a substrate).[7] Thermal-transfer printing has several advantages over direct-thermal printing. Printheads last much longer in the thermal-transfer process because the ribbon serves as a buffer between the label and the printhead. While thermal-direct printers have the advantage of not requiring ribbons, they are limited to special chemically treated paper, whereas thermal transfer permits a wide variety of label materials such as paper and various synthetics, including vinyls (polyolefin, polyethylene, polypropylene, polyvinylfluoride, etc.). Unlike direct thermal's chemically treated paper that fades as it ages or blackens when exposed to heat or sunlight, thermal-transfer materials can endure even harsh environmental conditions such as extreme heat, cold, and moisture. Fortunately, libraries do not require the endurance of Kapton labels, which can withstand up to 1,000 degrees Fahrenheit (but only for a few seconds) and are used for printed circuit-board production. Instead, libraries can use a less expensive material such as polyethylene that is of moderate stiffness, yet flexible enough to curve and adhere to a book spine.

Thermal-transfer printing also has distinct advantages over more popular

printing technologies such as impact (dot matrix), laser, and ink-jet printing. While multipart forms require impact printers, these break down more, require ribbons that are subject to drying out, provide poorer image quality, and are limited in the materials on which they can print. Laser and ink jet printers likewise provide poorer image quality, less durable print, and fewer label-material options than thermal-transfer printers. Ink jet printers have the further detriment of taking longer to print.[8]

The Printer

When using thermal-transfer printing, print quality depends on the printer, the ribbon type, and the label stock, and how well these work together. The printer's printhead controls the resolution of the print which is measured in dpi (dots-per-inch) or dots/mm (dots-per-millimeter). On the low end, thermal-transfer printers have a resolution of 203 dpi or 8 dots/mm (for example, Allegro 2, Sato CX200, Zebra 160S). Printers in the mid-range offer 300 dpi or 12 dots/mm (for example, Datamax's DMX- I-4308, Zebra 105Se). Recently Sato released a new ultra-high resolution thermal-transfer printer, the M84pro, with a resolution of up to 600 dpi or 24 dots/mm.[9] USU Libraries' Zebra 105Se has 300 dpi resolution, which has been found to be very satisfactory. One of the label formats includes very small text and a small bar code; even at this size, the text is quite legible and the bar code scans with no problems. An easily overlooked thermal-transfer printer attribute that ought to be considered is the amount of noise it produces as it prints. McClean pronounces thermal printing "silent" compared to other printing processes, though USUL's experience belies his assertion.[10] The Zebra 105Se produces quite an aggravating screech, especially when printing a long string of labels. However, library staff have become accustomed to the noise and those who work in far-flung areas of the department benefit from hearing the confirming sound of the printing in operation. An especially critical feature to check when considering a printer is whether it can provide the desired font type, size, and scalability.

The Font

Finding a thermal-transfer printer that would support a suitable font was difficult for USUL because library staff had stringent requirements. Of course attractiveness, scalability (the ability to adjust size vertically and horizontally), and crisp, dark letters were desired. But another requirement was that all letters and numbers be clearly differentiated, especially the letters i versus l versus the number 1, the upper case B versus the number 8, and the lower case b versus the number 6. Clearly differentiated characters are critical for getting books reshelved correctly. After a few months spent evaluating font samples provided by the vendor, the process was reversed, sending the vendor several fonts that were *known* to work and asking them to find a printer to support them. The company found a printer, the Zebra 105SE, that supported the most desired font, Century Schoolbook.

The Label

Selecting a label stock involved finding an appropriate durable material and an adhesive well suited to book surfaces. Since the company specialized in creating labels and bar codes, there were no limits in the choice of different label stock and adhesives. An adhesive was selected that chemically bonds with book surfaces within twenty-four hours. The decision was made to use Computype's TS503 label stock. A custom-order version of this label stock was requested, consisting of 1" x 1.5" labels of 2.8-mm white polyethylene in continuous rolls of one thousand labels. The adhesive on this label has a very aggressive initial tack (it sticks well when first applied) as well as a good chemical bond that develops subsequently. At 1.1 mm, it is thicker than a normal adhesive, giving better adhesion to books with a textured surface. According to Computype, the rubber-based adhesive has been age-tested with good results; it is the same adhesive used at the library with excellent results for an outsourced retrospective project in 1997–98.

Previously, pairs of labels were used, one 1" x 1.5" label for the spine and a larger inside-label that provided the author, title, and call number. Having these two different-sized labels resulted in some wastage when only a spine label (without the inside label) was needed. By taking advantage of font scalability, the size of the font was reduced and the printing turned sideways (landscape) so author, title, and call number could fit on the same 1" x 1.5" label used for the spine (figure 1). Another element—the bar code—was added, and it was discovered that the printer could print a scannable bar code on the former inside label, thus making this second label useful for inventory purposes. With this in mind, the decision was made to move the inside label to the outside back cover of the book to facilitate inventory scanning.

Creating a Formatting Program

Working with Computype's programmer via phone and e-mail, specifications were created to manipulate data from the library's online catalog so that it formatted and printed correctly on the label. Data streams generated by the online catalog's print-label command were captured and sent to the programmer to ana-

lyze. A data stream is simply the electronic message sent from the online catalog to the printer telling it how to format the label and what data to print. To generate these, an item record was called up in the catalog, a label was formatted using regular database commands, and then instead of choosing a printer to send the data to, "Generic, Text Only" was selected and the data was saved as a file that could be read using Notepad. These were then sent off to the programmer as e-mail attachments.

During a dialogue with the programmer that lasted several weeks, system-printing parameters were adjusted in the online catalog as well as the printer's program specifications. The company ran the program through its own testing regime and then sent the library an EPROM chip that was installed into the LabelMorphor for a test run. The EPROM chip is the brain behind the thermal printer; it contains the electronic program that transforms the data in the online catalog data stream into the desired label format. Library staff tested it for a month or so, requesting two programming changes that were accommodated simply by the company mailing new EPROM chips that staff installed with no problem. Total time for development of this printing system from its inception to final testing was about eight months; cost for the programming work was $1,700. Later, after using the program for a while, it was felt a batch-printing feature needed to be added; the library contracted with the company to add this enhancement for an extra fee of $250.

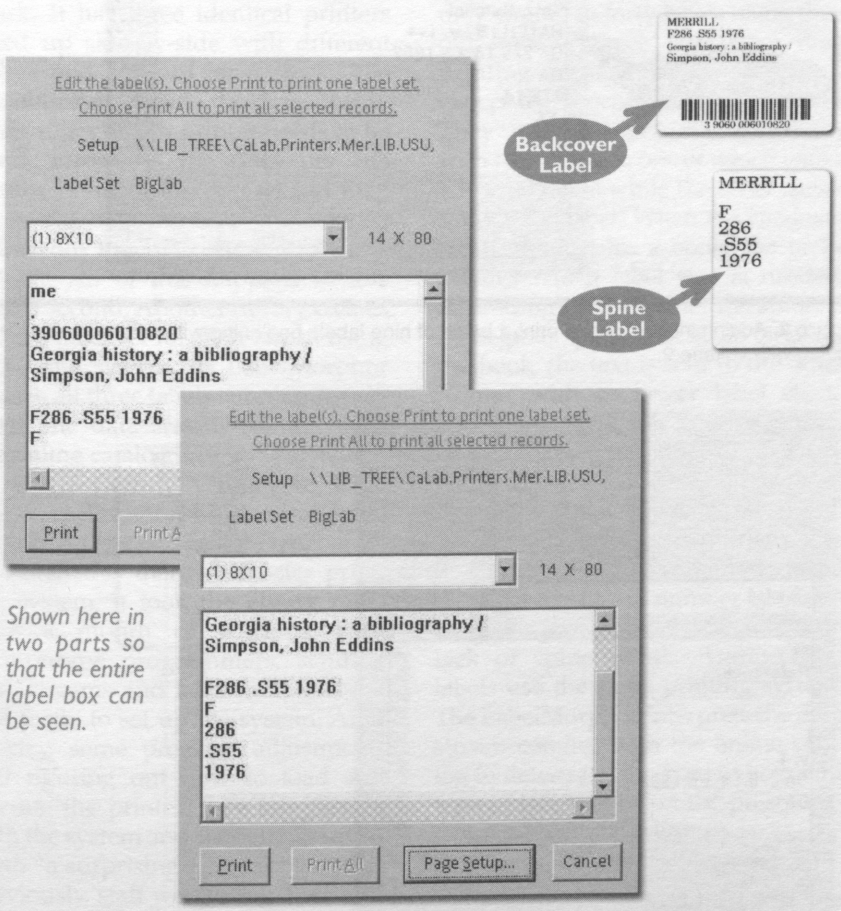


Figure 1. The one label format in the online catalog generates the two different labels above. The period (.) in the second line of the Print Spine Label box instructs the printer to leave the collection area of the label blank.

Batch Printing

The batch-printing enhancement enables more than just batch printing; it also can automatically increment the volume enumeration (volume 1, volume 2, and so on). On the volume-enumeration line that normally contains volume text such as "volume 1," the word "BATCH" is entered followed by a space, colon, space, and then a number indicating the desired number of labels. To increment volume enumeration, ++ is appended to the volume number (figure 2). Thus, to create ten labels for volumes 1 to 10, the statement: BATCH : 10 : v. 1++ is used. The program accommodates multiple levels of volume enumeration, but only the last level changes numbers incrementally. So, for volume 1, parts 1 to 3, the statement reads: BATCH : 3 : v. 1, pt. 1++ . To generate one hundred labels that were exactly the same, the statement is simply: BATCH : 100.

Other Functions

In addition to providing batch-printing capability, the printing system gives options such as a choice of printing spine labels alone or paired with accompanying back-cover labels (figure 3). To maximize font size, the printer uses a fairly large default font size, but if any line in the call number exceeds six characters, it automatically reduces the font for the entire call number. Extra long titles and author lines automatically wrap around. The program translates the library catalog's location codes into readable headings. For example, the

location code "me" would generate the text "MERRILL" at the top of the label to indicate that the Merrill Library housed the item. Likewise, collection codes are translated into recognizable collection names using the online catalog's label parameters. The location code "me" and collection code "mref" for reference books in the Merrill Library, for example, cause the label to print with the location "Merrill" on the top line and the collection "Reference" on the second line.

The thermal-transfer label printing system has worked flawlessly in the two years it's been in operation. Staff can print labels from any location in the library because the printer is accessible via the library network. The labels are consistently clear and dark. They stick so well to most book surfaces that there is rarely a need to use label protectors. Several online catalog upgrades, one of which was very substantial, have occurred, with no adverse effect on the printing capability.

Each label order includes printer ribbons and cleaning solution. Thermal printers should be cleaned every ribbon change to avoid problems such as unwanted lines and blank spots caused by dust on the platen or printhead.[11] Unlike dot-matrix printer ribbons, thermal-transfer ribbons neither dry out nor do they blacken fingers as they are changed. Because ribbons come with the labels and there is no longer a need for label protectors, ongoing label costs at the library are actually less than with the old dot-matrix printing system. Label stock, complete with ribbons and cleaning solution, costs 3.6 cents for each pair of labels; fifty thousand labels are ordered at a time to obtain this low price. By contrast, the former label stock entailed the following costs: 3.8 cents per pair of labels, plus 2.9 cents for a label protector, plus ribbon costs, for a total exceeding 6 cents per pair of labels. According to Computype, an economy version of this system has been developed for smaller

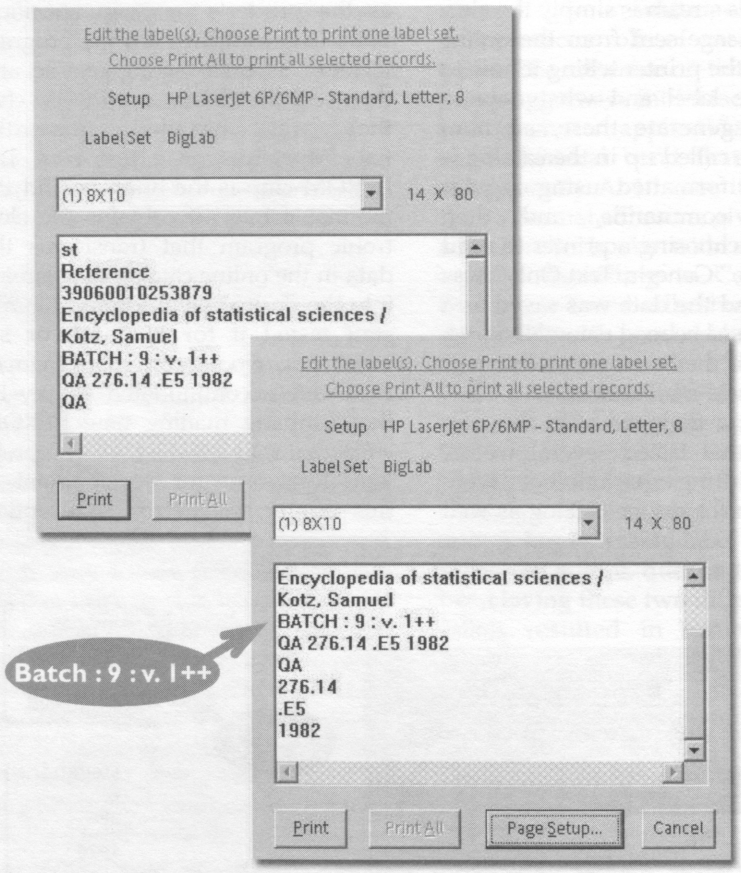


Figure 2. Adding instructions to print a batch of nine labels beginning with volume 1 and ending with volume 9.

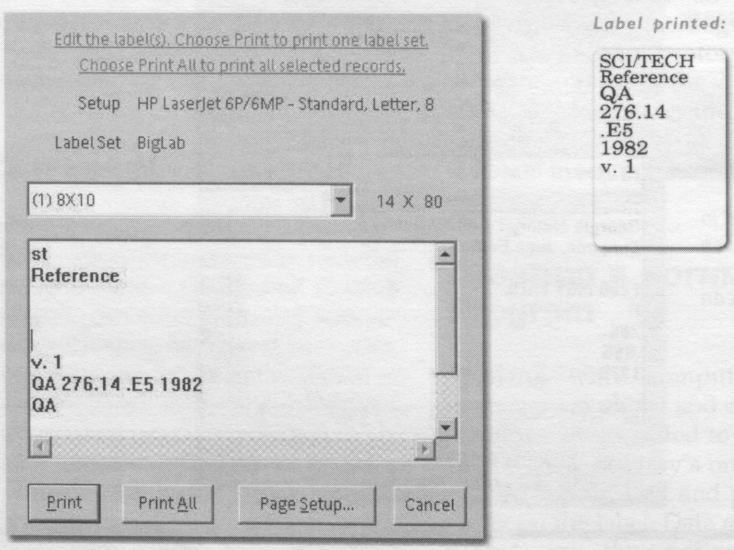


Figure 3. Deleting lines 3 to 5 from the label format instructs the printer to omit printing a back cover label.

libraries that have less demanding label requirements and workloads as well as smaller budgets. The technology is the same but a smaller, lower-resolution printer is used.

Based upon the success of this collaborative venture, Computype worked with the library's database producer Dynix to make this labeling system available to all Horizon and Dynix database users. Other database producers Computype has worked with include TLC, Innovative Interfaces, Inc. (III), and CARL Systems, Inc. (CSI); it currently is working on projects with Sagebrush, SIRS Mandarin, and SIRSI. Each library gives its own twist to its printing system. One library prints labels using three different colors of label stock. It has three identical printers lined up side-by-side with different colors and sizes of label stock. The LabelMorphor reads the data stream, determines which printer needs to be used, reformats the data per the requirements of the library, and then sends the data through an electronic switch box that directs it to the correct printer. All of this happens in less than a second. Another library creates a three-label set for their books using this same system. Its LabelMorphor creates all three labels simultaneously from one data stream coming from the online catalog.

Davis County Library in Utah uses a smaller, more economical Zebra printer. As one of the first Dynix users to install this thermal-transfer printing system, it took the library more than a month of working with Computype programmers, sending data streams and label samples back and forth, to set up the system. After making some program adjustments and figuring out how to load and operate the printer, they are pleased with the system and claim it has saved them "a surprising amount of work." Previously, staff were using a secondary program that necessitated extra steps in their workflow to enter and print out labels; now they just "spit them out" using print commands in their Dynix module during the regular cataloging process. While loading ribbon in the big Zebra printer used by USUL is an easy task, apparently it is not so easy for the smaller printer, according to Jerry Meyer of Davis County Library; he would prefer to have the ribbon provided in a cartridge for easier loading.

Orem Public Library in Orem, Utah, has designed an especially sophisticated system that accommodates about twenty-five different label formats for their various collections and media types. Because of the complexity involved, it took about three months to design and program their system. But the time was well spent: they now can print 160–250 labels in ten minutes compared to only twenty to forty labels using their old system. Once they have their printing system networked, they expect the process to be even faster. They use an electronic switchbox with two printers, one of which uses a 1.5" x 1.0" label while the other uses a 2.5" x 0.75" label. When the cataloger creates a label for a book, she or he decides which label size is needed, depending on the size of the spine. If the regular spine label is too large for the book, the text is sent to the other printer with narrower label stock. Text prints onto this narrower label using landscape mode so that the entire call number will fit on the label in such a manner that the label will run along the spine rather than across it. This flexibility has greatly reduced the need to put call number labels on the front or back of the book due to lack of spine width. Audio-visual labels use the same printing system. The LabelMorphor interprets the data stream coming from the online catalog to determine the type of item and type of label based on the preset formats created by the library.

Conclusion

Thermal-transfer printing technology is an affordable, effective solution for libraries looking for a way to create legible labels that both look good and endure. While requiring some initial investment of time and effort to design and write a program to fit each library's unique needs, the addition of a programmable LabelMorphor to interpret and format data from an online catalog results in a highly customizable printing system that will print accurate, legible, and attractive labels. While this article focuses on libraries using a customized program to integrate label printing with their online catalog, other off-the-shelf options employing thermal-transfer printing exist such as TAGPRINT PRO or Label Matrix, both of which advertise that they work with Windows and support thermal-transfer printers.[12] A search of the Web for companies offering thermal-transfer printing equipment and services yields many sites with useful information.[13] Whether a library is looking for a turnkey printing system or one customized for its use, thermal-transfer printing is an option worth exploring.

References and Notes

1. Don Samdahl, "Automated Spine Labeling at the University of Alabama Libraries," *Library Software Review* 11, no. 1 (Jan./Feb.1992): 13–15. Article describes Vernon's Label Printing Software using a dot-matrix printer to print using OCLC's s14 or s16 functions. Details problems using Selin labels as well as printers with platens. See also Kathy Fladland, "Using CAT ME Plus's Label Module with SELIN Labels," *OCLC Micro* 7, no. 4 (Aug. 1991): 29–34. Article describes using the CatME program with Selin labels; includes description of the Selin labeling process and the problem of wastage. Another article that describes a locally designed program for producing Selin labels directly from bibliographic records is John Duke, "Sprint: A Microcomputer Program for Producing Call Number Labels," *Information Technology and Libraries* 10, no. 1 (Mar. 1991): 61–67.

2. Emily Badertscher, "Yellow Bricks: Combination Lettering on Label Sets Using PASSPORT," *OCLC Micro* 8, no. 6

(Dec. 1992): 29–30. Article describes setup and procedures for producing double-high letters using a dot-matrix printer with OCLC PASSPORT.

3. Z. Kelly Queijo, "Getting Out of a Sticky Situation: Label Software to the Rescue," *Library Mosaics* 2 (May/June 1991): 19.

4. See Computype's Web page. Accessed Jan. 20, 2004, www.computype.com.

5. M. D. McClean, "Advances in Thermal-Transfer Printing Technology," *Surface Coatings International* 77, no. 5 (May 1994): 186–88.

6. Sidney Hill, "Shedding Old Labels," *Manufacturing Systems* 14, no. 8 (Aug. 1996): 22–32. For other descriptions about manufacturing and business applications of thermal-transfer printing, see: Janet Endrijonas, "A Total Solution," *Manufacturing Systems* 14, no. 1 (Aug. 1996): 44; Richard Babyak, "Getting Tough," *Appliance Manufacturer* 51, no. 3 (Mar. 2003): 33–34; David Cartlidge, "Personalizing and Identifying Products with Instant Decals," *Ceramic Industry* 152, no. 10 (Sept. 2002): 30–33; and Maggie DeWitt, "Demand Heats Up for Thermal-Transfer," *Business Forms, Labels and Systems* 39, no. 11 (Nov. 20, 2001): 66–68. Also, ID Technology's Web site provides case histories of thermal-transfer printing applications including a pottery company's UPC labels; a book-distribution facility's UPC and EAN bar code labels (they moved from laser to thermal-transfer labels, reducing label generation and application costs from $.06 per book to $.028 per book); a fragrance packager's switching from a two-label system of applying security tags and UPC labels to a system that could print UPC codes directly onto a security label in one step, providing a 137-percent return on investment for this automated system; and thermal-transfer labels used by a meat processor for inventory tracking and shipping.

7. Several Web sites offer detailed descriptions of thermal-transfer printing technology. See E-Labeling LLC, (www.e-labeling.net/FAQ.htm) and American Printing Converters (www.apclabels.com/TTRibbonProductInfo.htm).

8. For a chart comparing thermal printing with other printing, see American Printing Converters' Web site (www.apclabels.com/TTRibbonProductInfo.htm).

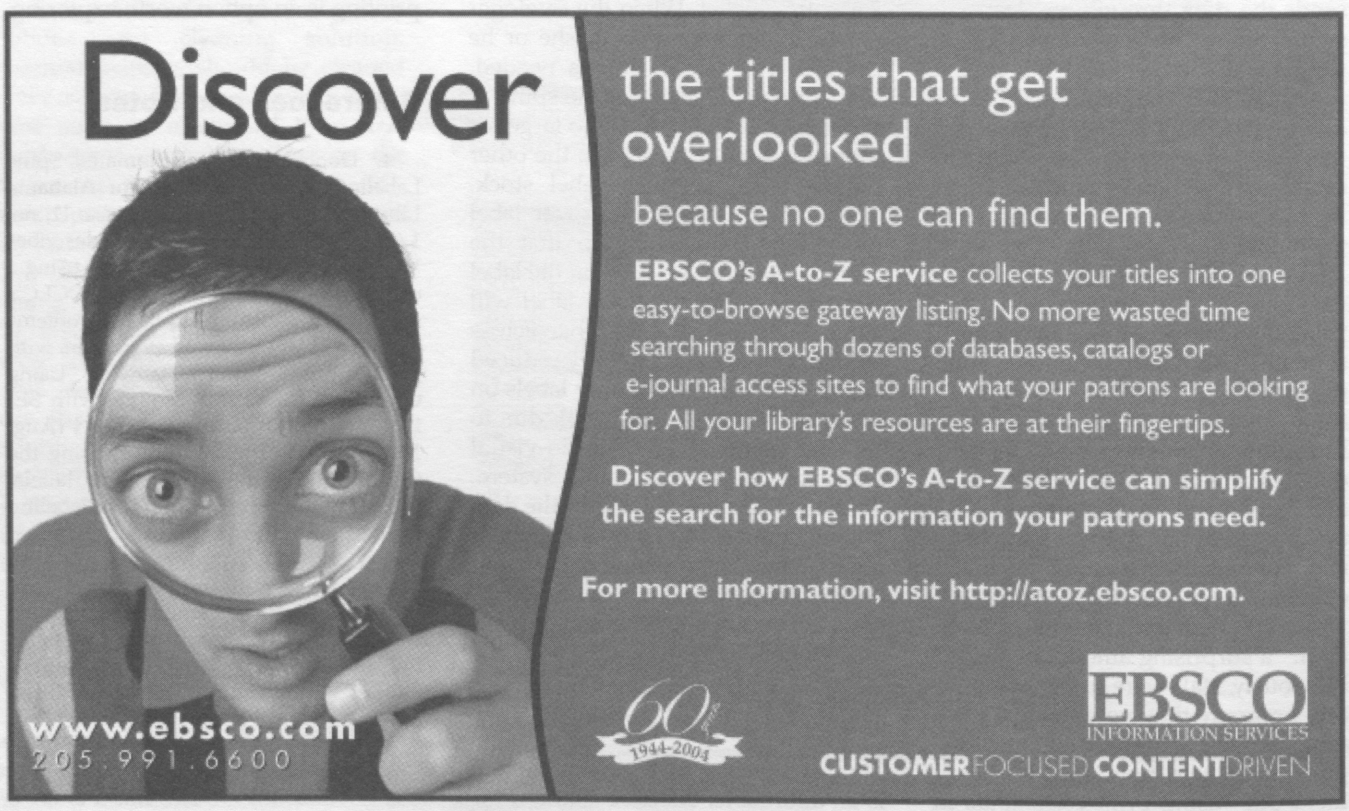
9. Simon Roberts, "New Heights for Printing Quality," *Frontline Solutions (Pan-European ed)* 12, no. 2 (Mar. 2002): 33.

10. McClean, "Advances in Thermal-Transfer Printing Technology," 188.

11. "Getting the Best Out of Your Thermal-Transfer Printing System," *Automatic I.D. News* 14, no. 4 (Apr. 1998): 26–27.

12. TAGPRINTPro (www.connectworld.net/bulkcable/identification4.html); Label Matrix (www.barcodes.com/htmlprod/class?83.html).

13. For example: Bar code Logics (www.barcodes.com); Computype (www.computype.com); Electronic Imaging Materials (www.barcode?labels.com/pagelib.htm); Label Match (www.labelmatch.com); Label Specialties, Inc. (www.labelspec.thomasregister.com); Lancer Labels (www.lancerlabels.co.uk); Product Identification, Inc. (www.prod-id.com/thermaltransfer.htm).



Book Review

Tom Zillner, Editor

Designing with Web Standards

By Jeffrey Zeldman. Indianapolis, Ind.: New Riders, 2003. 456p. $35 (ISBN 0-73571-201-8).

Designing with Web Standards opens with a statement intended to shock: 99.99 percent of today's Web sites are obsolete. The first part of the book explains why and outlines the remedy as Zeldman promotes Web standards with conviction, clarity, and humor.

Tracing the problem to the beginnings of the Web, the author explains that the Web was originally designed for text. HTML was designed to mark up documents structurally, <h> for headings, <p> for paragraphs, and so on. As the Web took off, Web designers began to create visually rich pages. To do this required using HTML in a way it wasn't intended—marking up Web pages to make them look a particular way, rather than for their structure.

The browser wars of the 1990s made things worse. Catering to the desire for visually exciting pages, and attempting to win market share, competing browsers implemented nonstandard, incompatible techniques designed to create visual effects. Pages created for one browser didn't work well—or worse, not at all—in another. This meant Web designers had to create multiple versions of pages in order to achieve the same look in different browsers.

The result was bloated pages that took too long to download, and pages that were difficult and expensive to create and maintain. Sites were often inaccessible to the increasing numbers of people accessing the Web with new types of Internet appliances such as personal digital assistants (PDAs) and Web-enabled cell phones. For persons with disabilities—the visually or motor impaired—these sites were especially inaccessible.

In 2001, a series of browsers began appearing that implemented Web standards that made it possible to design pages with leaner, cleaner coding. As a result, they loaded faster and were easier and cheaper to create and maintain. A core group of technologies designed by the World Wide Web Consortium (W3C), these standards allow designers to create a single version of a site that, with the help of appropriate style sheets, can be accessed by, among other things, multiple browsers, PDAs, and screen readers. Following these standards means not having to solve today's problems by creating others for tomorrow. It also means knowing there won't be some incompatible new technology in the future that will require expensive recoding to incorporate.

In the second part of the book, Zeldman walks us through the process of designing with Web standards by working through two projects. He describes moving from HTML 4.0 to XHTML ("the future of the Web"), using cascading style sheets (CSS) to control the look and layout of a site in a variety of browsers, while making judicious use of JavaScript and DOM (Document Object Model) to control a site's behavior. In this section, he also addresses, among other topics, DOCTYPE switching and Web accessibility.

While Web standards are a major step forward, they haven't solved all the Web's problems, as Zeldman makes clear. Browsers have their quirks and don't implement the standards in exactly the same way. Zeldman discusses the white-space and box-model problem, among others, and shows how to get around them. He devotes an entire chapter to typography. Unfortunately, different browsers don't agree on standard font sizes. Thus, the Web designer must choose between coding font sizes *relatively,* which allows user control but runs the risk that fonts will appear too large or small in particular browsers, or using *absolute* font sizes, which display uniformly in all browsers, but restrict the user's ability to modify the font size in Internet Explorer for Windows, the most used browser.

He ends with a brief section discussing and comparing the various browsers that are compliant with Web standards.

Zeldman is as qualified as anyone to write a book on Web standards. He can speak authoritatively about the history and development of Web standards, as he has been instrumental in promoting their implementation and use.

In 1998, Zeldman cofounded *A List Apart* (www.alistapart.com), a widely used "magazine for people who make Web sites," as well as the Web Standards Project (WaSP) (www.webstandards.org). WaSP was influential in promoting the term Web standards (called "recommendations" by the W3C, the standards-making body for the Web), and in urging browser manufacturers to make their browsers standards-compliant. WaSP has also worked directly and indirectly with companies marketing Web editors, including Macromedia (Dreamweaver) and Adobe (GoLive) in an effort to promote standards-compliance in those products.

His personal Web site, The Daily Report: Jeffrey Zeldman presents Web Design News (www.zeldman.com) is an important resource for the Web-design community. He is owner of a Web-consulting company, Happy Cog, where he applies Web standards in his practice.

Designing with Web Standards is an important book for anyone involved in Web design. It is both a wake-up call and a call to action. It deserves a place next to *Eric Meyer on CSS* and Joe Clark's *Building Accessible Web*

Editor's note: Interested in writing a book review for *ITAL*? Writing a review is a great way to break into print or add to your vita. It's also a great opportunity to share your expertise and opinions with your colleagues. You may review a book of your choice or we can provide you with one. If you are interested or have questions, please contact Tom Zillner (tzillner@wils.wisc.edu).

Sites on the Web designer's bookshelf.—*Thomas E. Lehman, University of Notre Dame Libraries, Indiana*

About Face 2.0

The Essentials of Interaction Design

By Alan Cooper and Robert Reimann. Indianapolis, Ind.: New Wiley Pub., 2003. 540p. $35 (ISBN 0-7645-26413).

Alan Cooper is a big name in user interface design, although he is probably less well known outside the world of software design than, for example, Jakob Nielsen. Perhaps his biggest claim to fame is the design of the initial version of Visual Basic (VB), a wonderful piece of workmanship. One could write an entire essay on the good and evil unleashed upon the world by offering the almost infinitely malleable user interface of VB, but this is not the place. Suffice it to say that software wrought from Alan Cooper's masterpiece is both praised and scorned. But Cooper is not resting on his laurels: he heads a design firm that is turning out first-rate products embodying the best in what he calls "interaction design."

Interaction design is more than just the look and feel of a piece of software; it encompasses both its appearance and behavior. It is defined in *About Face* as "the design of complex, user-focused behaviors of interactive systems" (page 31 of front matter). Software (and Web pages) must do more than just look good; they must also serve their functions well. Too often there is a mismatch between the user interface and the underlying functioning of software: either the interface is seductive and the underlying function is a morass, or, more commonly, an elegant (or at least useful) piece of software is hard to use and unintuitive because of lousy interface design. Why? Because for most developers, design of the user interface is secondary to design of the underlying code.

The actual interaction with the user is often phoned in at the end of a project—an afterthought. Cooper believes that user goals are paramount, and user goals are ill met by clever coding wrapped in ornate but unusable screen design. He presents a process called Goal-Directed Design (GDD) that starts with users and ends with them. The stages of GDD are research, modeling, requirements definition, framework definition, and refinement. *About Face* first provides a broad overview of each of these stages before moving on to more in-depth coverage of each.

Research, the first stage of interaction design, focuses most heavily on user goals, as revealed through ethnographic techniques that meet users in the wild, as they go about their working tasks. This observation is informed by some initial hypotheses about the behavior variables representing a spectrum that can distinguish various users. The authors point to the example of an e-commerce application (47). Among the possible variables are frequency of shopping (infrequent to frequent), desire to shop (hates-to-shop to loves-to-shop), and motivation to shop (bargain hunting to searching for just the right item). This articulation of behavior variables will be used in the next stage of the design process, but it is first employed here in order to ensure that the users who are interviewed exhibit behavior across the full range of variables.

Cooper and Reimann believe that it is crucial to observe users as they work and to interview them in the work context. So if they were designing an interface to an online catalog, they would watch and interview people interacting with the current catalog software. A key to the usefulness of the interviews is to find users who fit all of the various roles possible in the interaction. For example, users of a catalog might be casual beginners, expert researchers, or almost anywhere between these poles. It is important to capture the full range of possible user stances in order to ensure a good design.

Actual observation interviews are not set up until there is an initial "persona hypothesis," a preliminary idea of who uses the product, how their needs and behaviors vary, and what behaviors and environments need to be explored. This and other behavioral and environmental information are used to put together the interviews. The observation and interviewing must focus on goals rather than tasks. After all, the software should help the user achieve her goals, rather than simply providing her with more efficient or otherwise better ways to perform tasks. Tasks are means to achieve goals, not ends in themselves. The authors provide useful tips for conducting the interviews (48–52), although they do not really cover the concomitant techniques of observation. Other types of research (focus groups, marketing studies, and usability and user testing) are mentioned but given short shrift. It's the interviewing that sets the stage and provides the underpinning for the next stage, modeling users.

Cooper and Reimann call modeling "the most powerful interaction design tool" they use (55). It is here that the research conducted in the previous phase is codified and formalized through the construction of personas (or, if you prefer it more pedantic, personae). A persona is, as its name suggests, a set of goals and behavior patterns that is represented as a single person. How are personas constructed? First of all, the behavioral variables hypothesized in the research stage are examined in light of the user interviews. There may be some adjustment to their characteristics or gradation based on what was actually observed. Perhaps a few new variables were discovered along the way. If enough unexpected variables emerge, it may be necessary to perform some additional interviews with these in mind. In any case, at some point a set of behavioral variables emerge,

against which each of the interviewees has been measured. The next step is to map every subject's responses across the behavioral variables, looking for behavior patterns. These are evidenced by clusters of subjects falling on the same point in the spectrum of behavior across multiple variables. Using the authors' e-commerce example, perhaps multiple users would say they love to shop, do it all of the time, and are looking for bargains. This cluster could then be used as the raw material for a persona.

The persona itself incorporates the behavior patterns and also includes goals discovered during the research process, either through observation or targeted questions. Then the persona narrative is written. "The best narrative quickly introduces the person in terms of his job or lifestyle . . . and briefly sketches a day in his life, including peeves, concerns, and interests that have direct bearing on the product" (71). Cooper and Reimann believe it is important to name the persona and provide a photograph. In other words, this must seem like a sketch of a real person. Unfortunately, the authors provide no example of a persona narrative. This is an aspect of a larger problem that will be discussed later.

Next comes what the authors call "framework definition," (or "build the interaction design"). A set of scenarios is developed as a prelude to the actual design. The scenarios are stories about the people depicted in the personas solving the problems and meeting the needs articulated during the research phase. It's actually not quite as simple as that: there is a five-stage process from which a "context" scenario emerges. Context scenarios are stories about a day in the life of the users (personas) and how they interact with the system under design. (A good example of one of these scenarios can be found on pages 81–82.) The context scenario is not an end product; it leads into a needs-identification stage. This is an extraction process. The context scenario forms the input to the needs identification stage. The resulting user needs are expressed in terms of objects and actions.

The needs identified serve as input to the design stage itself—"defining the interaction framework" (83–88)—again a five-step process. One of the outputs of this stage is another scenario. The "key-path" scenario takes the desired interactions identified in the context scenario and spells out the actual tasks (actions and pathways) that will be embodied in the design product. It is during this set of stages that sketches and storyboards of the actual user interface finally emerge. The key-path scenarios are worked on iteratively, in top-down fashion, becoming more detailed as the design is elaborated, starting with key functions and moving progressively toward less important features off the beaten path. At the same time, the principles of good design presented elsewhere throughout the book are brought to bear, approaching the design in a bottom-up fashion. So the central core of the design stage is driven from both the top down and the bottom up, obtaining the best of both worlds.

Finally, there is refinement. The least coverage is spent on this in the first chapters, partly because much of the rest of the book is spent on tips and techniques that constitute a key piece of the refinement stage. However, this is also a stage at which another set of scenarios, "validation scenarios," are constructed (89). These are simply scenarios for "less-traveled interactions," and use the same method of construction as key-path scenarios.

It's significant to note that all of these stages of interaction design are iterative. Although it is most obvious, perhaps, in the design and refinement stages, at each of the other earlier stages there is a back-and-forth honing process. This does not seem to be stressed by the authors, but it is a key to ultimate success of the design.

The preceding has been a more-or-less detailed exposition of the stages of interactive design. Yet this only constitutes the first ninety pages of the book. And that's a problem with *About Face 2.0*—the design process is spelled out with just enough detail to convince the reader that this could be a useful process that should be incorporated into the activities of any software design team, yet there does not seem to be enough detail to recreate it effectively. One thing sorely missing is an entire case study, laying out the documents created at each step, the research, personas, and scenarios that embody the implementation of interaction design. The few examples that are provided of small pieces in the larger scheme—for example, the sample context scenario (81–82)—leave the reader hungry for more. A case study for a single piece of software, with each of the example pieces placed in the context of the concepts as they are spelled out, rather than presenting the whole of a case study as a separate lump in a separate chapter or appendix, would be the most helpful and effective method to make practitioners out of those who are believers. But this reviewer would settle for the lump version instead—anything that illuminates and elucidates interaction design as Cooper and Reimann envision it.

If the exposition of interaction design is the meat, the remaining sections and chapters are the potatoes. Although potatoes are denigrated in the Atkins age, in this book these chapters provide plenty of nutrition. One theme that echoes throughout becomes something of a hobbyhorse, but an important one: don't make the user feel dumb, especially through seemingly meaningless dialog boxes. Good examples are offered in chapters 10 ("Eliminating Excise") and 13 ("Rethinking Files and Save"). The authors make important points, particularly that dialog boxes often offer the implicit message "You have made a mistake; you are stupid." Chapter 13 ("Eliminating Errors") deals with this most directly and at greatest length, offering methods to avoid errors and resultant error messages. Better and harder work by programmers is the

key. One failure in this discussion: although Cooper and Reimann loathe error dialogs, they do offer an example of how they should look in case all other error suppression and correction actions fail. The example (444) is awful: a huge box with way too much information. Most users may find all of this overwhelming rather than helpful. While it attempts to be humane, it is humane with a vengeance. Still, the idea that most error dialogs can be avoided altogether is important and may be achievable to some degree.

An entire section of the book is devoted to direct manipulation and its permutations (part 5, "Mice and Manipulation"). There is a quite helpful discussion of the power of the mouse and the intricacies of the operations it can perform. There are separate chapters devoted to "Selection" and "Drag-and-Drop" (chapters 22 and 23), as well as a quite good chapter on manipulating objects (chapter 24). All of the material outlining direct manipulation material in relation to good design is first-rate. Even the coverage of 3-D manipulation is interesting.

Another whole section is devoted to "Controls and Their Behaviors," offering important insights and tips for dealing with windows, menus, toolbars and tooltips, and dialogs. Cooper and Reimann are particularly fond of tooltips, alluding to them in passing in multiple chapters, but more clearly spelling out their use and superiority to balloon help in pages 384–86. As elsewhere in the book, the material in this section is clear and thoughtful.

There are asides concerning Web applications throughout *About Face 2.0*, but there is also a fairly short chapter devoted exclusively to the Web (chapter 37, "Designing for the Web"). The authors first distinguish between Web sites and Web applications. Web sites are information-centric and users simply navigate or search the site, finding the specific pieces of information they seek. More of the chapter is devoted to Web applications, which are characterized by their heavily transactional nature. Because of this, Web applications are more solidly in the applications world than the Web world. The chapter's treatment of Web applications mainly serves to connect up concepts presented throughout the book with the specific functionality of the Web. For example, "sovereign-posture" applications are introduced much earlier as "programs that are best used full-screen, monopolizing the user's attention for a long period of time" (103). A word processor like Microsoft Word is a good example of such a desktop application. In this chapter, the authors discuss how sovereign Web applications might be realized through a Web-browser interface although they favor a non-browser-based implementation. Similarly, specific sorts of sites are connected to previously covered concepts. While the chapter is too short, particularly on specific examples, to make it a primary resource for designers of Web interaction, it does illustrate that the key concepts discussed elsewhere in the book can be brought to bear on Web-design projects. It's not simply about the appearance of static pages anymore. Sophisticated Web applications require sophisticated thinking. The chapter is helpful, and Cooper and Reimann provide pointers to several other books that have Web design as a primary focus.

When all is said and done, what Cooper and Reimann reveal again and again in this book is that "design concerns the whole of the product or service—not just what it does, but what it is and what goals it serves" (480). The user interface is, as the term suggests, where the user connects with the behavior of the software. Design of user interaction encompasses far more than cosmetic changes to the user interface. This is not something that can be grafted on at the end, after the underlying software has been written. Users' goals must be met, and the best way to meet them may be to design software from the outside in. Turning the coders loose certainly hasn't done the job.

About Face 2.0 is a book worth reading. Even though this reviewer has serious reservations about the chapters that spell out the interaction-design method, much can be learned from them. Additionally, the rest of the book is chock-full of important insights and useful examples. Although, as noted, Web-based interaction-design methods are not covered at great length, the principles spelled out here can be usefully applied in constructing Web pages and sites, particularly those that are transaction-rich.

Read this book if you develop software or Web applications. Read the first section in its entirety. Consider reading the remainder in its entirety as well, but if you don't have time, dip into the function-specific chapters as you need to. It's worth the time.—*Tom Zillner*

Index to Advertisers

Art & Tech	20	LITA	cover 2, 2, cover 3
EBSCO	36	Sea Change Corp.	15
InfoUSA	cover 4		

LITA ANNUAL CONFERENCE HIGHLIGHTS

LITA PRESIDENT'S PROGRAM

Monday, June 28, 2004, 2–4 P.M.

It's a Small World After All!

Three innovators from the University of Maryland's Human-Computer Interaction Lab will present and discuss the International Children's Digital Library (ICDL), a large-scale digital archive of literature for children ages three to thirteen. Faculty from the College of Information Studies and the Computer Science Department are collaborating to create a digitized children's library that will grow to more than 10,000 books available in 100 languages.

Speakers: Benjamin B. Bederson, Assistant Professor of Computer Science and Director of the Human-Computer Interaction Lab at the Institute for Advanced Computer Studies; Allison Druin, Assistant Professor in the College of Information Studies and member of the Human-Computer Interaction Lab; Ann Carlson Weeks, Professor of the Practice in the College of Information Studies.

PRECONFERENCES

For more information or to register, please visit www.lita.org/ala/lita/litaevents/annual2004orlandofl/LITAPreconfs.htm.

CIPA: Internet Filtering Using Open Source

9 A.M.–5:30 P.M.

Registration opens at 8:30 A.M.

The focus of this program is to demonstrate the installation of filtering solutions, in the light of the CIPA rules. While many commercial software solutions are readily available, there are also open source and free versions that are effective and can be customized to the specification of each of the libraries. This program will show the technical aspects of open source filtering software capabilities, and how to use the products.

Speaker: Cindy Murdock, Network Administrator, Meadville (Pa.) Public Library

Presented by the LITA Open Source Systems Interest Group and co-sponsored by the Public Library Association.

Usability Issues in Metasearch Interface Design

2–5 P.M.

Registration opens at 1:30 P.M.

A discussion of some of the user interface issues that are specific to the world of metasearch, including resource discovery, displaying different data formats in a single display, relevance ranking, presenting merged records in a understandable format, etc. The preconference will address issues of concern common to the library community in general, but will be geared toward members currently or soon to be dealing with implementation at their own organizations.

Speakers: Brenda Bailey-Hainer, Director of Networking and Resource Sharing, Colorado State Library; Stanley Wilder, Assoc. Dean, River Campus Libraries, University of Rochester; Oliver Pesch, Chief Architect, EBSCO Publishing; Ezra Schwartz, President, Art & Tech, Inc.

Presented by the LITA Program Planning Committee.

Portals in Libraries: A Symposium

9 A.M.–5:30 P.M.

Registration opens at 8:30 A.M.

Portals are transformational environments that address the problem of information glut by customizing information content to meet end-user needs. Libraries of all types are planning services they call portals. This preconference's mission is to enhance the importance of portals for libraries. Participants will depart with a good understanding of portal implementation issues, observe best practices, and learn about standards, interoperability issues, identity management, and more.

Speakers: Howard Strauss, Princeton University; Kathy Scardellato, Toronto Public Library; Hugh Jarvis, Gemma DeVinny, University of Buffalo; Roy Tennant, CDL; Art Rhyno, University of Windsor; Eric Lease Morgan, University of Notre Dame

Presented by the LITA Internet Portals Interest Group.

LITA

> "In a library such as ours... ReferenceUSA is among the most in-demand databases because of its scope and accuracy."
>
> *John Ganly*
> *Assistant Director*
> *New York Public Library*

Our reference information in print, on CD-ROM and via the internet gets used. Why? Because everybody needs our reference products for everyday life. For example:

- CONSUMERS
- PROFESSORS
- SALES PEOPLE
- LAW LIBRARIES
- ENTREPRENEURS
- STUDENTS
- JOB SEEKERS
- NEWS ORGANIZATIONS

*info*USA Inc.'s databases contain company name, address, phone number, SIC/NAICS codes, estimated sales volume, number of employees, credit rating code*, key executive names, and more. Virtually all U.S. businesses - public and private.

Internet Access – *Reference*USA® Enjoy real-time access to the world's finest database of 14 million businesses and 250 million U.S. residents with comprehensive coverage, accuracy and instant access to information. And that's just the beginning! With its power, speed and user-friendly interface, *Reference*USA becomes the most effective, easy-to-use reference tool available. Customized reports, immediate downloads and business profiles are a snap!

American Business Disc® is the ultimate reference tool for Libraries. Imagine having over 14 million businesses on a single CD-ROM! Search by Company Name, Address, SIC Code(s), Yellow Page Heading, Sales Volume, Number of Employees, Geographically by ZIP Code(s), City(ies), State(s), or even the entire U.S.

PowerFinder® contains more than 115 million U.S. businesses and residential listings, 800 numbers, fax numbers Web site URLs and census information. This set of six regional CD-ROMs gives you instant access to virtually every business and nearly every resident in the United States.

American Manufacturers Directory & CD-ROM contain every manufacturer in the U.S. with 20 or more employees, more than 168,000 total listings – all in two volumes. The CD-ROM contains all 645,000 manufacturers.

Big Businesses Directory & CD-ROM include America's largest companies, 193,000 companies that employ 100 or more people. This directory also includes 703,000 top Executives and Directors.

State Business Directories are the most popular directories in every reference department. Before your patrons research any other source, point them to our publications. Four easy-to-use sections include: Businesses by City, Businesses by Yellow Page Category, Major Employers, and Manufacturers by City and Product.

"In a library such as the New York Public Library-Science Industry and Business Library which serves over a thousand users each day, a resource such as ReferenceUSA is among the most in-demand databases because of its scope and accuracy."

John Ganly
Assistant Director
New York Public Library

ReferenceUSA℠
A division of *info*USA

For More Information Call Toll-Free:
(866) 813-3009
or E-mail: reference@infoUSA.com

Friends of Libraries U.S.A.

5711 South 86th Circle • P.O. Box 27347 • Omaha, Nebraska 68127
Phone: (402) 593-4523 • Fax: (402) 596-7688 • E-Mail: reference@infoUSA.com

* *Our Credit Rating Codes are indicators of probable ability to pay. They are based on business demographic factors such as number of employees, years in business, industry stability, bill paying history, barriers to entry, and government data. We recommend that these ratings be used primarily as a starting point and should not be the sole factor used in making a credit decision. You must obtain more information from bank and trade references, local credit bureaus, or other sources before extending credit. We will not be liable for any losses resulting from the use of this information.*

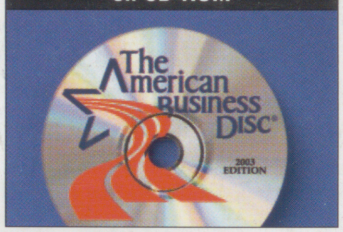
Internet Access

ReferenceUSA-online database

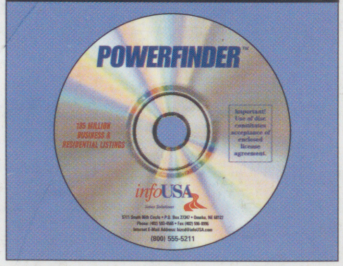


In Print and on CD-ROM

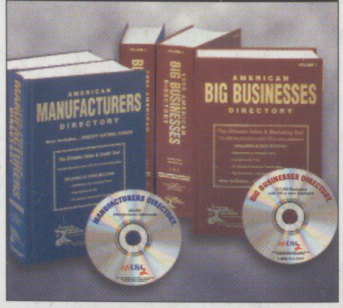
Over 14 Million Businesses on CD-ROM



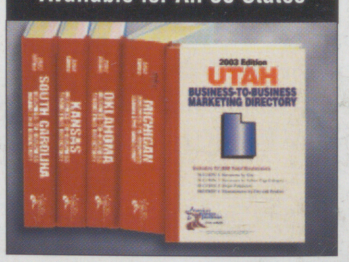
More Than 115 Million U.S. Business & Residential Listings



The Best Directories of America's Top Companies



Available for All 50 States



#23IT4