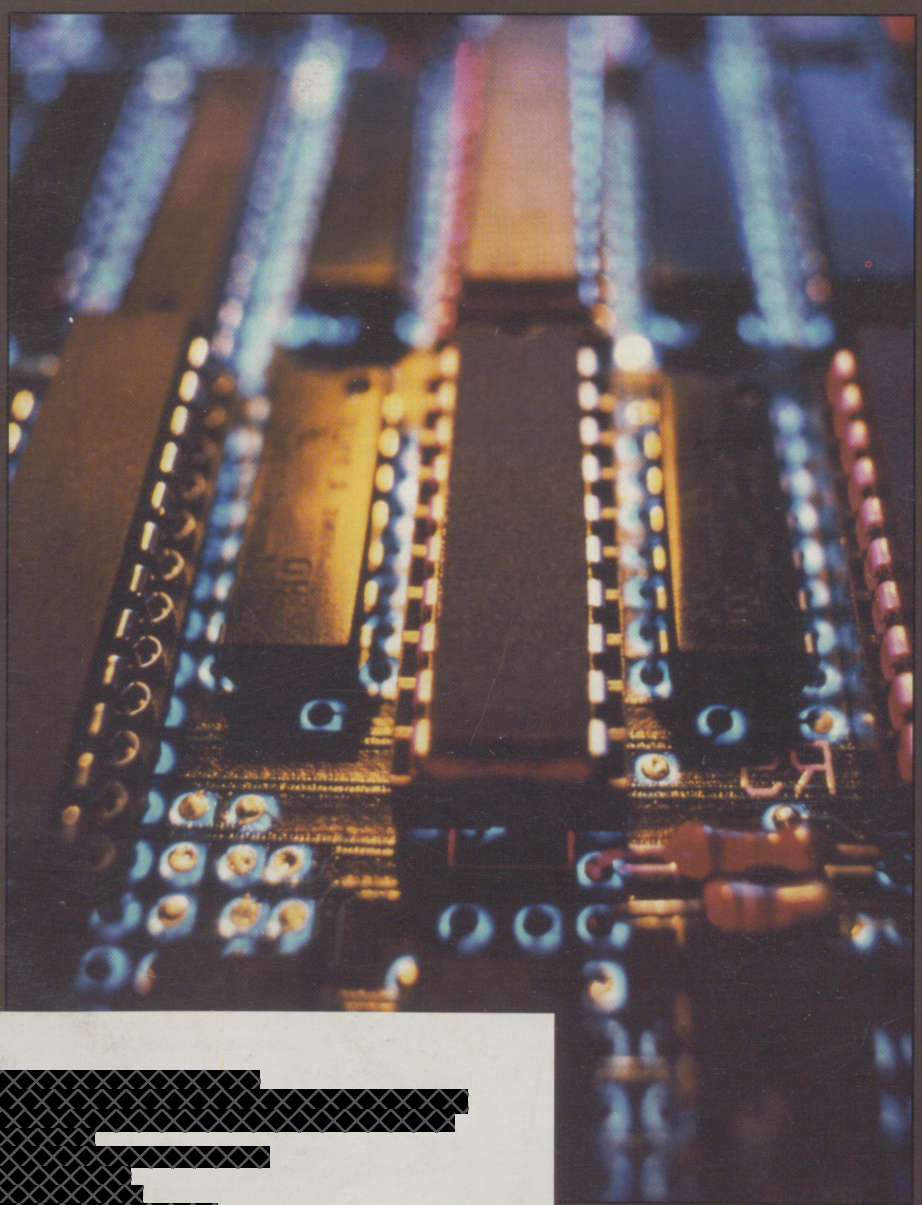
Information Technology AND Libraries

ISSN 0730-9295

Volume 23 Number 3

SEPTEMBER 2004



IN THIS ISSUE

- Retrieval of Chinese Language Titles in Pinyin
- Analyzing Web-Site Rank in Google
- Providing Access to Foreign Language Electronic Resources

AMERICAN LIBRARY ASSOCIATION

BOSTON 5

MIDWINTER MEETING

January 14, 2005 in Boston!

LITA presents two of our newest Regional Institutes as pre-Midwinter Meeting Workshops!

Establishing an Institutional Repository
Susan Gibbons, Presenter

Librarians considering or in the early stages of establishing an institutional repository will want to attend this day-long look at the technical, policy and human factors related to establishing an institutional repository (IR). Expanding on the presenter's Library Technology Report of the same title (July/August 2004), the institute will reflect the current state of IR development and best practices, as well as the lessons learned by early adopters. Susan Gibbons is assistant dean for Public Services & Collection Development at the University of Rochester, River Campus Libraries. She has been involved with the University of Rochester's institutional repository project since 2000, when she was the digital initiatives librarian.

Wireless Networks in Libraries
Marshall Breeding, Presenter

How can you incorporate wireless networks to enhance library services and provide more convenient access to electronic resources? This workshop provides in-depth, practical information on wireless networking technologies, including the technical underpinnings of wireless networking and general networking concepts that help put wireless networks into perspective. Participants will learn to separate the hype from the realistic expectations of the benefits of wireless technologies. Presented by Marshall Breeding, Library Technology Officer at Vanderbilt University and regular participant in LITA's Top Technology Trends program.

Both workshops take place January 14, 2005, from 9:00 A.M.–4:30 P.M.

lita

Registration fees:

LITA Member:	$ 195.00
ALA Member:	$ 255.00
Non-Member:	$ 305.00
Student/Retired:	$ 195.00

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: **John Webb,** Washington State University Libraries; **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; **Nancy Colyar,** Louisiana State University, 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 92 of the June 2004 issue or on the LITA Web site at www.lita.org/ital. Manuscripts of feature articles, communications, and tutorials should be addressed to John Webb, Assistant Director for Digital Services/Collections, Washington State University Libraries, Pullman, WA 99164-5610; jwebb@wsu.edu . 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, Angela Hanshaw; Stephanie Kuenn, Kristen McKulski, Karen Sheets, and Christine Velez), 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.∞

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.

Information Technology AND Libraries

Volume 23, Number 3 September 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.

95 Retrieval of Chinese Language Titles in Pinyin: A Comparative Study
JIE HUANG

102 Applying Geographic Information Systems to the Weber County Library System
JANAE KINIKIN

108 Jump Higher: Analyzing Web-Site Rank in Google
LISA ZHAO

COMMUNICATIONS

119 Providing Access to Foreign Language Electronic Resources
EILEEN LLONA, EMALEE CRAFT, KEIKO YOKOTA-CARTER, AND DAVID PHAM

123 Will Libraries' Web-based Survey Methods Replace Existing Non-Electronic Survey Methods?
GAY HELEN PERKINS

TUTORIAL

128 Segregating Library Users in a Microsoft Windows Client/Server Environment to Control Access to Public Printers
PERRY C. HORNER AND DENNIS ISBELL

133 Software Reviews

136 Index to Advertisers

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

LAMA presents

Creating Library Leaders for the 21st Century

What It Takes to Take the Lead

3RD NATIONAL INSTITUTE

November 18–20, 2004

Wyndham Palm Springs Hotel PALM SPRINGS, CALIFORNIA

Engage in two days of learning and dialog focusing on personal and organizational leadership:

- Leading from all organizational levels
- Becoming an effective change agent
- Developing your leadership capacity
- Transferring your leadership skills to new environments

KEYNOTE SPEAKER



Carla D. Hayden, Ph.D., 2004 American Library Association President
Dr. Hayden, Executive Director of the Enoch Pratt Free Library, Baltimore, Md., is a noted library leader and educator. Her strong and articulate voicing of ALA's opposition to sections of the USA PATRIOT Act led to her being honored by *Ms. Magazine* as one of its Women of the Year for 2003.

For more information, visit **www.ala.org/lama.**
Register online, by fax, or by mail.
Advanced registration ends October 22!

Major Sponsor
3M Library Systems
Sponsors
OCLC
ProQuest
Supporter
Jossey-Bass/John Wiley & Sons

INSTITUTE FACULTY

Camila Alire, Ph.D., 2005 ACRL President-Elect and Dean of Library Services, University of New Mexico General Libraries, Albuquerque, "Leading During Traumatic Times."

Luis Herrera, 2004 Public Library Association President and Director, Pasadena (Calif.) Public Library, "Leading Your Superiors and Your External Community—Infusing Values."

John Lubans Jr., Visiting Professor, School of Library and Information Sciences, North Carolina Central University, Durham, "Leading Your Organization—Infusing Values and Building a Leadership Infrastructure."

Donald E. Riggs, Ph.D., Vice President of Information Services, Nova Southeastern University, Ft. Lauderdale, Fla., "Leadership Styles and Techniques."

Julie Todaro, Dean of Library Services, Austin (Tex.) Community College, "Transferring and Adapting Leadership to New Environments."

Pat Wheeler, Manager, Six Sigma Master Black Belt, 3M Library Systems, St. Paul, Minn., "3M Six Sigma—Achieving Breakthrough Performance for Our Customers, Shareholders, and Employees."

Retrieval of Chinese Language Titles in Pinyin: A Comparative Study

Jie Huang

This study attempts to investigate the Pinyin retrieval of Chinese language materials within online catalog databases. The Peking University Library (PKUL) database is studied in comparison with those of the Online Computer Library Center (OCLC) and the Research Libraries Information Network (RLIN). It is found that both title-keyword and known-title searches at PKUL yield better results because of its fundamental difference in retrieving bibliographic records in Pinyin. It adopts a non-segmented string match in retrieval. In so doing, it greatly raises the precision rate without negatively affecting the recall rate, and bypasses the controversy over, and potential human errors in, word division in cataloging and retrieving.

The conversion from Wade-Giles to Pinyin of Chinese language bibliographic records has finally been accomplished in North American libraries. Now the question facing the library community is how to improve the current way of handling the Pinyin records so as to provide a better service for the end users of those records. With this question in mind, the author turned to China, where the Pinyin romanization was originally invented, and compared the database of Peking (Beijing) University Library (PKUL) with those of the Online Computer Library Center (OCLC) and the Research Libraries Information Network (RLIN). The findings of this study point to a way in which the North American library community could improve its service of Chinese language bibliographic records to its patrons. The two sections following the literature review report the study comparing OCLC and RLIN with PKUL respectively in title-keyword and known-title searches. The concluding section presents the summary and implications of the study.

Background and Literature Review

For decades, the issue of word division kept the Library of Congress (LC) from converting Chinese romanization from Wade-Giles to Pinyin. In Pinyin, where the Chinese language is romanized with alphabetic letters, the normal practice is to separate words from each other with a space, as opposed to Wade-Giles, which separates all the individual syllables (each represented by a Chinese character in writing) within and between words with a space. The rationale behind aggregating syllables into lexical units in Pinyin is to reduce ambiguity resulting from homophony in the Chinese language. There are as many as forty-seven thousand to eighty-five thousand distinct characters in written Chinese, which share only around thirteen hundred syllables, including tonal distinctions, in spoken Chinese.[1] When tonal distinctions represented by diacritical marks are eliminated, which is the common practice in the library community, the number of base syllables is reduced to approximately 410.[2] The many-to-one relationship between characters and syllables in Chinese leads to great homophonous ambiguity as Chinese characters are romanized into Pinyin representation, especially after the tonal distinctions are excluded. Homophonous ambiguity, however, is solved 95 percent of the time when individual syllables are aggregated into word units.[3] The problem with syllabic aggregation is, on the other hand, that word boundaries are not always clear in Chinese. Sometimes differing interpretations lead to different segmentations of lexical units. Therefore, there arises the issue of word division in the Pinyin romanization.[4]

In the absence of an international standard on word division, and out of concern over inconsistencies in Pinyin aggregation in creating and retrieving records, LC decided to separate all individual syllables from each other except in the cases of personal names, geographic locations, and certain proper nouns.[5] A strong argument against the monosyllabic format (i.e., syllable division) consists in the fact that, as mentioned earlier, it greatly increases ambiguity caused by homophony and thus reduces readability of Pinyin records.[6] However, the large catalog databases such as OCLC and RLIN now offer the Show Vernacular option to display the character version of Chinese records. These days, with the widespread introduction of personal computers and their ability to display fonts in a wide range of vernacular scripts, readability of Pinyin records in monosyllabic format is no longer a serious problem on the assumption that the patrons who search Chinese records can read Chinese characters, as well as Pinyin romanization.

Arsenault has investigated the issue of word division in relation with retrieval efficiency (shorter complete-task time) and effectiveness (higher success rate of retrieving items sought) in OPAC known-title searches.[7] His experimental studies suggest that aggregation of monosyllables into lexical units does improve retrieval efficiency, especially in keyword searches, while effectiveness remains mainly unaffected. His research, however, also shows that even native-speaker subjects vary in their interpretation of how syllables should be aggregated into lexical units. The variation leads to a marked increase in the average number of queries required to

Jie Huang (lilyh@ou.edu) is Cataloger and Assistant Professor in the Catalog Department at University of Oklahoma Libraries, Norman.

find each record, although there is no noticeable drop in the success rate (the number of items found). Arsenault's findings suggest that the benefits gained from aggregation outweigh the inconvenience generated by the inconsistencies in aggregation format in cataloging and retrieving. He argues that it would be unreasonable to dismiss the use of aggregated Pinyin simply because consistency is difficult to achieve.[8]

The author attempted to find out how the libraries in China approach the issue of word division.[9] PKUL, the largest university library in Asia according to its Web site, was selected for this purpose, and its database was studied in comparison with those of OCLC and RLIN. At PKUL, readability related to word division in Pinyin is not an issue since all the Chinese language records are displayed exclusively in vernacular script of Chinese characters. Pinyin only serves as a retrieving mode in conjunction with character-based search modes. The most remarkable finding is that PKUL bypasses the problem of Pinyin aggregation in search-term inputting and yields much better search results than OCLC and RLIN.[10]

Title-Keyword Searches

This section reports an example of comparative studies in title-keyword searches to illustrate how PKUL is more user-friendly than OCLC and RLIN. For this study the keyword searched is *shuǐdào* 水稻 "paddy rice," a compound word composed of two syllables. The author conducted title-keyword searches respectively in OCLC, RLIN, and PKUL by using both polysyllabic and monosyllabic input methods. A total of 1,264 records were retrieved in six searches, and the results were examined to distinguish relevant from irrelevant records. The criterion for determining if a record is relevant or irrelevant is to see if the record title contains the keyword searched. The examples in figure 1 illustrate relevant and irrelevant records. The first line presents a title divided into syllables, with each syllable separated from the next by a space. The second line presents the same title with syllables aggregated into words. Beneath it, the third line provides a word-by-word gloss and the fourth and last line is an English translation in double quotes. The bold type highlights the parts where homophony takes place.

Figure 1a is a title about the cultivation of paddy rice, so it is a relevant record. Figure 1b *shuidao* means "watercourses" or "waterways," which is homophonous with the word that means "paddy rice." It is therefore retrieved but is an irrelevant record. In figure 1c, *shui* "tax" is part of a compound *fushui* "taxation" whereas *dao* "introductory" is a part of another compound *daolun* "introduction." Thus, it is again an irrelevant record. Likewise, the titles in figure 1d–f all contain syllables homophonous with those of *shuidao* "paddy rice"; they are all irrelevant records too.

After relevant records were distinguished from irrelevant ones, the percentages for recall and precision were calculated. The search results obtained from the three databases for the keyword *shuidao* or *shui dao* "paddy rice" are presented in table 1. "OCLC *shuidao*" yields a 100 percent for precision, but its recall percentage is very

a. ***Shui dao*** *zai pei*
Shuidao *zaipei*
paddy-rice cultivation
"The cultivation of paddy rice" (水稻栽培)

b. *Ru hai* ***shui dao*** *ji hua*
Ru hai ***shuidao*** *jihua*
into ocean water-courses plan
"Plan for watercourses into the ocean" (水稻栽培)

c. *Fu* ***shui dao*** *lun*
*Fu****shui dao****lun*
taxation introduction
"An introduction to taxation" (赋税导论)

d. ***Shui*** *li dian li ji shu bao* ***dao***
Shui*li dianli jishu bao****dao***
hydraulic electric technology report
"Report on hydraulic and electric technology"
(水利电力技术报道)

e. *Cong huo* ***dao shui***
Cong huo ***dao shui***
from fire to water
"From fire to water" (从火到水)

f. *Wo bu zhi* ***dao*** *wo shi* ***shui***
*Wo bu zhi****dao*** *wo shi* ***shui***
I not know I am who
"I don't know who I am" (我不知道我是谁)

Figure 1. Examples of Relevant and Irrelevant Records

Table 1. Results for Searches on *Shuidao* and *Shui Dao* in the Three Databases

Database/ Keyword	Retrieved	Relevant	Precision %	Recall %
OCLC *shuidao*	13	13	13/13=100.0%	13/200=6.5%
OCLC *shui dao*	505	200	200/505=39.6%	100.0%
RLIN *shuidao*	127	98	98/127=77.2%	98/112=87.5%
RLIN *shui dao*	481	112	112/481=23.3%	100.0%
PKUL *shuidao*	69	63	63/69=91.3%	100.0%
PKUL shui dao	69	63	63/69=91.3%	100.0%

poor, only 6.5 percent, which is derived from the total retrieved (13) divided by the total relevant in the database (200). Note that the total relevant in the database (200) is derived from the total retrieved for "OCLC *shui dao*" (505) minus the total irrelevant retrieved (305).

It is worth noting here that the Chinese records in OCLC are cataloged with non-aggregated Pinyin. However, thirteen records were retrieved after the search on *shuidao* (two syllables aggregated into one unit) was conducted. Examining all thirteen records revealed that they all include an Other Titles field that is rendered in aggregated Pinyin, while their Title proper is rendered in non-aggregated Pinyin. That is why they were retrieved after the aggregated *shuidao* was entered.

In RLIN, as shown in table 1, the aggregated *shuidao* yielded a total of 127 retrieved records, ninety-eight of them being relevant. Precision reaches 77.2 percent while recall is as high as 87.5 percent. The relatively high percentages in both recall and precision with RLIN is due to the fact that "a great proportion of the Chinese-language records in that database contain aggregator characters in the Romanized fields."[11] It is worth noting that when the two syllables were separated in another search in RLIN, although recall improved from 87.5 percent (ninety-eight relevant records) to 100 percent (112 relevant records), precision fell considerably from 77.2 percent to 23.3 percent. That is to say, one has to go through 481 records to find 112 of them relevant. This result shows that the argument that non-aggregated titles would provide more access points and, therefore, have a greater possibility of being found is actually flawed, because it really means destroying the balance between recall and precision, resulting in impractically large numbers of titles retrieved with poor precision rate.[12] It is therefore not user-friendly at all.

The most remarkable finding is that, with PKUL, the results remained exactly the same whether the two syllables were joined (*shuidao*) or separated (*shui dao*). In both cases, the total retrieved is sixty-nine, with sixty-three of them being relevant. Thus, both recall and precision percentages are very high, 100 percent and 91.3 percent respectively. These results show that, with this database, it does not matter whether the end user inputs the Pinyin keyword with polysyllabic or monosyllabic method.

A question to ask at this point is: Why does PKUL yield much higher recall and precision percentages than OCLC and RLIN, ignoring the issue of aggregation in entering keywords in title-keyword searches? Readers are again referred to figure 1, where the bold-typed parts in the example titles can be abstracted into the configuration in figure 2, characterized by presence or absence of two features: sequence and adjacency. If X precedes Y, then they are sequential (Sequence: Yes). If, instead, Y precedes X, then they are not sequential (Sequence: No). Adjacency here means X and Y are not separated by another element or other elements (XZY), excluding the space.

Configuration	Sequence	Adjacency
a. XY (= XY1)	Yes	Yes
b. XY (= XY2)	Yes	Yes
c. X Y (= WX YZ)	Yes	Yes
d. X...Y	Yes	No
e. Y X	No	Yes
f. Y...X	No	No

Figure 2. Configuration by Sequence and Adjacency for Figure 1

This configuration of X and Y is interpreted as follows. In figure 2a, "XY" or "XY1" stands for *shuidao* "paddy rice" in figure 1a, which is a correct target for retrieval. In figure 2b, "XY" or "XY2" stands for *shuidao* "waterway; watercourse; water route" in figure 1b, which is not a correct target for retrieval but is a "reasonable error" because X and Y are both sequential and adjacent. In figure 2c, "X Y" or, more exactly, "WX YZ" stands for *fushui daolun* "taxation introduction" in figure 1c, which is not a correct target for retrieval either but is again a "reasonable error" because X and Y are both sequential and adjacent (ignoring the space). In figure 2d–f, however, the combinations of X and Y are not correct targets for retrieval, and they are "unreasonable errors" due to the absence of sequence and adjacency.

In both OCLC and RLIN, when "XY" (with no space in between) is entered in a title-keyword search, the databases will treat it as one keyword. On the other hand, when "X Y" (with a space in between) is entered, the databases will treat it as two separate keywords, regardless of their sequence or adjacency. Therefore, all the records with the configuration of figure 2a–f, instantiated by figure 1a–f, will be retrieved. That is why the precision rate is low for both OCLC and RLIN when Pinyin keywords are entered in a non-aggregated fashion. These databases retrieve all the records whose titles in part match each of the four cases in figure 3. Obviously, only those titles that in part match figure 3a have the potential of being a targeted record whereas all those matching cases figure 3b–d will be untargeted records.

In PKUL, in contrast, when both "XY" and "X Y," with or without a space between the two syllables, are inputted in a title-keyword search, the database will impose two conditions simultaneously. That is, both sequence and adjacency have to be fulfilled if a particular record is retrieved. In other words, the PKUL database will only retrieve those records with the configuration of figure 3a while filtering out figure 3b–d. That is why its precision percentage is very high compared with OCLC and RLIN.

Configuration	Sequence	Adjacency
a. XY	Yes	Yes
b. X...Y	Yes	No
c. YX	No	Yes
d. Y...X	No	No

Figure 3. Generalized Configuration by Sequence and Adjacency

a. *fu shui dao lun*	(W-X-Y-Z)
b. *fushui daolun*	(WX-YZ)
c. *fushuidaolun*	(WXYZ)
d. *fu shuidao lun*	(W-XY-Z)
e. *fushuidao lun*	(WXY-Z)
f. fu shuidaolun	(W-XYZ)

Figure 4. Six Ways of Entering the Title in Pinyin

It is worth pointing out that, although a quantitative method was used for this study, the difference found between PKUL on the one hand and OCLC and RLIN on the other is qualitative rather than quantitative. That is, there is absolutely no difference in retrieving results with PKUL no matter whether the two syllables of the keywords are aggregated or not, but this is not the case with either OCLC or RLIN. The same results are expected if more keywords are tested. Readers are referred to Huang and Huang and Haynes for discussions of other title-keyword searches.[13] Suffice it here to say that similar results are observed across those tests.

Known-Title Searches

With the findings discussed in the previous section, the author extended the study to known-title searches in PKUL. The question was: Do the advantages of PKUL observed in title-keyword searches apply to known-title searches? To answer this question, the author searched a number of known titles in the PKUL database. The title in figure 1c, *Fushui daolun* (赋税导论) "An Introduction to Taxation," is an example of those searched. This title consists of two compound words "WX" and "YZ" represented by four syllables. The author entered it in the six ways using Pinyin searches, as in figure 4.

The parentheses contain symbolic configurations with hyphens representing spaces. Of these ways of entering the title, figure 4a divides the title into four individual syllables; that is, it is in non-aggregated Pinyin. On the other hand, figure 4b is in aggregated Pinyin because it divides the title into two words. In figure 4c–f, however, the four syllables of the title are aggregated in obviously wrong ways. What is interesting is the fact that all six different ways of segmenting the title, four of them being obviously wrong, led to the same result: the same title was retrieved in all six searches. That is, the PKUL database treats six-way segmented titles as the same thing because they all satisfy the conditions of sequence and adjacency, ignoring the spaces. Moreover, a known-title search can also be accomplished by inputting just the initial letters of the syllables in the title, instead of spelling out the whole title. For example, the author entered "fsdl," with or without spaces between these letters, and retrieved the title in figure 1c without bringing out any other untargeted titles. This initials-only Pinyin option makes searches much more convenient to end users. It also helps reduce errors caused by misspelling search terms, which are not uncommon even among native speakers.[14]

For comparison, the author conducted similar searches in OCLC selecting separately Title and Title Phrase search modes. All six-way segmented titles in figure 4a–f were inputted in both modes, but only figure 4a successfully retrieved the targeted title. In the Title mode, figure 4a in fact retrieved three records including *Fu shui dao lun* (赋税导论) "An Introduction to Taxation." The other two are two different editions of a book on rivers, flood dams, and reservoirs in China, titled *Lu shui ke tan: fu lu* (潞水客谈：附录), which shares two homophonous syllables, *shui* and *fu*, with the targeted title. Clicking the link More Details revealed that the Other Titles field displays the remaining two homophonous syllables *dao* and *lun*. In the Title Phrase mode, which is available only in the Advanced Search, figure 4a, and only figure 4a, retrieved the targeted title *Fu shui dao lun* without bringing out with it any other untargeted records.

Then, the author used both Title and Title Phrase modes to make similar searches in RLIN. The results were similar to those from OCLC. After figure 4a was entered in the Title mode, the targeted record was retrieved with thirty-five other untargeted ones. In the Title Phrase mode, figure 4a retrieved the targeted title without bringing out other records. All the other searches with figure 4b–f were unsuccessful in both modes. It is worth mentioning that the reason that figure 4b, in which four syllables are aggregated into two words, failed to retrieve the targeted record is probably due to the absence of aggregator character, which joins syllables into words, in this record.

Conclusion

The findings of this study suggest that PKUL has three important advantages over OCLC and RLIN in its way of handling romanized Chinese records. First, it simultaneously imposes the conditions of sequence and adjacency in Pinyin retrieval and, in so doing, greatly raises the precision rate without negatively affecting recall. For the term "XY" in title-keyword searches, as the symbolic configuration in figure 3 indicates, PKUL only retrieves records that in part match figure 3a while filtering out all those that in part match figure 3b–d. Only titles in part matching figure 3a have the potential of being the targeted records, whereas all those in part matching figure 3b–d are doomed to be untargeted records. This merit of PKUL also extends to its known-title searches. Secondly, PKUL is not sensitive to the distinction between aggregated and non-aggregated Pinyin in retrieving. In title-keyword searches, for a keyword "XY," whether the patron enters "XY" or "X Y," the results remain the same. In known-title searches, whether the patron separates all syllables of the title or puts them into a continuous string, and no matter how the patron segments the title, the results will remain unaffected. In other words, different ways of segmentation of the title will not alter the results of retrieval; only those titles that have the exact syllables in the exact order will be retrieved. Finally, PKUL can also process initials-only Pinyin search terms. Instead of spelling out the whole search term, patrons can input only the initial letters of its syllables to retrieve the targeted item. This option can not only save inputting time, but also avoid errors caused by misspelling. All these merits make PKUL more user-friendly than OCLC and RLIN. They are worth adopting to improve the handling of Chinese language records in the North American library community.

In order to learn more about the PKUL database, the author visited PKUL in summer 2002. It was found that PKUL adopted the SIRSI Unicorn system in 1999.[15] In trying to adapt this system into the Chinese environment, PKUL takes the non-segmented string-match approach to retrieval in Pinyin. Admittedly, the PKUL way of retrieving Chinese-language records in the Pinyin mode is not perfect. It cannot avoid untargeted records that satisfy both conditions of sequence and adjacency, as instantiated by figure 1b and c and symbolically represented by figure 2b and c. Nevertheless, it does filter out all those irrelevant records, as instantiated by figure 1d–f and represented by figure 2d–f, and, for that matter, maintains a comparatively high precision rate. According to Nie and Shen, PKUL plans to introduce tonal distinction and lexical segmentation in its second phase of adapting the SIRSI Unicorn system to the Chinese environment.[16] This planned move certainly deserves our further attention.

With the capacity of OPACs to show vernacular scripts and that of computers to read them, the current monosyllabic, non-aggregated Pinyin records in the databases of OCLC and RLIN, now equipped with their character counterparts, no longer pose a serious problem, as they used to, of low readability caused by homophony. From the perspective of cataloging, keeping the current monosyllabic, non-aggregated Pinyin format means the avoidance of launching a very costly task of converting all the Chinese language records into the polysyllabic, aggregated Pinyin format. Furthermore, this also means avoidance of potential inconsistencies in word division in converting the existing records and in creating future records. From the perspective of retrieving, the inconsistencies in word division in cataloging will only be compounded by much more such inconsistencies in inputting search queries. End users will benefit tremendously if OCLC and RLIN adopt PKUL's non-segmented string-match approach to Pinyin retrieval, because this approach is not at all sensitive to inconsistencies or errors in word division and greatly improves precision at no cost of recall. It is user-friendly.

References and Notes

1. Clément Arsenault, "Word Division in the Transcription of Chinese Script in the Title Fields of Bibliographic Records," *Cataloging and Classification Quarterly* 32 (2001): 109–37.
2. Ibid.
3. Clément Arsenault, "Word Division in the Transcription of Chinese Script in the Title Fields of Bibliographic Records" (Ph.D. diss., University of Toronto, 2000), 38.
4. See the discussions in Arsenault, "Conversion of Wade-Giles to Pinyin: An Estimation of Efficiency Improvement in Retrieval for Item-Specific OPAC Searches," *Canadian Journal of Information and Library Science* 23 (1998): 1–28; Arsenault, "Word Division in the Transcription of Chinese Script in the Title Fields of Bibliographic Records"; Arsenault, "Testing the Impact of Syllable Aggregation in Romanized Fields of Chinese Language Bibliographic Records." In *Dynamism and Stability in Knowledge Organization*, eds. Clare Beghtol, Lynne C. Howarth, and Nancy J. Williamson, (Wèurzburg, Germany: Bergon Verlag, 2000), 143–49; Arsenault, "Word Division in the Transcription of Chinese Script in the Title Fields of Bibliographic Records"; Arsenault, "Pinyin Romanization for OPAC Retrieval: Is Everyone Being Served?" *Information Technology and Libraries* 21 (2002): 45–50; Linda Groom, "Converting Wade-Giles Cataloging to Pinyin: The Development and Implementation of a Conversion Program for the Australian National CJK Service," *Library Resources and Technical Services* 41 (1997): 254–63; Jie Huang, "The Issue of Word Division in Cataloging Chinese Language Materials" (master's thesis, University of Oklahoma, 2002); Jie Huang and Kathleen J. M. Haynes, "The Issue of Word Division in Cataloging Chinese Language Titles," *Cataloging and Classification Quarterly*, forthcoming; Victor H. Mair, "Pinyin Orthographical Rules for Libraries," *Chinese Librarianship: An International Electronic Journal* 10 (2000): 1–3. Accessed Apr. 14, 2002, www.

whiteclouds.com/iclc/cliej/cl10mair.htm; Mair, "Pinyin Orthographical Rules for Libraries: A Follow-up," *Chinese Librarianship: An International Electronic Journal* 11 (2001): 1–7. Accessed Apr. 14, 2002, www.whiteclouds.com/iclc/cliej/cl11mair.htm; Mair, "Pinyin Orthographical Rules for Libraries: A Recent Literature Review," *Chinese Librarianship: An International Electronic Journal* 11 (2001): 1–3. Accessed Apr. 14, 2002 , www.whiteclouds.com/iclc/cliej/cl11mair2.htm; Philip A. Melzer, "Pinyin Romanization: New Developments and Possibilities," *Journal of East Asian Libraries* 109 (1996): 91–92; Meltzer, "Pinyin Romanization: Word Division Recommendation," *Chinese Librarianship: An International Electronic Journal* 2 (1996): 1–3. Accessed Apr. 14, 2002, www.whiteclouds.com/iclc/cliej/cl2phil.htm; Meltzer, "Library of Congress Converting to Pinyin for Chinese Romanization," *Chinese Librarianship: An International Electronic Journal* 4 (1997): 1–2. Accessed Apr. 14, 2002, www.whiteclouds.com/iclc/cliej/ cl4phil2.htm.

5. Meltzer, "Pinyin Romanization: Word Division Recommendation," *Chinese Librarianship: An International Electronic Journal* 2 (1996): 1–3. Accessed Apr. 14, 2002, www.whiteclouds.com/iclc/cliej/cl2phil.htm.

6. See Arsenault, "Word Division in the Transcription of Chinese Script in the Title Fields of Bibliographic Records"; Huang, "The Issue of Word Division in Cataloging Chinese Language Materials"; Mair, "Pinyin Orthographical Rules for Libraries"; and Mair, "Pinyin Orthographical Rules for Libraries: A Recent Literature Review."

7. Arsenault, "Word Division in the Transcription of Chinese Script in the Title Fields of Bibliographic Records"; Arsenault, "Testing the Impact of Syllable Aggregation in Romanized Fields of Chinese Language Bibliographic Records"; Arsenault, "Word Division in the Transcription of Chinese Script in the Title Fields of Bibliographic Records"; and Arsenault, "Pinyin Romanization for OPAC Retrieval: Is Everyone Being Served?"

8. Arsenault, "Word Division in the Transcription of Chinese Script in the Title Fields of Bibliographic Records."

9. Huang, "The Issue of Word Division in Cataloging Chinese Language Materials."

10. See also Huang and Haynes, "The Issue of Word Division in Cataloging Chinese Language Titles."

11. Arsenault, "Word Division in the Transcription of Chinese Script in the Title Fields of Bibliographic Records," 116.

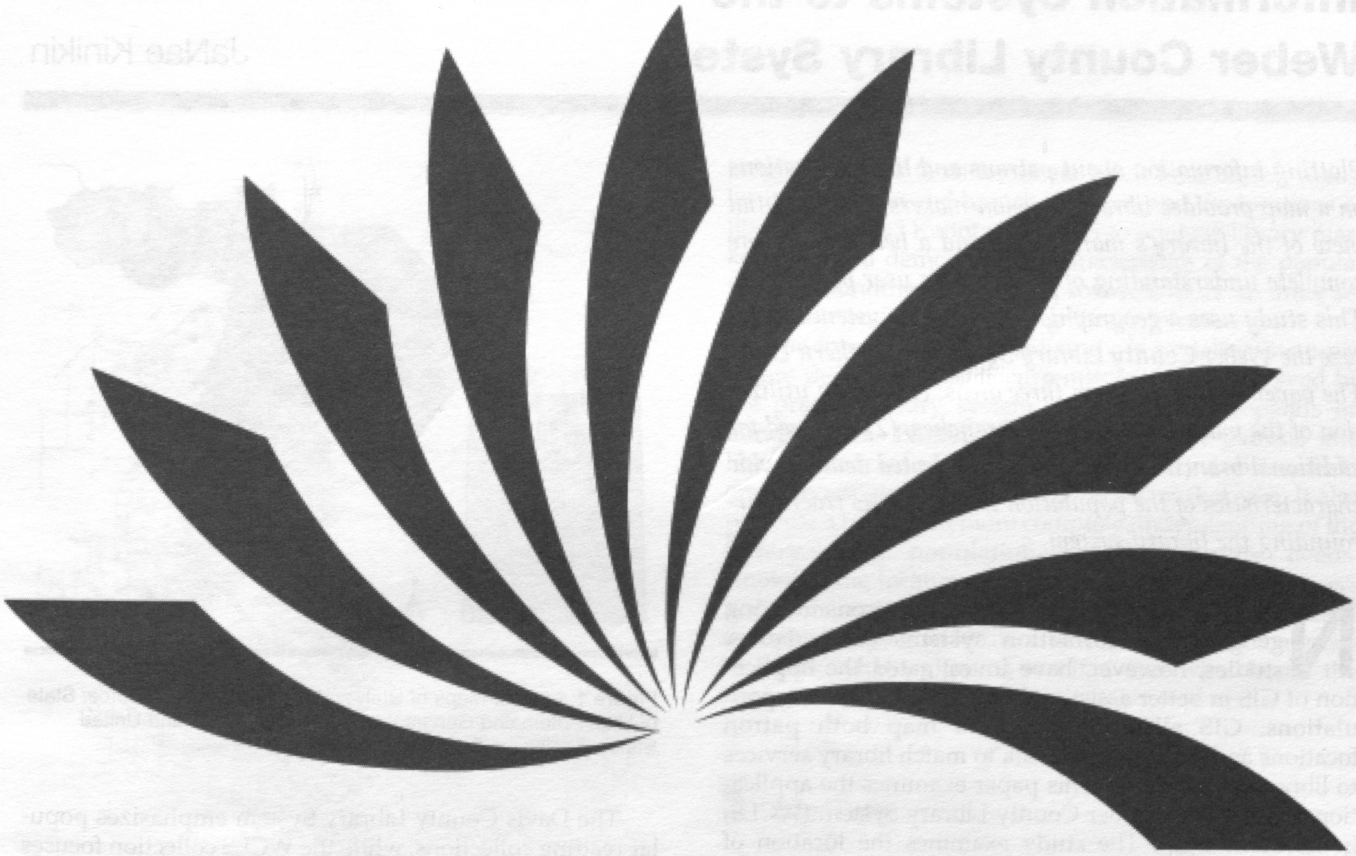
12. See, for example, William E. Studwell, Rui Wang, and Hong Wu, "A Tale of Two Decades: The Controversy over the Choice of a Chinese Language Romanization System in American Cataloging Practice," *Cataloging and Classification Quarterly* 18 (1993): 117–24; Karl K. Lo and R. Bruce Miller, "Computers and Romanization of Chinese Bibliographic Records," *Information Technology and Libraries* 10 (1991): 221–33.

13. Huang, "The Issue of Word Division in Cataloging Chinese Language Materials"; Huang and Haynes, "The Issue of Word Division in Cataloging Chinese Language Titles."

14. Arsenault, "Word Division in the Transcription of Chinese Script in the Title Fields of Bibliographic Records"; Arsenault, "Pinyin Romanization for OPAC Retrieval: Is Everyone Being Served?"

15. Nie Hua and Shen Zhenghua 聂华、沈正华, Unicorn Xitong Zhongwen Jiansuo Jizhi Shixing Fang'an Jieshao, Unicorn 系统中文检索机制试行方案介绍 [A Report on the Try-out Plan for the Unicorn Chinese Retrieval System] (Beijing: Peking University, 2002).

16. Ibid.



Content Transformation Solutions

- **SGML, HTML, XML**
- **Imaging**
- **eBook & PDF**
- **DTD Development**
- **Web-site Development**
- ***and so much more...***



Information Publishing Group, 1930 SouthWest 48th Lane, Cape Coral, FL 33914
239.549.2384. • iris.hanney@techbooks.com • www.techbooks.com

Applying Geographic Information Systems to the Weber County Library System

JaNae Kinikin

Plotting information about patrons and library locations on a map provides library decision-makers with a spatial view of the library's market area and a better and more complete understanding of the library's user population. This study uses a geographic information system to analyze the Weber County Library System in northern Utah. The paper concentrates on three areas: (1) patron utilization of the main library and its branches, (2) the need for additional branch libraries, and (3) selected demographic characteristics of the population in the census tracts surrounding the library system.

Numerous studies focus on library patrons utilizing geographic information systems (GIS). Fewer studies, however, have investigated the application of GIS in better assisting libraries to serve their populations. GIS allows libraries to map both patron locations and demographic data to match library services to library users' needs. This paper examines the application of GIS to the Weber County Library System (WCLS) in northern Utah. The study examines the location of patrons in relation to the library system and determines whether or not there is a need for additional library branches or services.

Background

The focus of this study is on the combined area of Weber and Davis counties located in northern Utah. Although WCLS primarily serves Weber-county residents, a reciprocal agreement between the library systems in the two counties allows residents from both counties to use either system. Approximately 196,000 residents live in Weber County, 40 percent of whom live in Ogden, the industrial hub of the county and location of the main WCLS library. The ethnic make-up of the county is predominately white (87.7 percent), although there is a significant Hispanic population (12.6 percent). Davis County has a population of approximately 238,994 residents, is predominately white (94 percent), and has fewer Hispanics (5.4 percent) than Weber County.[1] The majority of the population in both counties resides in a central corridor between Interstate 15 and Highway 89 (figure 1).

JaNae Kinikin (jkinikin@weber.edu) is Science Librarian at Stewart Library, Weber State University, Ogden, Utah.

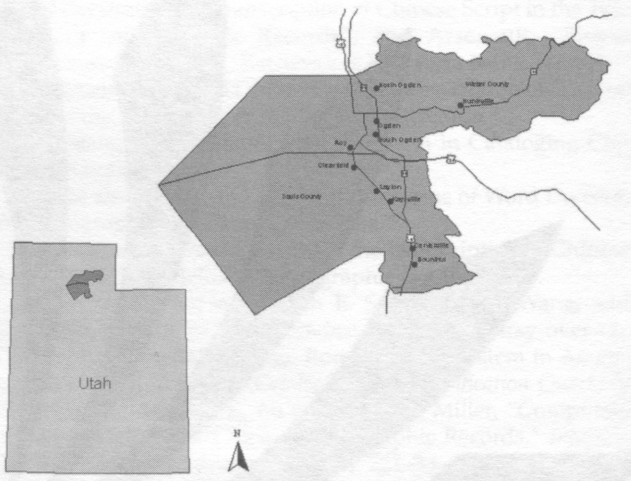


Figure 1. Locator Maps of Utah and the Study Area (Source: State of Utah Automated Geographic Reference Center and United States Geological Survey)

The Davis County Library System emphasizes popular reading collections, while the WCLS collection focuses on lifelong learning. Additionally, WCLS has a much larger collection, containing more research and reference materials. As a result, the collections serve different purposes and may attract different patron populations.

The main WCLS library is located in downtown Ogden, the north branch in the city of North Ogden, the Ogden Valley branch in Huntsville, and the southwest branch in Roy (figure 2). Two new library branches have been proposed. One will be located in the southern section of the county, and its location has already been determined. The other proposed branch will be located in the northwest, although a site has not yet been selected (figure 2).

Literature Review

A review of the literature revealed a large amount of information on library-location theory and public library user characteristics. A lesser amount of information was found regarding the application of GIS in public library planning.

The most important factor found regarding library location was accessibility, including patron distance from the library, and the selection of optimal locations when constructing new library buildings.[2] This study focuses on the location of current patrons of the library system, as well as analyzing areas for the construction of additional branch libraries. Hawkins, Morris, and Sumsion

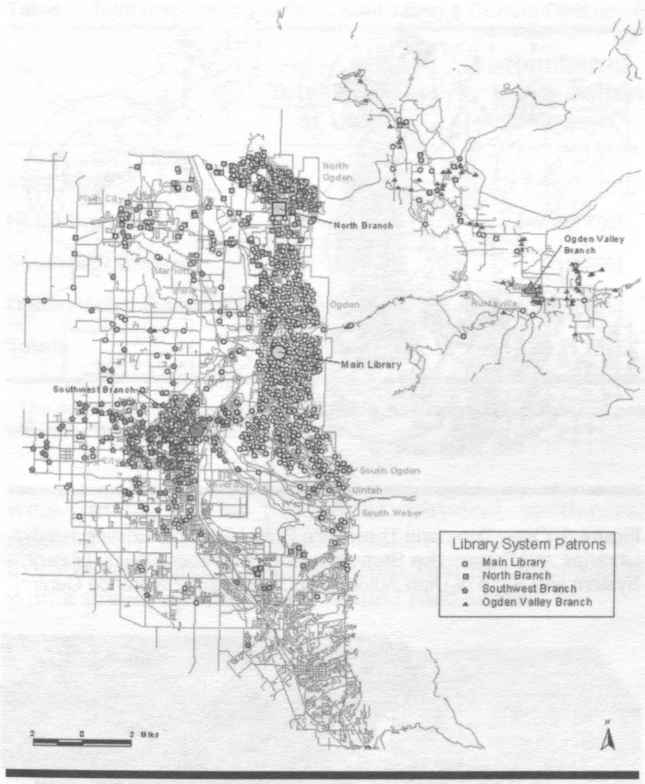


Figure 2. Main and Branch Libraries and Patron Locations (Source: Weber County Library System and State of Utah Automated Geographic Reference Center)

examined reasons why people use both main and branch libraries and, again, accessibility was a major issue. In their study, patrons' responses included "easy to get to" and "nearest my home."[3]

Additionally, Koontz found the demographic factors of income and ethnicity to be important determinants of library usage.[4] O'Connor and Fortenbaugh studied socioeconomic indicators and library use in the state of New Jersey and found that "[a]s a community's median family income increases, library funding increases."[5] The present study mapped existing libraries and possible sites for new branches. It also examined median household income and Hispanic ethnicity in the surrounding census tracts.

Studies have also analyzed the application of GIS to assist libraries in better serving their patrons. Ottensmann conducted a study of public library branches for the Indianapolis-Marion County Public Library System and presented a number of examples of GIS applications in a library setting, including mapping where library patrons live and their travel distance to libraries.[6] Jones used GIS to map patron address information to provide a visual display of library patrons.[7] Clark described the use of mapping techniques to differentiate block boundaries, roads, location of main and branch libraries, and the locations of library patrons.[8] Koontz used GIS to analyze library market areas and demographic characteristics of the population surrounding libraries to provide a more accurate set of parameters for determining library location.[9]

The intent of the present study is to determine areas where the population is currently being underserved by the present library system, as library use depends on library users. Plotting information about patrons and library locations on a map provides library decision-makers with a spatial view of the library's market area. It also presents a better and more complete understanding of the library's user population and the patrons' needs. Knowing the location and characteristics of library users and their reasons for visiting the library helps planners determine where to locate a new library, what materials to select, what programs to offer, and what hours the library should be open.

Method

This project followed the methods Ottensmann used in his study of the Indianapolis-Marion County Public Library.[10] Data for the project were collected during the week of February 16–22, 2003. Consultation with WCLS's director and technology director determined that this week would provide typical library patron circulation patterns. The automation librarian at WCLS extracted from the library system's patron database the addresses of patrons who had checked out at least one item from the main or a branch library during the selected week in February.[11] Because address locations for infrequent library users and those who only use library resources within the library were not available, the study excluded these patrons. Library users with only post-office boxes were also excluded because their addresses could not be plotted.

ArcView 3.3 GIS software distributed by the Environmental Systems Research Institute was used for the project. The mapping process was begun by obtaining the Census Tiger File 2000 Street Network from the Automated Geographic Data Center.[12] A layer was then added that included the Weber County main library and its branches to the project. After that, the addresses extracted from the patron database for each library were matched to the street network of Weber and Davis Counties using ArcView's geocoding feature. This process created four layers, one for each of the three library branches plus the main library. The four layers contained library patron residential locations for each library designated by points. Addresses that did not

match the street network were hand-plotted. To differentiate the layers, patrons borrowing materials from a particular library and the library itself were assigned corresponding symbols (figure 2).

Buffers—concentric circles drawn at selected intervals from each library location—were added to the map to determine the number of patrons living within a certain distance of each library. In a literature survey on distance and public library use, Palmer found that most library users travel two miles or less when visiting a library, so buffers of one, two, and three miles were drawn.[13] Locations for the projected southeast branch and a possible location for a northwest branch were also plotted. Circles were drawn around these two points to examine overlap with the existing library branches, as well as to determine areas from which additional library users might originate (figure 3).

To determine population demographics of the existing system and the prospective branch libraries, data from the 2000 Census was added to the map. Two variables—median household income and Hispanic population—were examined (figures 4 and 5). This information was included to help analyze whether or not the Hispanic population is concentrated in certain locations, and to ascertain this population's proximity to library services. Income was analyzed since it is an important determinant of library usage and support.[14]

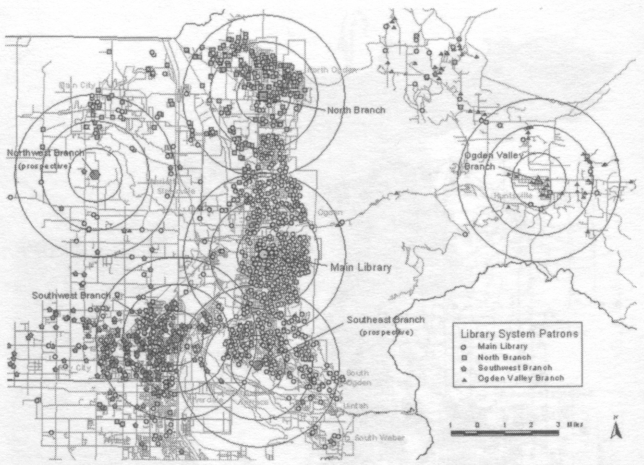


Figure 3. One-, Two-, and Three-Mile Buffers around Existing Libraries and Prospective Branches (Source: Weber County Library System and State of Utah Automated Geographic Reference Center)

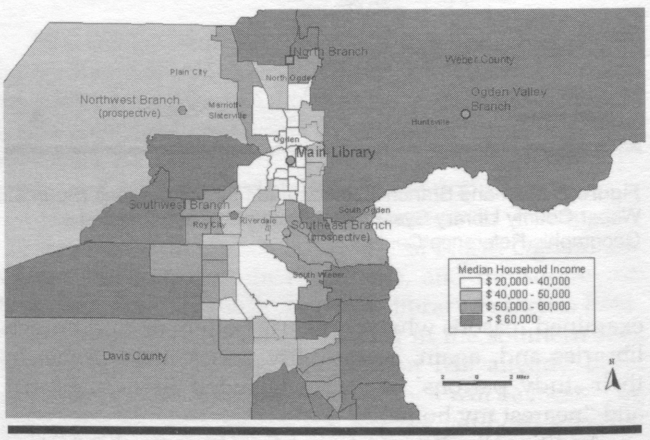


Figure 4. Median Household Income for the Weber County Library System (Source: Weber County Library System and Geolytics Census Data)

Analysis and Discussion

Straight-line Distance Traveled

Figure 2 shows that the majority of patrons use the library closest to their residence; however, it also illustrates that some library users travel longer distances to visit other libraries. A small percentage of main (2 percent) and southwest (2 percent) patrons traveled ten or more miles to visit their respective branches, while a larger percentage of north (13 percent) and Ogden Valley (16 percent) branch patrons did so. Additionally, in this study 3 percent of north-branch patrons traveled twenty or more miles to visit that branch. Several factors might account for patrons traveling longer distances to visit a library far from their primary residence. A specific library may be closer to their place of employment and more easily accessible. One library's collection may be more appealing to a patron, with more books or other resources about a subject of interest to that individual. One library may offer a service or event that appeals to an individual patron, such as a book talk on a certain topic, better Internet access, or a special exhibit.

Patron Distance

The main library serves almost twice the number of patrons as the north and south branches. Each of these branches serves a similar number of patrons. The majority of patrons for the main, north, and southwest branches live within a three-mile radius. The majority of patrons for the Ogden Valley branch live outside a three-mile radius (table 1).

The Ogden Valley branch only displays eighty-six patrons. This low number may be due to the fact that many patrons who frequent this branch live within the city limits of Huntsville and therefore have post-office boxes that

Table 1. Total Number of Patrons Living within a Certain Distance of Each Branch

	Total Number of Users	Number of Users within One Mile*	Number of Users within Two Miles*	Number of Users within Three Miles*	Number of Users outside Three-Mile Radius
Main Library	904	146 (16%)	340 (38%)	490 (54%)	414 (46%)
North Branch	420	72 (17%)	219 (52%)	276 (66%)	144 (34%)
Southwest Branch	423	77 (18%)	237(56%)	317 (75%)	106 (25%)
Ogden Valley Branch (OK)	86	11 (13%)	25 (29%)	36 (42%)	50 (58%)
Totals	**1833**	**306 (17%)**	**821 (45%)**	**1119 (61%)**	**714 (39%)**

* The number of users within each buffer is cumulative (the number of patrons within the one-mile buffer is included in the count of the number of patrons within the three-mile buffer.)

were unable to be plotted as physical residences. Additionally, the Ogden Valley branch is located in a relatively isolated area—a valley surrounded by mountains—with a small population and limited road access.

Need for New Branches

An examination of table 2 reveals that 18 percent of the library system's patrons in this study reside within a three-mile radius of the prospective southeast branch. Although there would be some overlap between the southwest branch and a new southeast branch (demonstrated by figure 3), the southeast branch would attract new patrons who live beyond the three-mile radius of the current branch. Table 2 shows the number of patrons who live within one to three miles of the designated location for the new southeast branch and are currently using other branches.

Table 3 shows the number of patrons currently using other branches who live one to three miles from a prospective location for a new northwest branch. Very few patrons in the northwest area of the county currently use the existing library system (figure 3), which may be due to the distance required to travel to a library branch or to the limited population in that area. Before pursuing a new branch in this area, more research (including surveys and population projections) should be done to determine whether or not there is a genuine need for library services in this area.

Demographic Characteristics of Patrons

Figure 4 portrays the median household income by census tract. This map shows that the population with the lowest median income surrounds the main library. The north and Ogden Valley branches serve a population with a higher median income. Traditionally, individuals with higher incomes may be more willing to support the construction of new library branches.[16] The prospective locations for both the northwest and southeast branches are in areas with a median household income of at least $33,000. The median household income around the main library is between $20,000 and $33,000.

Figure 5 shows the location of the Hispanic population in relation to WCLS. The majority of the Hispanic population resides in downtown Ogden near the main library. This high percentage indicates a need for the purchase of a greater number of Spanish-language books and Hispanic programming at this library. Moving away from downtown Ogden, the Hispanic population declines significantly, with the smallest percentage surrounding the Ogden Valley branch in Huntsville. The north and southwest branches have a surrounding Hispanic population between 5.2 percent and 9.8 percent. Although not a high percentage, in some cases minorities use the library more heavily.[15] A survey conducted at these two branches will help determine if there is a need for additional Spanish-language services. The prospective southeast branch will also have to consider the Hispanic population.

Conclusions

The current study examined the application of GIS to WCLS and was limited to those patrons who had borrowed items during a selected week in February 2003. GIS was used to plot the locations of patrons and the library system to determine which patrons utilize which library branch. Next, the need for additional library branches was examined by analyzing the locations of the patrons using the current library system. Finally, two demographic characteristics, income and ethnicity, were added and analyzed.

This study found that an additional branch in the southeast area of the county would likely be supported.

Further inquiry should be made at the north and southwest branches to see if there is a need for Spanish services. At the current time, the population levels in the northwest do not indicate a need for a branch in that area. As the population of the area continues to grow, a new branch may be required.

Limitations

The data for the study were collected during the week of February 16–22, 2003. The study focused on patrons who checked out items during that week and excluded patrons who use the library infrequently or who only use library resources within the library. Additionally, the study excluded library users with only post-office box addresses, which cannot be plotted on a street network.

Future Research

This study was limited to those patrons borrowing books from the main library and library branches; additional data could be collected on library users who only use materials in the library, and on potential library patrons. These data could be plotted to provide library planners with additional information on how to serve the library's patrons more effectively and to target new library users. A survey of library-card holders in areas not currently served by the present system could be conducted and the data plotted to determine locations for new branches. Driving time versus travel distance for patrons could be analyzed to determine the amount of time required for patrons to drive to their nearest branch.

Demographic data from the United States Census Bureau could be incorporated to target specific patron populations. Library-card holders with young children might be invited for a summer reading program. Teenagers could be asked to participate in an after-school literacy program. Additional information could be collected on patrons through surveys. Those with particular interests could be identified for a book discussion group or notified when books related to their interest(s) became

Table 2. Patrons Utilizing Current Library System Residing within One-, Two-, and Three-Mile Radius of the Prospective Southeast Branch Library

Branch Patrons Are Currently Utilizing	One Mile	Two Miles	Three Miles
Main Library	51	166	278
North Branch	2	4	6
Southwest Branch	4	18	43
Ogden Valley Branch	0	2	3
Totals	**57 (3%)**	**190 (10%)**	**330 (18%)**

Table 3. Patrons Utilizing Current Library System Residing within One-, Two-, and Three-Mile Radius of the Prospective Northwest Branch Library

Branch Patrons Are Currently Utilizing	One Mile	Two Miles	Three Miles
Main Library	0	8	22
North Branch	0	7	23
Southwest Branch	1	4	11
Ogden Valley Branch	0	0	1
Totals	**1 (0%)**	**19 (1%)**	**57 (3%)**

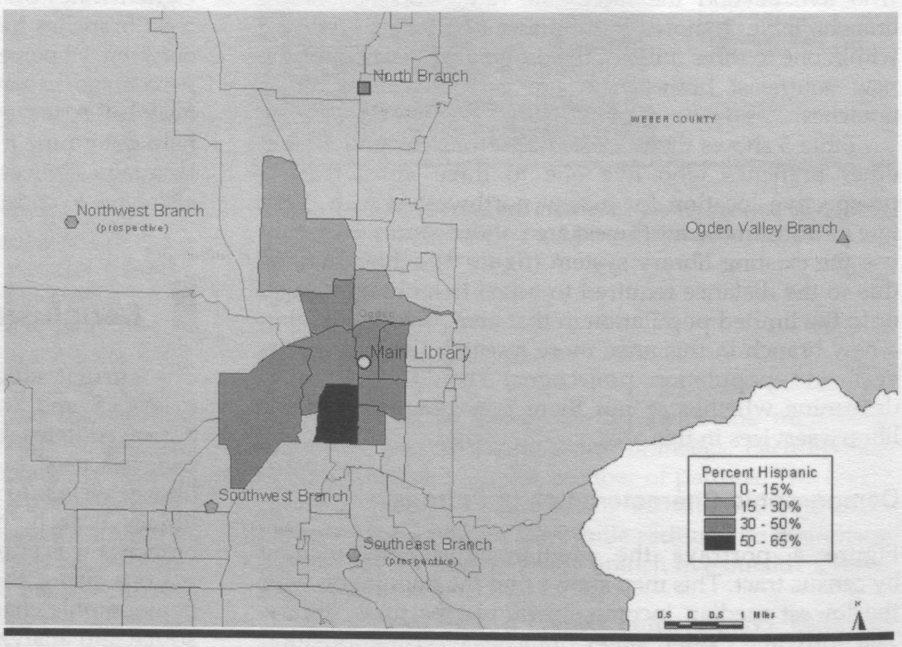


Figure 5. Hispanic Population for the Weber County Library System (Source: Weber County Library System and Geolytics Census Data)

available in the library. The library might offer courses or workshops designed for people speaking English as a second language. A multitude of possibilities exist in applying GIS in the public library to provide better services to library patrons.

Acknowledgments

The author would like to thank the following people from WCLS for their assistance in preparing this paper: Linda Wangsgard, director; Theron Stock, automation librarian; and Scott Jones, technology director. A special thanks also goes to Cameron Lindsley for his help with ArcView, and Ingrid Weinbauer for her helpful comments and suggestions on the paper and the maps. Finally, the author would like to thank Kathy Payne, Fran Zedney, and Keith Hench for their editing help.

References

1. Bureau of the Census, American Factfinder [online]. Weber County, Utah Statistics. Accessed Mar. 30, 2003, http://factfinder.census.gov/.

2. Susan E. Palmer, "The Effect of Distance on Public Library Use: A Literature Survey," *Library Research* 3 (1981): 315–54; Christine M. Koontz, "Public Library Site Evaluation and Location: Past and Present Market-Based Modeling Tools for the Future," *Library and Information Science Research* 14 (1992): 379–409.

3. Margaret Hawkins, Anne Morris, and John Sumsion, "Socioeconomic Features of UK Public Library Users," *Library Management* 22 (2001): 258–65.

4. Christine M. Koontz, *Library Facility Siting and Location Handbook* (Westport, Conn.: Greenwood, 1997).

5. Daniel O. O'Connor and Robert Fortenbaugh, "Socioeconomic Indicators and Library Use," *Public Libraries* (May/June 1999): 156–64.

6. John R. Ottensmann, "Using Geographic Information Systems to Analyze Library Utilization," *Library Quarterly* 67 (1997): 24–49.

7. Alan D. Jones, "Where do all the Good Books Go? Geographic Information Systems and the Local Library," *Australian Library Journal* (Nov. 1993): 241–49.

8. Philip M. Clark, "Thematic Mapping, Data Mapping, and Geocoding Techniques for Analyzing Library and Information Center Data," *Journal of Education for Library and Information Science* 36 (fall 1995): 330–41.

9. Koontz, "Public Library Site Evaluation."

10. Ottensmann, "Using Geographic Information Systems to Analyze Library Utilization."

11. Theron Stock, "Re: Question on GIS project." E-mail to JaNae Kinikin. Apr. 12, 2003.

12. State of Utah Division of Technology and Information Services. 2003. Automated Geographic Reference Center. Accessed Mar./Apr. 2003, ftp://ftp.agrc.state.ut.us/NAD83/.

13. Palmer, "The Effect of Distance on Public Library Use."

14. Koontz, *Library Facility Siting and Location Handbook;* O'Connor and Fortenbaugh, "Socioeconomic Indicators and Library Use."

15. Koontz, Library Facility Siting and Location Handbook.

16. O'Connor and Fortenbaugh, "Socioeconomic Indicators and Library Use."

USABILITY,
ACCESSIBILITY,
YOUR LIBRARY.

User Interface Consulting

OPACs
LIBRARY PORTALS
DIGITAL COLLECTIONS
METASEARCH ENGINES

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

Jump Higher: Analyzing Web-Site Rank in Google

Lisa Zhao

What makes a Web site rise to the top of a search-engine retrieval? The deluge of information and the short attention spans of users have made this question vital. The research reported in this paper aims to analyze characteristics of Web sites that may have enabled them to rise to the top of a search-engine retrieval. The author searched Google by using selected keywords—"cataloging department"—over a period of ten weeks, and then examined characteristics of the top twenty retrievals each week to determine which characteristics had the greatest influence on retrieval rank.

The number of Web sites has grown rapidly. According to the Office of Research, Online Computer Library Center (OCLC), from 1998 to 2002, the annual increase of public Web sites (sites providing free, unrestricted access—or at least a significant portion of their content—to all) was between 35 percent and 55 percent and the number grew from 1,457,000 to 3,080,000.[1] Private sites (sites whose contents are intended for a restricted audience; restriction can be explicit as in fee payment or authorization, or implicit) increased 12 percent (315,000) in 1998 and 29 percent (2,489,000) in 2002. Provisional Web sites (sites in transitory or unfinished states, offering content that is, from a general perspective, meaningless or trivial) increased over 30 percent every year, from 864,000 in 1998 to 3,143,000 in 2002.

In such a rapidly enlarging ocean of information, how to reach the top in search-engine retrievals is a serious matter for information providers. "Search engines are a gateway to information services including digital libraries and IR."[2] But Web searchers will not look at a Web site buried in a retrieval and are not patient enough to look through even a fraction of their search results. Huberman observed that "the average number of pages surfed at a site was almost three, users typically requested only one page."[3] Several analysts have estimated that "the average users' patience lasts between four to ten seconds."[4] Echoing the proliferation of online information, users' search behavior is changing to a short style. In their research, Spink, Jansen, and Bateman found that "users (in average) ask short queries and that they view two to three pages of the results."[5] An advice from the Web site of grokdotcom.com also pointed out that "your average prospect will view two to three pages before moving on, so at best, you're two clicks away from dead in the water unless you help them get where they want to go quickly."[6] When Web searchers have such short attention spans, as Bury stated, "the ranking of search results is very important, as most users won't bother to look beyond the first twenty results or so. . . if the search engine doesn't return the mostly relevant results in the first twenty, the user gives up on the search and loses faith in that search engine. It's obviously important to a Web site to be listed in the first twenty results for relevant search terms, otherwise that Web site will receive very little traffic from the search engines."[7]

Lisa Zhao (chzh@uic.edu) is Assistant Catalog Librarian in the Catalog Department at University of Illinois at Chicago.

Web Sites, Search Engines, and Their Features

Google is currently a well-known and well-used search engine. According to the Search Engine Watch, Google has been "twice-voted most outstanding search engine by Search Engine Watch readers."[8] On September 2, 2003, Google reported it indexed 3.3 billion textual documents.[9] Claiming "to serve over 70 million searches each day through its own site and its affiliates, Google can arguably be called the most important and powerful search engine of today."[10]

Each search engine keeps secret the algorithms it uses to rank Web sites. These ranking algorithms are sophisticated, may combine many factors, and are tuned in various ways. Although some literature has focused on the ranking algorithms of search engines, no one has really given an authoritative description of them. Google has made public one of its page-ranking systems—PageRank (PR). Google's PR, in brief, can be expressed as a mathematical formula (figure 1) explained by Callan.[11]

The formula in figure 1 shows that not only is the PR of linking pages important for ranking, but also how many other sites are also linked or voted for from that page. From the viewpoint of the mathematics, in the formula, $C(t1)\ldots C(tn) \downarrow \rightarrow PR(A) \uparrow$—that is, the fewer links the sites or pages that link to a site or page have, the higher PR that site or page will receive. You may wish to have other sites or pages linked to your site or page. The problem is that you cannot control how many links the other site or page can have. The formula can also be interpreted in another way: $PR(t1)\ldots PR(tn) \uparrow \rightarrow PR(A) \uparrow$—the higher the PR of the other sites or pages that link to you, the higher the PR of your site or page. The weight that PR is assigned within Google's algorithm has always been debated. Gerhart revealed, "Google does not stick with one algorithm like other search engines. Instead, Google switches between a few different algorithms from month to month."[12]

PR(A) = (1-d) + d(PR(t1)/C(t1) + ... + PR(tn)/C(tn))

PR(A): the PageRank your page will get after being linked from someone else's site (t1)

C(t1): the amount of total links that (t1) has.

d: damping factor is widely known to be 0.85.

t1: someone else's site.

PR(t1): the PageRank of the page t1 which links to you.

Figure 1. Formula for Expressing Google PR

In figure 1, the tn factor also merits discussion. This factor represents the number of sites that link to or vote for the examined site. Bradley explains, "some engines pay particular attention to the number of other pages that link to the pages that it retrieves, and they work on the assumption that if lots of people link to a page it will in some way be a 'better' page than if very few people link to it. (The flaw here, of course, is that new pages will have fewer links than older pages in many cases). Google is a good example of a search engine that utilizes this technique, and many people find that it gives particularly good results."[13]

Besides PR, Google takes many other things into account when ranking Web pages. The location of keywords in a Web site and the number of times the keywords appear in a Web page are two other factors. Sullivan comments, "one of the main rules in the ranking algorithm involves the location and frequency of keywords on a Web page. . . Pages with the search terms appearing in the HTML title tag are often assumed to be more relevant than others to the topic."[14] Regarding frequency, he has written, "frequency is the other major factor in how search engines determine relevancy. A search engine will analyze how often keywords appear in relation to other words in a Web page. Those with a higher frequency are often deemed more relevant than other Web pages."[15]

Considering the keyword appearance, another commentator has stated, "search engines will also check to see if the search keywords appear near the top of a Web page, such as in the headline or in the first few paragraphs of text. They assume that any page relevant to the topic will mention those words right from the beginning."[16]

Contents inserted between HTML <TITLE> tags are also considered an effective factor to rankings. "Make sure your target keywords appear in the crucial locations on your Web pages. The page's HTML title tag is most important. Failure to put target keywords in the title tag is the main reason why perfectly relevant Web pages may be poorly ranked."[17]

The keyword density is another factor recognized in ranking Web sites. "Google considers keyword density in the body of a page for determining relevancy too, so make sure your keywords and phrases appear a couple of times throughout the whole page. Don't go overboard though, a density of 6–10 percent seems to work best."[18]

The purpose of the HTML meta tag is to allow the designer "to provide additional text for crawler-based search engines to index along with your body copy." While the HTML meta tag may have helped in the past, "most crawlers now ignore the tag."[19] "Search Engine Features for Webmasters" reports that Google does not consider meta description and meta keywords in ranking Web sites. This research, therefore, did not examine the meta keywords.[20]

Given the variety of factors that can affect page ranking of search engines, it is clear that "ranking is a complicated process, not helped by the fact that engines all do it differently . . . The more blunt view is that it's a total mess and is compounded by the secrecy surrounding ranking."[21]

The Test

From November 13, 2002, to January 15, 2003, the author conducted the same search on Google once a week (ten weeks in all) to determine what sites appear among the top twenty retrieved and how characteristics of these sites relate to their ranking. The author studied the following factors that presumably affect where a Web-site match appears: (1) the weight of Google PR; (2) the popularity of Web sites; (3) the number and density of keywords; (4) keyword positions in the title of the home page content, in the HTML title, in the uniform resource locator (URL); and (5) the structure of domains.

In this study, the search was conducted in the Google simple-search screen, and "cataloging department" was chosen as the search keywords. The investigator, in other research, had searched Google for cataloging-department Web sites and became intrigued with how Google ranked the sites retrieved. While "cataloging department" may seem like a simple search, other studies have found that most searches are short and simple. Nowicki has discovered almost half of her study's respondents use single-word queries to search for information.[22] Spink, Jansen, and Bateman. also had similar findings: "respondents were not proposing to use many search terms or employ complex search strategies."[23]

The author chose "cataloging department" as the search keywords in this research rather than "catalog department" because the former retrieved top-twenty matches that were all related to library cataloging departments while using "catalog department" retrieved at least three commercial catalog departments that had nothing to do with library cataloging.

One factor, click popularity—the number of times a Web site has been viewed in a fixed time span—is important to page ranking. Unfortunately, the present study could not harvest these data for each Web site tested. Some institutions do not record such data and some others do not reveal these counts.

Results and Analysis

Web Sites on the Top-Twenty List

During the ten weeks of the present study, twenty-four different Web sites made it to the top-twenty list (appendix A). Only three positions: the first, fourth, and seventh, were held by the same Web sites throughout the test. The remaining positions were held by different Web sites and changed at least once in the test period. In other words, all but three Web sites that appeared in the top-twenty list changed their position at least once during the test.

The twenty-four Web sites and their position changes in the top-twenty list during the ten weeks are shown in figures 2 and 3. The figures use the acronyms and abbreviations of the Web sites' institutions (see appendix A for the full names). The order shown in figures 2 and 3 was based on the first week's result. In the ten-week test period, the ranks of the first twelve Web sites shown in figure 2 varied less than the second twelve sites shown in figure 3. The first thirteen institutions appeared in the top-twenty list in every week of the test period. None of the top twelve changed its position more than ±3 places weekly and none appeared lower than fourteenth. No Web site in the thirteenth to twenty-fourth range (figure 3) remained at the same position throughout the whole test time. Only three Web sites in figure 3 stayed on the list in every week of the test. Some sites in figure 3 appeared in the top twenty only a few times in the ten-week period. For example, BSU appeared only three times at positions eighteen and nineteen; SUNYSB showed up three times at the twentieth position.

Overall, the fact that just twenty-four Web sites appeared among the top twenty in the present study's ten-week observation shows there is stability in the ranking.

Google's Ranking Algorithm—PR

Google's PR of the studied Web sites was measured by a function of the Google tool bar that was downloaded from the Google site. The relationship between the Google PR and the place of appearance in the top twenty is shown in figures 4 and 5. The Google PR of most sites fell between four and seven. Some Web sites in the top-twenty list shared the same Google PR value. At positions one through ten, shown in figure 4, Web sites had their PR position at either five or six while the value of the PR at positions from eleven to twenty ranged from two to seven, as shown in figure 5. The highest Google PR value in the test was seven and it appeared at positions fifteen and seventeen instead of at preceding positions.

Most of the time, the top fourteen Web sites had their PR value larger than or equal to the mean value (4.5) of the tested PR in the ten-week period. Changes in the PR value at each position of the first fourteen Web sites were no bigger than ±1. The PR of other sites after the top fourteen had less stable values from week to week.

In general, especially after position ten, once the position of appearance falls, PR begins to decline despite the finding that a few sites with lower position of appearance (i.e., a larger ranking number) had a higher PR.

Popularity of the Site

The popularity of a Web site is defined as the number of Web sites that link to a given site. Google's PR does not reflect the popularity of a Web site because it involves other components in the calculation. Thus, when many sites link to the considered site, the site may or may not receive a high PR value. A low PR does not necessarily imply low popularity of a Web site.

The popularity of a Web site is comparable with the journal impact factor for rating the print journals. "The journal impact factor is a measure of the frequency with which the 'average article' in a journal has been cited in a particular year. The impact factor will help you evaluate a journal's relative importance, especially when you compare it to others in the same field."[24] In figure 1, calculation of the Google PR, the factor tn is used to represent the popularity of the retrieved site.

The present study found the number of sites that link to the studied Web sites individually by using each site's URL as the keyword to search in Google. Since 80 percent of Web sites in the top twenty remained in every week of the test period, the popularity was tested only in the last five weeks.

The results are shown in figures 6 and 7. The popularity of the Web site at the number one position is far ahead of other sites. As stated earlier, the first position was held by only one institution in the ten-week test period. Figures 6 and 7 show that the number of Web sites that link to the examined site decreased as ranking decreased. Google apparently is considering the popularity of Web sites in its ranking system. However, as previously mentioned, the flaw in using popularity as a factor for ranking is that newer pages will have fewer links than older pages.

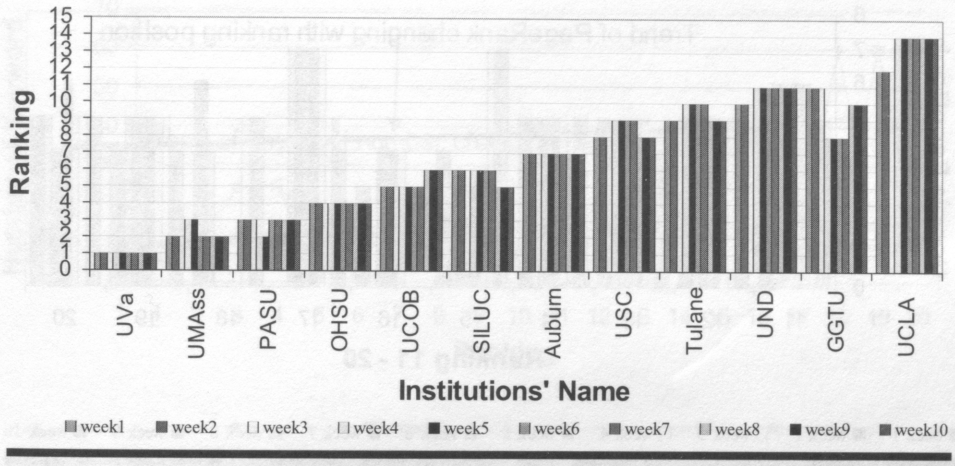


Figure 2. Top Twelve Institutions and Their Ranking in Ten Weeks

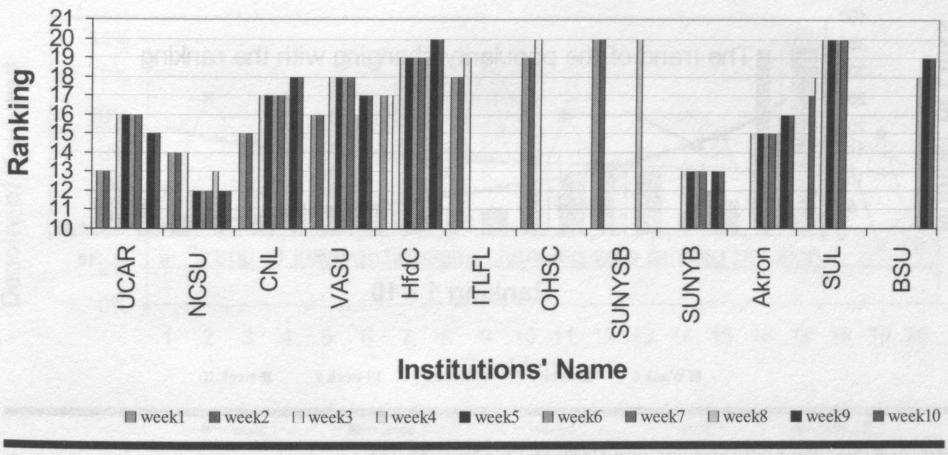


Figure 3. Top Thirteenth to Twenty-fourth Institutions and Their Ranking in Ten Weeks



Figure 4. Google PR—First Ten Positions in Ten Weeks

Number and Density of Keywords

The basic Google search was used in the present study with the keywords "cataloging department" without quotation marks. Thus, Google sought Web pages containing both words whether together or separate. The investigator counted the number of times the keywords "cataloging" and "department" appeared in the home page (that is, the first page) of each studied Web site.

In the test period, only one out of the examined twenty-four home pages changed its content once. (The search result first reached the old site and after about ten seconds it changed to the new site automatically. This study examined only the newer site.) Given this stability of content the investigator counted the number and calculated the density of keywords only once and then compared the keyword number and density with rank during the odd-numbered weeks.

The relationship between the number of search terms in each home page and ranking is shown in figure 8. For most home pages in this test, the lowest number of keywords was two and highest was fewer than thirty. There was only one special case in which the keyword number went up to sixty at position eighteen and nineteen in the first and third weeks. But this home page did not remain in those positions for very long. After appearing in the first three weeks, it disappeared from the top-twenty list as can be seen in figure 3.

Keyword density is decided not only by the number of keywords, but also by the total number of words in a home

page, calculated as (number of keywords in the home page/ number of words in the home-page content). The test recorded the number of both keywords and total words in the home page. The relation of the keyword density and the ranking position can be seen in figure 9. The result shows that the keyword density of most home pages is between 4 percent and 13 percent, which is not too far from the model of "a density of 6–10 percent seems to work best."[25]

Comparison of figures 8 and 9 shows that a high number of keywords does not always signify a high density of keywords. For example, the Web site at position eighteen in the first week had the highest number of keywords in its home page but did not have the highest keyword density. Many segments of the curves in figures 8 and 9 were overlapped because, as mentioned before, either some positions were held by the same institutions throughout the test period, or the tested values of the sites were very close. The highest-ranked Web site did not have the highest number and density of keywords. The lowest-ranked Web site did not have the lowest number and density either.

The results did not show an observable effect of the keyword number on the ranking. In the top-twenty list, as shown in figure 8, the trend of the number of keywords in the home page changed little with the ranking. As claimed by Google itself, "Google goes far beyond the number of times a term appears on a page and examines all aspects of the page's content (and the content of the

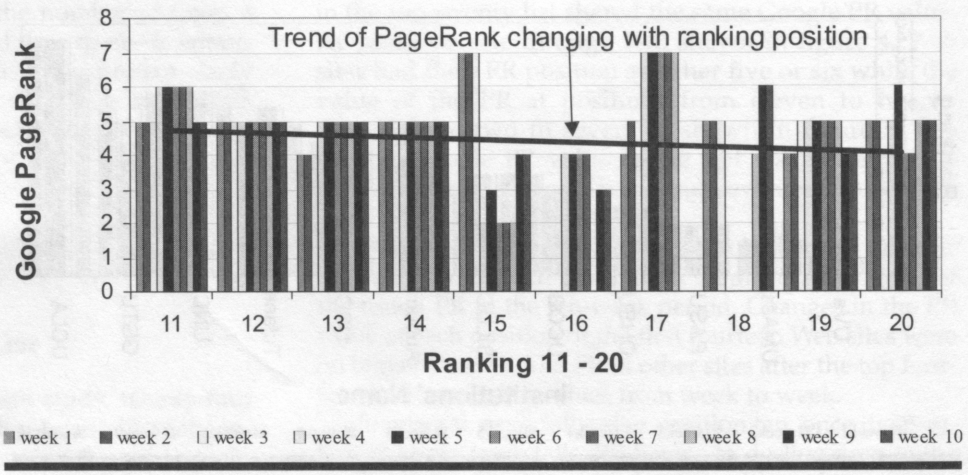


Figure 5. Google PR—Second Ten Positions in Ten Weeks

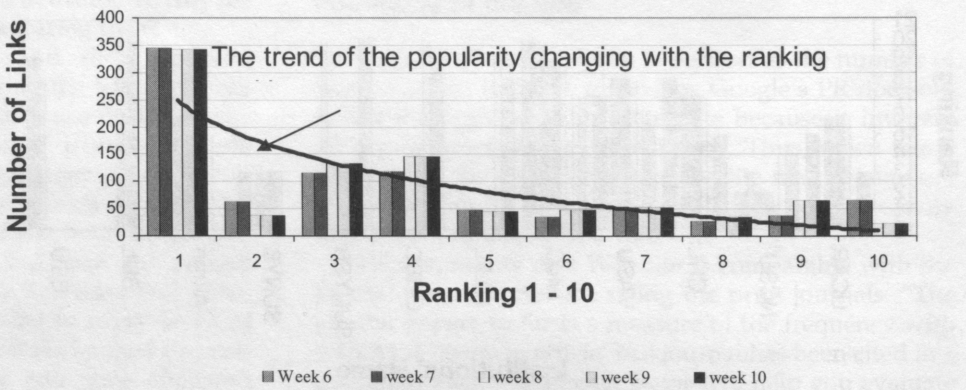


Figure 6. Popularity of Domains and Ranking One through Ten

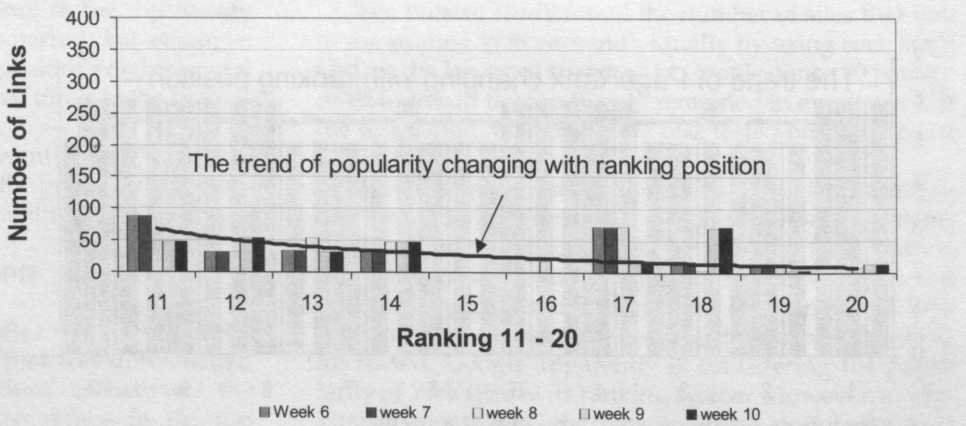


Figure 7. Popularity of Domains and Ranking Eleven through Twenty

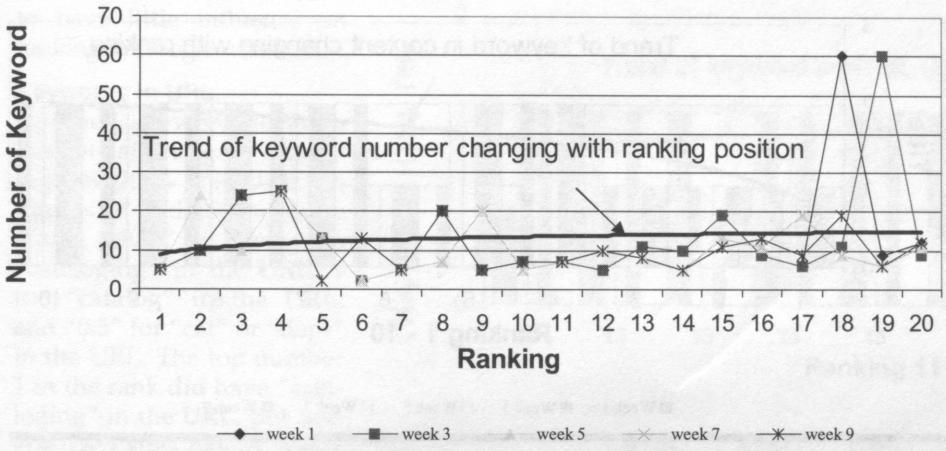


Figure 8. Keyword Number and Ranking Positions in Five Weeks

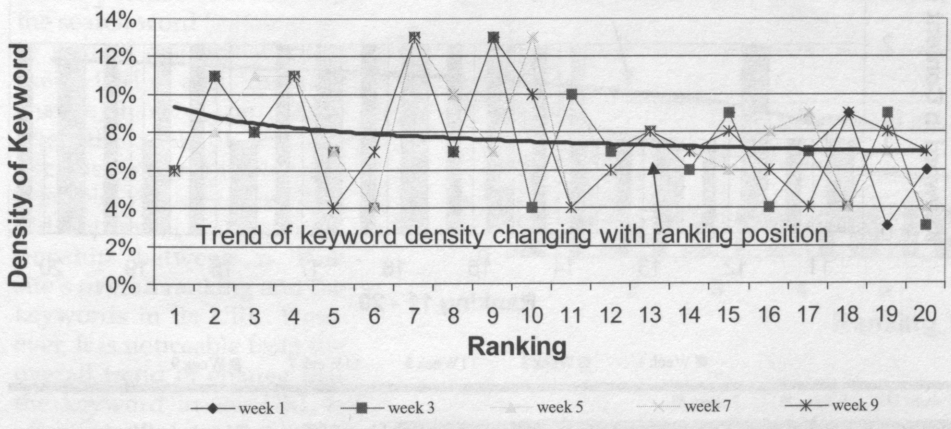


Figure 9. Keyword Density and Ranking Positions in Five Weeks

pages linking to it) to determine if it's a good match for your query."[26] In contrast, figure 9 shows that the ranking of a Web page declines as the keyword density also declines.

Keyword Positions

The investigator tested whether the keywords "cataloging department" appeared in three places: the content title of the home page, the HTML title, and the URL, and assessed their effects on ranking. Since most home pages did not change their content in the ten-week period, the investigator recorded each place and compared it with ranking positions in the odd-numbered weeks.

Keywords in the Content Title

The content title is the title that appears at the top of the home-page content. Technically, it is usually inserted between the HTML heading tags <H1> and </H1> or <H2> and </H2> at the beginning of the page. H_n of the HTML tags specifies headings in a document.

All twenty-four home pages in the test had either "cataloging department" or "cataloging" in their content title. The relation of the page ranking and the keywords found in the title of the home-page content is shown in figures 10 and 11. Those pages having the keywords "cataloging department" in their content title were scored 2, those pages that had "cataloging" were scored 1. Of the twenty-four home pages, eighteen sites (75 percent) had the keywords "cataloging department" in their content titles; six pages (25 percent) had "cataloging." Among the eighteen sites, one site had changed to only having "cataloging" in its content title in the middle of the study. It was surprising to find that the Web sites at the three highest positions did not have the keywords "cataloging department" in their content titles. More home pages at lower ranks (i.e., a higher ranking number in the figure) had both keywords in their content titles than those at higher ranks (i.e., a lower ranking number in the figure). Figures 10 and 11 depicted an inverse relationship between the presence of keywords in home-page content titles and the Web site's ranking.

Keyword in HTML Title

In Web design, one of the HTML required tags is <TITLE>. . .</TITLE>. This pair of tags contains the document title. The text inserted between these two tags does not appear within the browser window. It is usually visible in the browser's title bar and in the page source.

Web pages with search terms appearing in the HTML <TITLE> tags are considered more relevant to the topic than those Web pages that do not have any search term in the HTML <TITLE>, according to Sullivan's online article "How Search Engines Rank Web Pages."[27] The present study examined the content between <TITLE> and </TITLE> in the HTML view of each home page. The examination revealed that the Web designers did not always put the same words in the content title and in the HTML title. The test results are shown in figures 12 and 13. A score of 2 was assigned to pages that had "cataloging department" in the HTML titles; 1 was assigned to pages that enclosed either "cataloging" or "department" in the HTML title; and 0 to those contained neither of the two keywords.

Comparing figures 12 and 13 with figures 10 and 11, it is not hard to find the differences between the content title and the HTML title. The three highest pages had only the keyword "cataloging" in the home-page content title (figure 10). The second- and third-highest pages had changes of keywords in their HTML titles (figure 12). The seventh-highest position was held by the same institution during the ten-week test process. Surprisingly, this page did not have either keyword in its HTML title.

Half of the studied home pages in the top ten positions did not have both search keywords in their HTML titles while most of the pages ranking from eleven to twenty did. In short, the presence of keywords in the HTML title seems

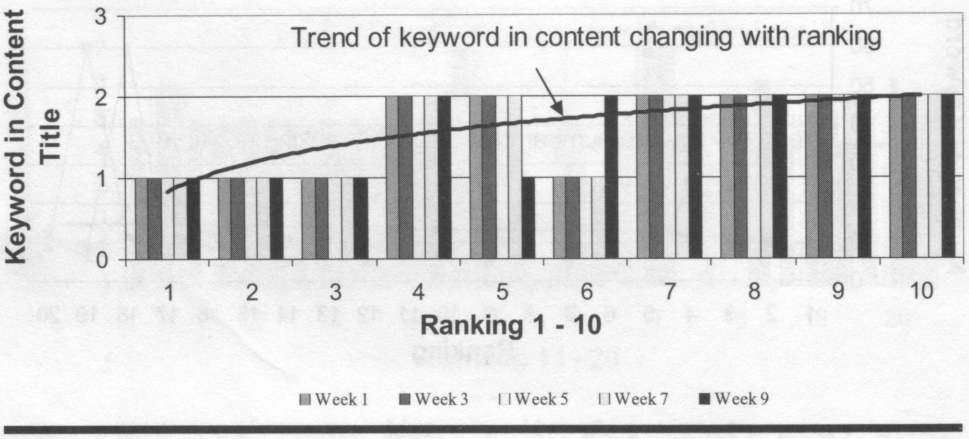


Figure 10. Keywords in Title of Home-Page Content and Ranking One to Ten in Weeks

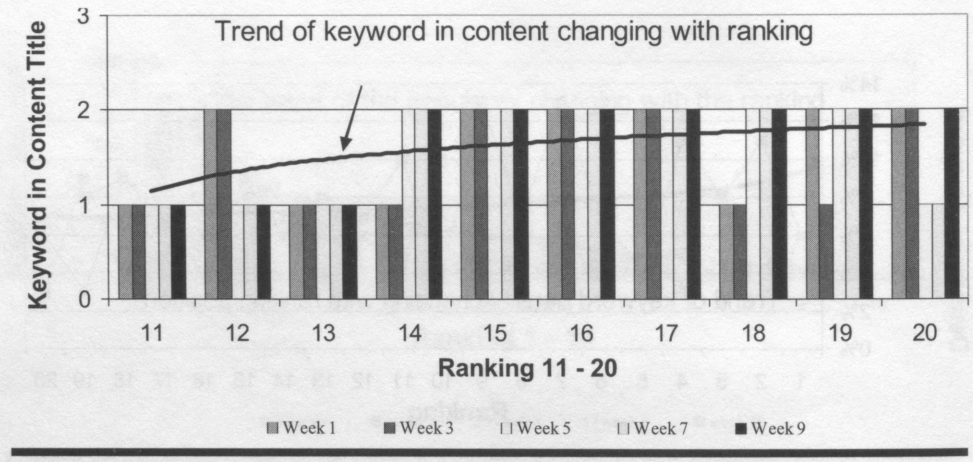


Figure 11. Keywords in Title of Home-Page Content and Ranking Eleven to Twenty in Weeks

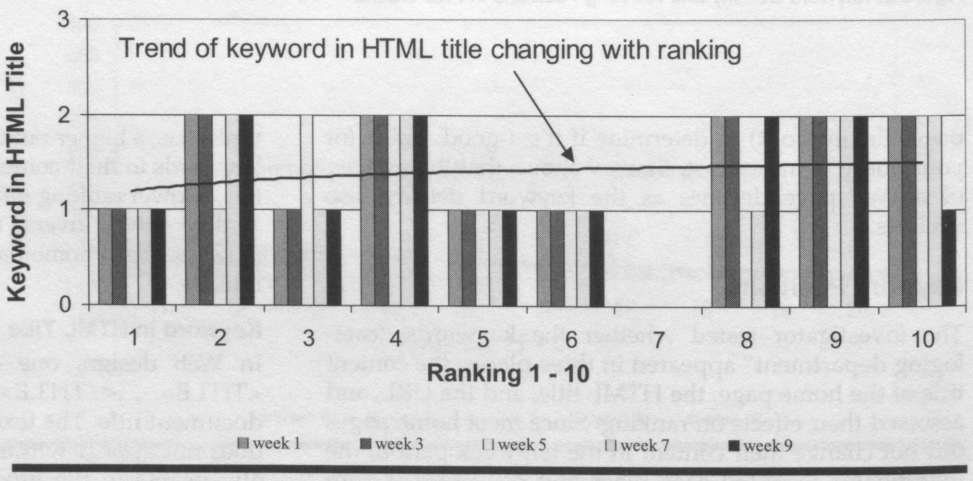


Figure 12. Keywords in HTML Title for the First Ten Positions in the Rank

to have little influence on ranking in Google retrievals.

Keywords in URL

Another keyword location possibly affecting the searching result is the URL. In figures 14 and 15, the value 2 was given to sites with "cataloging" in the URL; 1 for "catalog" in the URL; and "0.5" for "cat" or "dept" in the URL. The top number 1 in the rank did have "cataloging" in the URL. But several Web sites having lower keyword value in the URL had higher rank than those sites with higher keyword value in the URL. All sites before position fourteen had the search word "cataloging" or part of a search word in their URLs and did not change sharply. The change after the rank of fourteen became more visible and unstable. Week by week data did not reveal a strong relationship between a Web site's overall ranking and the keywords in its URL. However, it is noticeable from the overall trend of change that the keyword in the URL is affecting the ranking. The less the keyword appears in the URL, the lower the ranking.

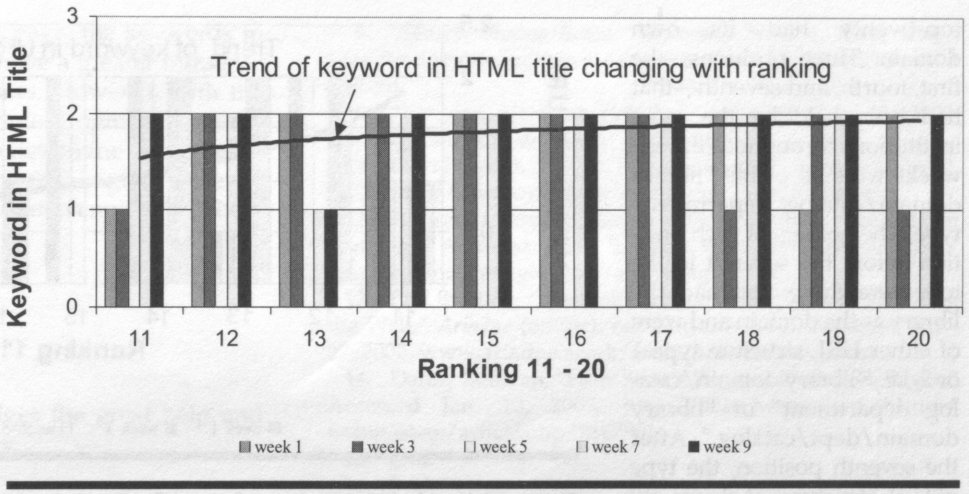


Figure 13. Keywords in HTML Title for the Second Ten Positions in the Rank

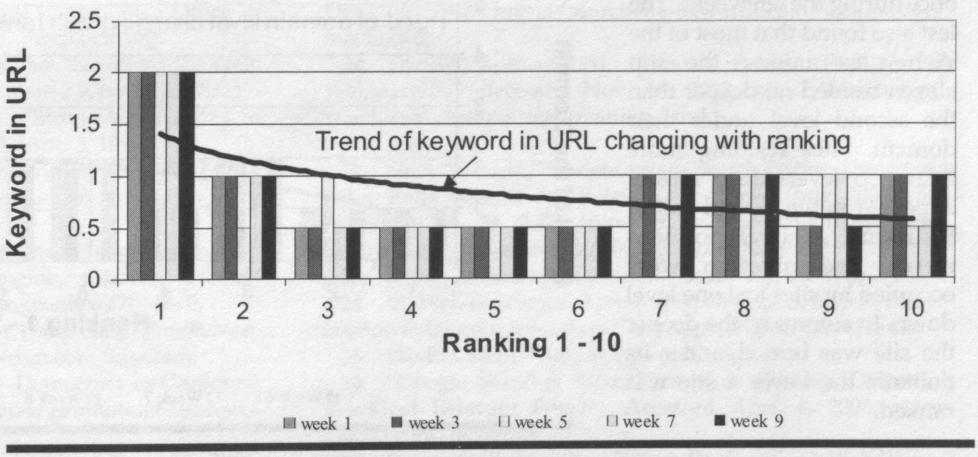


Figure 14. Keywords in URL in the First Ten Positions

The examination of keywords in the content title, HTML title, and URL shows that none of the three has a strong effect on overall rank of a Web site by Google, although a decline in URL keyword strength does coincide somewhat with overall rank.

The Domain Structure of Web Sites

The relationhip between a Web site's location under its domain and its rank in Google is the last factor the present study examined. The twenty-four tested Web sites that appeared among the top twenty retrieved by Google during the ten weeks had nine types of URL structures:

1. library domain/catalog department (7) (the number of Web sites that have such URL structure) (this is level 1, i.e., the cataloging dept. site resides one level down from its domain)
2. library domain/dept/catalog (6) (includes library domain/technical services/catalog department) (level 2)
3. library domain/libraries/cataloging (1) (level 2)
4. library domain/dept/branch/cat (1) (level 3)
5. library domain/libraries/units/dept/cat (1) (level 4)
6. institution domain/~catalog department (1) (level 1)
7. institution domain/~techserv/cat (1) (level 2)
8. institution domain/library/catalog department (2) (level 2)
9. institution domain/library/dept/catalog department (4) (level 3)

Figures 16 and 17 show the relation between the level from domains and the ranking in the last five weeks. Among the twenty-four Web sites, sixteen had their libraries' domains. No cataloging department in the

top-twenty had its own domain. Three positions—the first, fourth, and seventh—that had been held by the same institutions throughout the ten weeks were all of the "library domain/catalog department" type. The holders of each position before the seventh in the test were those that had the library as the domain and were of either URL structure type 1 or 2, i.e., "library domain/catalog department" or "library domain/dept/catalog." After the seventh position, the type of URL structures at eleven out of the last thirteen ranking positions changed more than once during the ten weeks. The test also found that most of the Web sites among the top eleven resided no deeper than the second level under their domain. Sites residing more than two levels deep mostly appeared in the second half of the results, although positions eleven and nineteen were occupied by sites just one level down. In summary, the deeper the site was buried under its domain, the lower a site was ranked.

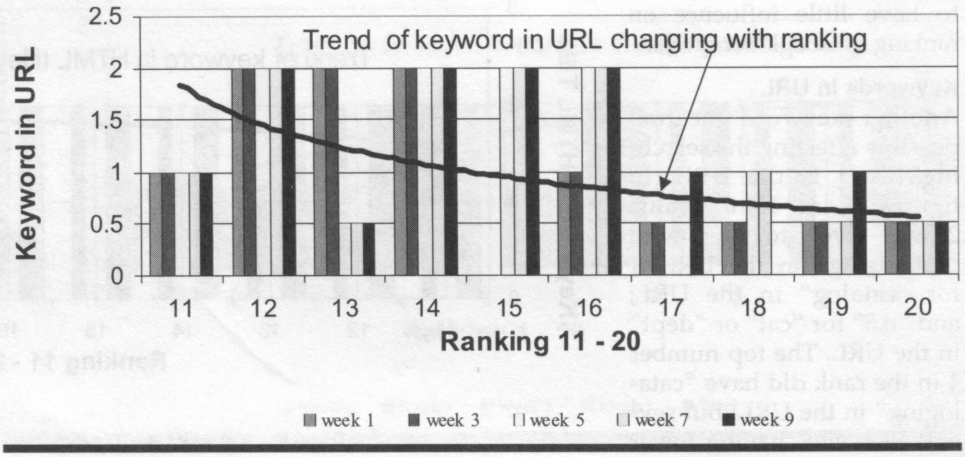


Figure 15. Keywords in URL in the Second Ten Positions

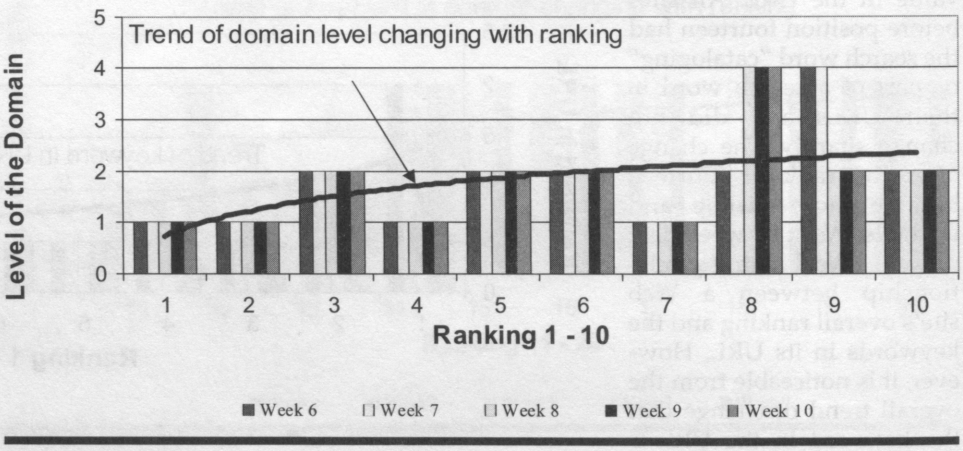


Figure 16. Domain Level and Ranking One to Ten in Five Weeks

Conclusion

The data collected in the ten weeks of this study present a complicated picture. The study's findings suggest that a higher rank in a Google retrieval requires a combination of factors. The sites with highest rankings had a better value on factors tested. Having only one or two high specifications resulted in a site barely making it to the front in the top list. Five out of the eight factors considered in this study—the Google PR, the popularity of Web sites,

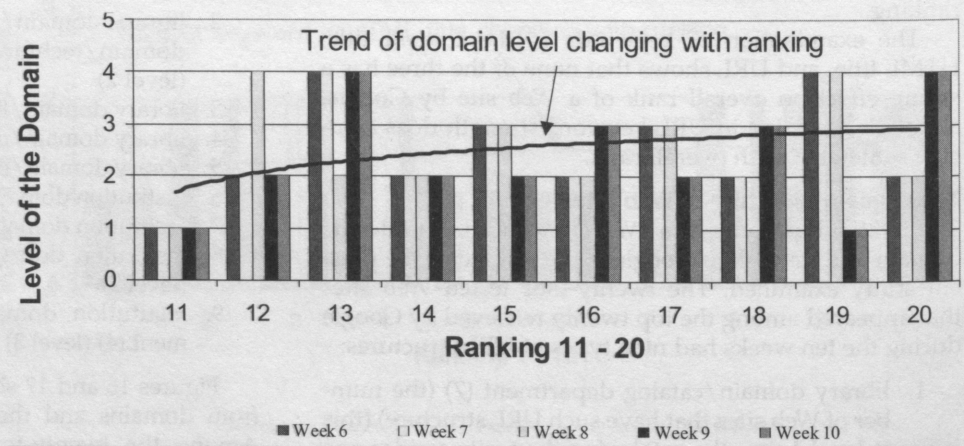


Figure 17. Domain Level and Ranking Eleven to Twenty in Five Weeks

density of keywords in the home page, the keywords in the URL, and the domain level—have a greater influence on the Web-site ranking than others. However, with the secret ranking algorithm and rapid development of search engines, it is almost impossible to determine what makes a Web site rise to the top of a search-engine retrieval. Future studies of page ranking by search engines will help to shed more light on this issue.

Acknowledgment

The author gratefully acknowledges the great help and comments of Dr. Stephen Wiberley.

References

1. OCLC Online Computer Library Center, Inc., Office of Research. Accessed Sept. 4, 2003, http://wcp.oclc.org.
2. A. Spink, B. J. Jansen, and J. Bateman, "Searching the Web: A Survey of EXCITE Users," *Internet Research* 9, no. 2 (1999): 117.
3. Bernardo A. Huberman et al., "Strong Regularities in World Wide Web Surfing," *Science* 280, (Apr. 3, 1998): 96.
4. "MarketScapes Top-10 Tips for Building Your Website." Accessed Nov. 8, 2002, www.marketscapes.com/news/ms_story5.htm.
5. A. Spink, B. J. Jansen, and J. Bateman, "Users' Searching Behavior on the Excite Web Search Engine," *CiteSeer.* Accessed Sept. 4, 2003, http://citeseer.nj.nec.com/context/70148/0.
6. "KISS Your Visitors if You Want Them Back" *Future Now.* Accessed Nov. 12, 2002, www.grokdotcom.com/kiss.htm.
7. Tony Bury, "Google PageRank—Democracy or Corporate Muscle?" Accessed Jan. 16, 2003, www.promotionbase.com/article/999.
8. "The Major Search Engines." Accessed Sept. 30, 2002, http://searchenginewatch.com/links/major.html.
9. "Search Engine Sizes." Accessed Sept. 4, 2003, www.searchenginewatch.com/reports/article.php/2156481#current.
10. "Ranking High at Google." Accessed Oct. 20, 2002, www.apromotionguide.com/google.html.
11. David Callan, "Search Tips on Google." Accessed Nov. 2, 2002, http://bucarotechelp.com/market/googletips1.asp.
12. Andrew Gerhart, "The Weight of Google PageRank in Ranking." Accessed Nov. 2, 2002, www.topsitelistings.com/searchengines/google_pagerank_weight.htm.
13. Phil Bradley, "The Relevance of Underpants to Searching the Web," *Ariadne* (online), no. 24 (June 2000). Accessed Nov. 27, 2002, www.ariadne.ac.uk/issue24/search-engines/.
14. Danny Sullivan, "How Search Engines Rank Web Pages." Accessed Jan. 10, 2003, http://searchenginewatch.com/webmasters/article.php/2167961.
15. Ibid.
16. Ibid.
17. Ibid.
18. David Callan, "Search Tips on Google."
19. Danny Sullivan, "How to Use HTML Meta Tags." Accessed Dec. 10, 2002, http://searchenginewatch.com/webmasters/meta.html.
20. Danny Sullivan, ed., "Search Engine Features for Webmasters," Accessed Nov. 10, 2002, http://searchenginewatch.com/webmasters/features.html.
21. Ibid.
22. Stacy Nowicki, "Student versus Search Engine: Undergraduates Rank Results for Relevance," *Portal* (July 2003): 510. Accessed Oct. 29, 2003, http://muse.jhu.edu/journals/portal_libraries_and_the_academy/v003/3.3nowicki.html.
23. Spink, Jansen, and Bateman, "Searching the Web," 125.
24. ISI Web of Science. Accessed April 4, 2003, http://isi0.isiknowledge.com/portal.cgi?DestApp=WOS&Func=Frame.
25. David Callan, "Search Tips on Google."
26. "Google Searches More Sites More Quickly, Delivering the Most Relevant Results. Accessed April 4, 2003, www.google.com/technology/.
27. Danny Sullivan, "How Search Engines Rank Web Pages."

Appendix A. The Twenty-four Web Sites That Made It to the Top Twenty in the Present Study

University of Virginia Library: Cataloging Services (UVa)
University of Massachusetts, Amherst: Cataloging (UMAss)
Penn State University Libraries: Cataloging Services (PASU)
Ohio State University Libraries Cataloging Department (OHSU)
University of Colorado at Boulder: Cataloging Department (UCOB)
Southern Illinois University at Carbondale Morris Library: Cataloging Group (SILUC)
Auburn University Libraries: Cataloging Department (AubU)
University of South Carolina Libraries: Cataloging Department (USC)
Tulane University, Howard-Tulane Memorial Library, Cataloging Department (TlnU)
University of Notre Dame Cataloging Department (UND)
Georgetown University Cataloging Department (GGTU)

UCLA Charles E. Young Research Library Cataloging Department (UCLA)
University of Calif. Riverside Library Cataloging Department (UCAR)
North Carolina State University libraries Cataloging Department (NCSU)
China National Library Acquisitions and Cataloging Department (CNL)
Virginia State University Johnston Memorial Library Cataloging Department (VaSU)
Haverford College Libraries Cataloging Department (HfdC)
Internet Library for Librarians Cataloging: Departmental Contacts, Policies, and Procedures (ITLFL)
Ohio State Libraries Cataloging Department Documentation (OHSC)
State University of New York at Stony Brook University Libraries/West Campus Cataloging Department (SUNYSB)
State University of New York at Buffalo Libraries Central Technical Services Cataloging Department (SUNYB)
University of Akron Libraries Cataloging Department (UAL)
Stanford University Libraries Cataloging and Metadata Services (SUL)
Boise State University Albertsons Library Cataloging Department (BSU)

lita thanks the following sponsors for participating in the LITA National Forum Sponsor Showcase

eBrary
Endeavor Information Systems
Ex Libris
Fretwell-Downing
Innovative Interfaces
Mobius
Northern Micrographics
OCLC
Serials Solutions
Sirsi
Techbooks
Veicon Technology

LITA also thanks EBSCO for their contribution in support of the Forum

Providing Access to Foreign Language Electronic Resources

Eileen Llona, Emalee Craft, Keiko Yokota-Carter, and David Pham

It may be commonly thought that CD-ROMs are a dead technology, but for many information resources produced outside of the United States, CD-ROMS may offer the only method of access to bibliographic and full-text data in non-English languages. For this discussion, CD-ROM technology refers to programs that provide bibliographic or full-text access on removable media, usually using proprietary search and interface software. Realistically, this can include DVD technology as well as older programs from floppy disks. Most products, however, presently come in CD-ROM format.

In the United States, more and more electronic resources are being made available through Web interfaces and subscriptions; by contrast, electronic resources in other countries are very often limited to CD-ROM form.[1] Electronic resources include both bibliographic databases as well as full-text digitized resources. They often include proprietary software for searching and displaying the data. Foreign-language products are written in the language of the source country, most often using localized code pages rather than Unicode. The value of these resources is often high for researchers, as they augment or replace traditional print resources. Examples include digitized newspapers from Japan, census statistics from China, digitized versions of Islamic scriptures, and election data from Russia. In the case of Japanese electronic resources, although some online full-text databases are available, CD-ROM is still the major format. Archives of newspapers, journals, and reports have been published in CD-ROMs, replacing hard copy reprints of these materials. In 2002, two major newspaper companies in Japan started publishing their archival newspaper in digital format. The whole set of one newspaper, *Yomiuri Shinbun,* consists of 126 CD-ROM disks.

For researchers in the United States, these resources are considered quite important in humanities and social-sciences research, and often represent the only feasible means for performing research on these cultures. Since primary materials are usually held by a limited number of libraries in the United States, traveling to the subject countries may be the only other alternative for researchers to do their work. Demand may be relatively low, but many resources are selective and offer valuable information or representation of cultural materials from other world regions. Since the frequency of use of these foreign-language materials outside the subject country is quite low, the cost per user is much higher than the cost of a more widely used resource such as the *New York Times* database. Not many research institutions can afford the luxury of these materials. However, the importance of holding the material is also highly valued when the alternative is to send researchers to travel to the source countries to use this material.

When obtaining resources from other countries, collection development librarians in the United States sometimes have a choice that includes print and electronic versions of a resource. While some foreign vendors publish databases on the Internet, the price for a subscription is often beyond what academic libraries can afford. Also, resources are often in the form of a one-time product representing an entire resource, bundled with retrieval and interface software (such as the *Yomiuri Shinbun* digitized newspapers). Since it is the content that is most valuable, format becomes the second choice. CD-ROM technology is still heavily used for delivery of specialized software and products.[2] The choice, therefore, is often between a print resource (such as an index or full-text resource) and its digital equivalent, if available. Since a digital form often includes superior searching methods and allows resources to be shared more readily, this format is often chosen. Cost and licensing issues must be considered, as either or both of these elements can sway a decision to stick with print, or not obtain the resource at all. If an electronic version is obtained, it may be difficult to provide easy access to the user. Programs are usually written using proprietary software and require specialized language settings. A program must be installed to view the data, so the library must make a commitment to installing and maintaining these programs and workstations. Realistically, while some of these resources do provide valuable information that is unavailable in other formats, access to that information remains a problem.

Licensing before Technology: Negotiating with Foreign Publishers

Another issue that plays a large role in obtaining electronic resources is licensing restrictions. In the case of Japanese vendors, the Japanese government has placed regulations on

Eileen Llona (ellona@u.washington.edu) is International Studies Computer Services Librarian, Digital Initiatives; **Emalee Craft** (Emalee.Craft@asu.edu) is Information Technology Coordinator, Information Technology Services; **Keiko Yokota-Carter** (kyokotac@u.washington.edu) is Japanese Studies Librarian, East Asia Library; and **David Pham** (dpham@u.washington.edu) is Senior Computer Specialist, Information Technology Services at University of Washington Libraries.

licensing agreements to protect consumers from damaged products and unethical conduct of companies. Therefore, more Japanese companies have set up legal departments and increased their sensitivity regarding the sale of their products. Licensing issues attached to electronic resources are sometimes impediments to the purchase of or subscription to Japanese (or other language) electronic resources. Universities may manage budgets to acquire materials by applying for grants, asking for donations, or even dividing payments over a period of years. However, negotiation of license issues may not be solved on an individual basis, because it deals with legal issues involving other countries.

One of the common legal issues for libraries is license agreements between a university and foreign vendors. Problems may arise over where jurisdiction lies and who is responsible for legal fees in case of legal disputes. Each provider and subscriber specifies its own jurisdiction for legal settlement. For example, the University of Washington (UW) specifies *King County, Washington State;* Japanese companies specify *Tokyo, Japan.* In addition, Japanese companies require all legal fees associated with disputes to be paid by subscribers. Usually, the institutional policies of American universities do not allow representatives to sign agreements with a foreign jurisdiction or that oblige the university to pay such legal fees, making it difficult to obtain subscriptions.

In *Yomiuri*'s case, the vendor refused to negotiate on treatment of a legal clause at the initial stage of negotiation. When communications through international telephone calls, faxes, and e-mails failed, the Japanese-collection librarian, a native speaker of Japanese, flew to Japan on a special mission on behalf of the director of the UW libraries. The librarian met the representatives of *Yomiuri* newspaper and negotiated with them personally, and was able to persuade them to delete the clause with a promise of maintaining a cordial, professional relationship with their business partners.

This face-to-face meeting was a breakthrough. The Japanese company, of course, conducts its business based upon business calculation, but before real negotiation could be established, it was essential that a high-ranking representative of the university express—in their native Japanese—the respect and faith necessary to build a trustworthy business relationship with them.

In this sense, working with foreign electronic resources not only involves technology issues, but also legal and cultural issues. Both information-technology staff and collection-development librarians offer diverse knowledge and experience that contribute in different ways towards providing access to these resources. Knowledge of the social-cultural situation of a provider's country, as well as technology issues, is necessary in bringing these resources to American researchers. Such team building between systems librarians and subject librarians is crucial to improve access to foreign electronic resources.

Common Practices Used for Providing Access to Foreign Language CD-ROM Resources in Libraries

Once the decision has been made to obtain a CD-ROM resource, libraries must provide access. Often this involves nothing more than making the disk(s) available for circulation. Often, however, these resources are quite expensive, and librarians are reluctant to circulate the materials. In such cases, libraries have to set up equipment for users to work with these resources. In the case of foreign-language media, this often involves language settings, and sometimes language-specific versions of operating systems (the Japanese version of Windows 98, for example). Individual workstations may be set up with the required language settings, often with separate workstations for each language. Obviously, this becomes unmanageable if every language is to be supported.

Foreign-Language CD-ROM Resource Access at the UW Libraries

The use of foreign-language CD-ROM resources is mainly at the East Asia Library at UW. There are about 200 CD-ROM resources in three languages (Japanese, Korean, and Chinese). In addition, Chinese resources may be in either traditional Chinese script, or simplified Chinese script. These require different language settings. All of the resources will run on some variant of the Microsoft Windows operating system.

At the UW–Seattle campus, library public PCs are managed with a one-size-fits-all approach across the seventeen Seattle-campus libraries. Part of the public PC management is a tight security model at the local desktop. While the strong security provides a stable platform for public use, as well as access control to meet licensing requirements, these security settings create configuration problems for programs requiring customized settings (such as language settings and proprietary software). In order to provide access to East Asian-language CD-ROM resources, workstations must be configured for language-specific code pages, keyboards, and font settings. In addition, much of the proprietary software conflicts with local security-policy settings. Thus, setup of local workstations and programs often requires tedious troubleshooting and configuration.

To provide access to these non-networked resources, four public workstations have been modified to accommodate language-specific settings and locally installed CD-ROM programs. This has required extensive troubleshooting due to the proprietary nature of the programs conflicting with the workstation security settings. When the public workstation model is updated, locally mounted software must be reinstalled. Obviously, this model is difficult to sustain over time.

Consider, for example, the problems involved in providing access to the *Yomiuri Shinbun* set of digitized Japanese newspapers. The Japanese-studies librarian purchased this multivolume set, which comes with a proprietary search engine and represents a very important collection for Japanese-studies researchers. Manual research methods using print and microfilm of these newspapers require many hours. The electronic version provides good search capabilities and instant access to full-page images of the newspapers. Due to the size of this valuable and expensive resource, the option of local installation on a public workstation is not feasible.

The alternative to installing proprietary CD-ROM resources on local public machines is to lend the materials to users. Problems with this solution include potential loss of materials, and requirements that users have the machines and settings needed to run the software. With a set as large as the *Yomirui Shinbun* product, circulation is not possible due to item costs and security issues.

Since the current alternatives do not lead to sustainable access, other methods of providing access to these resources have been explored. Networking is the obvious choice, with Internet-enabled access being most ideal. A review of networking options is provided by Kumar;[3] newer technologies can also be utilized to provide networking of foreign-language resources, as has been done at the East Asia Library.

Networking of Foreign-Language CD-ROM Resources at the UW Libraries

English-language CD-ROM resources have been made available over the UW libraries network for several years. CD-ROM resources at UW have evolved in several stages. In 1999, UW replaced its locally developed X-Windows-based Willow information system with its Information Gateway (the Gateway).[4] During the employment of Willow as the libraries' primary information search interface, CD-ROMs were not generally offered as networked resources. With the inception of the Gateway, however, access to electronic resources of varied types was integrated through the use of a digital-resource registry.[5] From the time of the Gateway's inception until early 2003, CD-ROM-based products, mostly English-language resources, were offered online through a hybrid implementation of Windows-based technologies. The CD-ROM images were copied in full to a Windows NT-based network. The interface for CD-ROMs was set up on client machines located in the libraries' public areas. Any proprietary software that came as part of a CD-ROM was loaded onto a central public image and then distributed via image casting to hundreds of public PCs.

Most CD-ROM resources, and all other electronic resources, were organized through the Gateway. Through a series of HTML-embedded scripts, a user connected to a CD-ROM application by launching it from the Web-based Gateway. These scripts would open onto the desktop with locally installed application software and appear to the user to be very similar to using a CD in the local drive. The scripts, along with network share limits, also were employed to insure that individual licensing agreements were met.

Although this system took advantage of the technology of the time and allowed centralized management of hundreds of electronic resources, it had serious drawbacks supporting CD-ROM resources. It was difficult or impossible to update any local software once the public image was prepared and deployed, and most notably for international studies resources, almost no support was possible for non-roman-character-based (NRCB) programs.

Beginning in 2002, the libraries began to investigate other technologies to support non-Internet-based electronic resources, including NRCB products. Starting with a system for primarily English-language items, the libraries set up a Windows 2000 server with Terminal Services in Application Mode and went live on public workstations in early 2003. This most recent networked setup is similar to versions used in libraries since the mid-1990s when Microsoft licensed CITRIX and developed the technology further.[6] Preparing for Terminal Services included installing Windows 2000 Advanced Server with Terminal Server set to Application Mode.

Application Mode means that server computing resources and licensing are optimized to allow multiple simultaneous client connections. Various authentication and resource-specific configurations can be set on the server side, allowing for security settings without interfering with program-specific requirements. The most important contribution of Terminal Services in providing access to foreign-language electronic resources is the ability to set all of the language configurations at the server, once. Fonts, input methods, and code pages are set once on the server, and delivered to the user via thin client architecture. Thus, the user's machine does not need to have

language-specific settings or software. Maintenance on public workstations is significantly reduced, and security is improved. Most importantly, access to these resources is vastly improved.

The Gateway continues to be the primary interface for English-language CD-ROM resources. However, instead of launching in the local public PC desktop with locally installed software, links to CD-ROM-based resources launch a universal ActiveX Terminal Services client from within Internet Explorer, so that users with an Internet connection can access networked resources after authenticating. Because network domain permissions can be employed with the Terminal Services client, domain groups make it easier to manage resource-licensing restrictions.

While Terminal Services is a great advance in providing anytime-anywhere access to foreign-language CD-ROM resources, there are constraints. Since most non-English resources are still produced using language-specific code pages (not Unicode), switching from one language to another is not yet possible without rebooting the server. Hence, language-specific servers must be set up and maintained. At the UW libraries, one Japanese server running Terminal Services has been set up, and a simplified Chinese server currently is being configured as well. While machine costs are relatively low, ongoing maintenance and labor remains a significant consideration in deploying Terminal Services for several languages. At the present, there are no immediate solutions to this problem. While the Unicode character set is available, there are few non-English resources that utilize this character set. Since code pages are loaded at boot time in the Windows environment, switching between languages and code pages requires a reboot. It is hoped that future implementations of the Windows operating system will provide for simultaneous-code page support.

Conclusions

While many electronic resources are being delivered via Internet subscriptions, libraries are still faced with providing access to CD-ROM resources, particularly for foreign-language-based materials. Thin client technology such as Terminal Services goes part of the way to improving access to these one-off resources. However, providing access to resources in multiple languages still presents problems, and not all languages can be served equally. Libraries must make decisions, both at the collection development and the information-technology levels, on which languages to support and how to support them. As Unicode becomes more widespread, it is hoped that language-specific programs will utilize this technology, thus making language-specific requirements less cumbersome.

Migration to Internet-based resources continues in the United States, and is becoming more common in other countries. However, issues of preservation and access will remain cloudy for this format for some time to come. If subscription prices become too high, or publishers go out of business, what becomes of the resource? On the other hand, the proprietary nature of programs and data made available on removable media such as CD-ROMs presents its own set of sustainability questions. In five or ten years, will operating systems be available to read currently produced programs? What commitments do libraries need to make to keep legacy equipment and software available for such resources?

In the foreseeable future, it appears likely that libraries will have to continue to support standalone resources that represent unique data sets and cultural-heritage data. Business models that provide instant Internet access are not feasible for many smaller companies and countries. Delivery mechanisms for providing such data are most likely to continue to be through proprietary programs and software that can be packaged and delivered on a one-time basis. In order for libraries to provide access to these often rare, unique data sets, some infrastructure should be in place for making such access sustainable.

Acknowledgments

William Jordan and Michael Reynolds of the UW Libraries Systems Department provided extraordinary technical support for providing the networking solutions described.

References

1. Ellen H. Hammond, "Japanese Electronic Database Use in North America: Widespread and on the Web: Report of the Consortial Licensing Task Force, North American Coordinating Council on Japanese Library Resources (NCC)." Accessed November 13, 2003, www.library.pitt.edu/ncc/CLTFSurvey2000.pdf.

2. Marshall Breeding, "Does the Web Spell Doom for CD and DVD?" *Computers in Libraries* 19, no. 10 (1999): 70–75; Barbara Burg and Amy M. Kautzmann, "Web versus CD-ROM: Format Follows Function," *Searcher* 6, no. 1 (1998): 59–63.

3. K. Satish Kumar and P. O. Jayaraman, "Networking of CD/DVD ROMS: An Overview," *DESIDOC Bulletin of Information Technology* 22, no. 6 (2002): 3–12.

4. Debra S. Ketchell et al., "Willow: A Uniform Search," *Journal of the American Medical Informatics Association* 3, no. 1 (1996): 27–37.

5. William Jordan, "My Gateway at the University of Washington Libraries," *Information Technology and Libraries* 19, no. 4 (2000): 180–86.

6. F. Bartow Culp, "Streamlining Electronic Access with Windows Terminal Server," *Library Hi Tech* 20, no. 3 (2002): 302–7.

Will Libraries' Web-based Survey Methods Replace Existing Non-Electronic Survey Methods?

Gay Helen Perkins

This paper reviews and tabulates the literature of the strengths of Web-based survey methods versus non-electronic survey methods in educational, business, and other institutions. The relative importance of the strengths of each method is examined. As indicated by Perkins and Yuan's study comparing Web-based and non-electronic library-satisfaction survey results, earlier nonlibrary, institutional findings may have applications in the library. Of particular current interest is survey respondents' representation of a large sample at Western Kentucky University (WKU) for the Web-based Association of Research Libraries (ARL) LibQUAL+ survey. Finally, there is a summary of comments from a small sample of respondents from three library electronic discussion lists—the Library and Information Technology Association (LITA-L), the Association of College and Research Librarians (ACRL-FRM), and the Reference and User Services Association (RUSA-L) to discuss how electronic/Internet survey methods may, in part, replace current non-electronic survey methods.

Review of the Strengths of Web-based versus Non-Electronic Surveys

During the last few years, there have been several reviews and discussions of the strengths and limitations of Web-based and non-electronic surveys. These include those that are applicable to library and information science; colleges and universities; secondary-school education; personnel and other applied settings; and of course, general survey and survey-research industry.[1] The appendix summarizes some of the more generalizable strengths for the two types of surveys, and presents these strengths according to the survey instrument, the use of computer or other access, sampling of respondents, administration and procedures of the researcher, and transmission and feedback of the data analysis.[2]

Relative Importance of the Strengths of Each Method

To simplify and make the appendix more useful to the practitioner, the author asked the director of institutional research at Western Kentucky University (WKU) to rank, based on his experience, the five most important strengths of Web-based surveys and the five most important strengths of non-electronic surveys.[3] The five most important strengths of Web-based surveys were as follows: (1) the sampling instrument is available 24-7 at a location convenient to the respondent; (2) less time required for delivery of the instrument to participants, the administrator's receipt of responses, data entry, and analysis and feedback; (3) the Web-based instrument allows inclusion of text, images, and sound; (4) direct and accurate electronic transmission (coding and analysis) of quantitative and qualitative data; and (5) potential for customized feedback, including survey results, supplied directly to the respondent, and based on response content.

The five most important strengths of non-electronic surveys were as follows: (1) computer literacy by the participants is not required, (2) fewer concerns about preservation of confidentiality than when a Web-based instrument is used; (3) the greater control that can be exercised over the administration of non-electronic instruments in a classroom and the correspondingly higher rate of completed responses; (4) some individuals in a sample may not have access to the hardware and software required to complete a Web-based instrument; and (5) a non-electronic instrument is generally more accessible to groups with limited financial resources, some racial and ethnic minorities, individuals of lower education, older age groups, or others who may be underrepresented because they lack computer or Internet access.

LibQUAL+ at WKU: A Discussion of Representation

To study a sampling characteristic of a Web-based survey, the author looked at the LibQUAL+ survey's representation of demographic groups, compared with that of the general population. As an example, 308 universities, including WKU, participated in the 2003 administration of the ARL's LibQUAL+ survey, a twenty-five-item, Web-based assessment instrument which measures library service quality.[4] The ARL and Texas A&M provided a central server with a scalable Web interface, and WKU libraries, with the help of WKU's institutional research and administrative computing departments, sent the LibQUAL+ instrument to randomly sampled e-mail addresses representative of the user population: 1163 undergraduate students (33.83 percent), 802 graduate students (23.33 percent), 738 faculty members (21.47 percent), and 735 staff members (21.38 percent). Only 292 WKU participants, or 8.49 percent, responded. The representation of the respondent sample was as follows: 101 undergraduate students (34.59

Gay Helen Perkins (Gay.Perkins@wku.edu) is Professor and Business Librarian at Western Kentucky University Libraries.

percent), 69 graduate students (23.63 percent), 87 faculty members (29.79 percent), and 35 staff members (11.99 percent). One can readily see that in the final respondent sample, undergraduate and graduate students were fairly represented. However, faculty members were overrepresented, and staff members were underrepresented. Thus, statistics based on total WKU data might be biased.

Electronic Discussion Group Members' Discussion of Electronic/Internet and Paper-and-Pencil Survey Administration in Libraries

Finally, to give information about what librarians considered strengths of electronic/Internet and paper-and-pencil surveys, members of three library electronic discussion groups were e-mailed the question, "Will administering a survey questionnaire via the Internet (both Web-based and e-mail platforms) replace the paper-and-pencil administration in libraries?" Out of twenty usable responses by members of the three electronic discussion groups, there were seven favorable comments about electronic/Internet surveys, five favorable comments about paper-and-pencil surveys, and eight favorable comments about each of the two categories. Additionally, there were two "other" responses unrelated to the topic at hand.

Favorable comments about the Internet survey mode included (1) higher probability of respondents' completing one or two pop-up online questions; (2) fast, meaningful data analysis; (3) more effective administration and compilation of surveys; (4) use of online surveys for interest in walk-ups to library; (5) continuation of an individual's usual response style, whether online or on paper; and (6 and 7) two very brief positive responses.

Favorable comments about the paper-and-pencil survey mode included (1) the more personal touch and greater probability of responding; (2–4) responding without need for computer literacy or assistance (three comments); and (5) respondents' common use of Internet for e-mail only, rather than for accessing home page and survey.

Favorable comments about each of the Internet versus paper-and-pencil survey modes included (1) online survey system's more efficient quantitative and qualitative data analysis versus paper-and-pencil's readiness for walk-ins without waiting to use terminals; (2) LibQUAL+'s ability to coax responses from students motivated by the prospect of a Palm Pilot prize, versus students' tendencies to delete surveys distributed via mass e-mailings; (3) e-mail's greater use for correspondence versus deletion of bulk e-mails; (4) ease of computer use versus paper-and-pencil's responses without need for computer literacy or comfort; (5) need for both modes to reach all segments of population; (6) optional computer kiosk for surveys versus paper-and-pencil designed for responding at given time and place; (7) need for all survey modes since some population segments irritated by e-mail and surveys; and (8) online surveys' use by students in academic libraries versus paper-and-pencil or face-to-face modes in public libraries where some patrons do not use computers or bypass home page with surveys.

As readily noted, there have been many reviews and discussions of the strengths of both electronic/Internet and non-electronic surveys. These reviews cover applications in library and information science, colleges and universities, secondary-school education, personnel and applied settings, and the general survey-research industry, and they make a strong case for advantages of each type of survey. Further, Perkins and Yuan's study comparing Web-based and non-electronic library-satisfaction survey results points out some of the strengths and limitations of a Web-based survey, and suggests that nonlibrary, institutional findings may have applications in the library.[5] A small sample of respondents from three library electronic discussion groups suggests that some librarians are not yet ready to give up on the paper-and-pencil survey mode.

In conclusion, findings mentioned in the tabled literature review, in the use of Web-based surveys at WKU, and in the library electronic discussion groups' responses would suggest that Web-based surveys may have the potential to replace non-electronic surveys where 24-7 administration in varied locations and fast data entry, analysis, and feedback are paramount. However, where certain population segments cannot access computers or their specific software applications, where specific response rates are necessary, and where human resources may not have computer literacy, non-electronic surveys will still be necessary in the near future.

Acknowledgments

The author thanks Bob Cobb, director, WKU institutional research, for ranking the strengths of Web-based and non-electronic surveys for this manuscript. Thanks also go to Michael Binder, dean, WKU libraries and Brian Coutts, head, WKU Department of Library Public Services for reading the manuscript. Thanks to Charles H. Smith, science librarian, WKU Libraries, who suggested asking library electronic discussion group members about Web-based surveys replacing paper-and-pencil surveys; and to the electronic discussion group respondents for participating. The author also thanks Consuella Askew Waller and ARL. Finally, the author thanks Ruth Carman and Christina Antoine for the excellent word processing.

References

1. Yin Zhang, "Using the Internet for Survey Research: A Case Study," *Journal of the American Society for Information Science* 51, no. 1 (Jan. 1, 2000): 57–68; John H. Krantz, Jody Ballard, and Jody Scher, "Comparing the Results of Laboratory and World Wide Web Samples of the Determinants of Female Attractiveness," *Behavior Research Methods, Instruments, and Computers* 29, no. 2 (1997): 264–69; Karen A. Pasveer and John H. Ellard, "The Making of a Personality Inventory: Help from the WWW," *Behavior Research Methods, Instruments, and Computers* 30, no. 2 (1998): 309–13; Robert N. Davis, "Web-Based Administration of a Personality Questionnaire: Comparison with Traditional Methods," *Behavior Research Methods, Instruments, and Computers* 31, no. 4 (1999): 572–77; Philip G. Handwerk, Cristi Carson, and Karen M. Blackwell, "Online versus Paper-and-Pencil Surveying of Students: A Case Study," Association for Institutional Research (AIR) 2000 annual forum paper, ERIC, ED 446512; Giuseppe Riva, Tiziana Teruzzi, and Luigi Anolli, "The Use of the Internet in Psychological Research: Comparisons of Online and Offline Questionnaires," *CyberPsychology* 6, no. 1 (Feb. 2003): 73–80; Craig Mertler, "Demonstrating the Potential for Web-Based Survey Methodology with a Case Study," *American Secondary Education* 30, no. 2 (Spring 2002): 49–61; Jeffrey M. Stanton, "An Empirical Assessment of Data Collection Using the Internet," *Personnel Psychology* 51 (1998): 709–25; Allen I. Kraut and Lise M. Saari, "Organizational Surveys: Coming of Age for a New Era," *Evolving Practices in Human Resources Management*, eds. Allen I. Kraut and A. K. Korman, (San Francisco: Jossey-Bass, 1999); Allen I. Kraut, "An E-mail Letter to a Friend," *MetroNews*, (New York: The Metropolitan New York Association for Applied Psychology, 2001); Lori Foster Thompson et al., "From Paper to Pixels: Moving Personnel Surveys to the Web," *Personnel Psychology* 56, no. 1 (Spring 2003): 197–228; William C. Schmidt, "World Wide Web Survey Research Benefits, Potential Problems, and Solutions," *Behavior Research Methods, Instruments, and Computers* 29, no. 2 (1997): 274–79; Mick P. Couper, "Web Surveys: A Review of Issues and Approaches," *Public Opinion Quarterly* 64 (2000): 464–94.

2. Ibid; UCLA Center for Communication Policy, "The UCLA Internet Report: Surveying the Digital Future—Year Three." Accessed Feb. 2003, www.ccp.ucla.edu.

3. Personal communication with Bob Cobb, Director of Institutional Research, Western Kentucky University, Nov. 11, 2003.

4. Association of Research Libraries, "LibQUAL+—Charting Library Service Quality." Accessed Oct. 22, 2003, www.libqual.org.

5. Gay Helen Perkins and Haiwang Yuan, "A Comparison of Web-Based and Paper-and-Pencil Library-Satisfaction Survey Results," *College and Research Libraries* 62, no. 4 (2001): 369–77.

Appendix. Strengths of Web-based Versus Non-Electronic Surveys

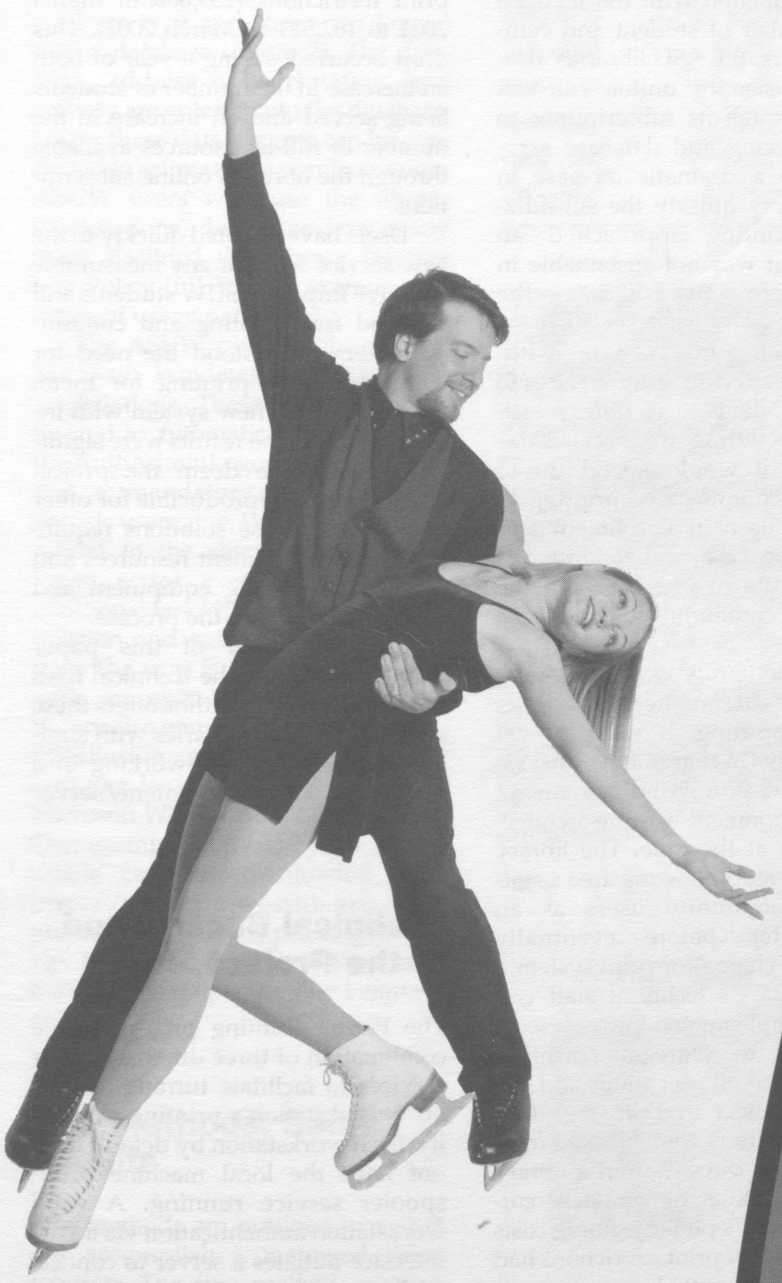
Web-based Survey Strengths	Non-Electronic Survey Strengths
Instrument	
Instrument allows text, image, and sound transmission.	Creation of non-electronic instrument easier, without development of Web pages, potential respondents' database.
Potential for Web-based interactive instrument that allows viewing relevant questions only.	Better control over non-electronic instrument's standardized appearance, without browser settings, processors, monitors, hardware platforms.
	Better control of non-electronic instrument's standardized testing environment, with consistent time and location and no multiple hardware, software, Internet connections.
Computer Access	
	Non-electronic instrument can include groups with limited financial resources, some racial and ethnic minorities, individuals of lower education, and older age groups who are underrepresented because they do not have computer or Internet access.
	Entire sample can access non-electronic instrument, whereas some individuals may not have access to Web-based instrument's hardware, software applications.

continued on page 126

Appendix. Strengths of Web-based Versus Non-Electronic Surveys (continued)

Web-based Survey Strengths	Non-Electronic Survey Strengths
Sampling	
Web-based instrument accesses larger and geographically broader sample (71.1 million U.S. users at home or outside home —2002) with desired special interests for better generalization of results.[1]	Control of sample representative of general population, with non-electronic administration (not self-selection of Web addresses for individuals or electronic discussion groups). Control of non-electronic instrument completion and higher complete response rate in classroom. Control of single non-electronic instrument with single response, by individual respondent or member of group without sophisticated software. Of those not comfortable with Web-based instrument, concern about confidentiality of Web-based instrument.
Administration-Human Resources	
Fewer human resource services—less photocopying of instrument; mailings; typing; scanning; cleaning, coding, and verifying of data.	Fewer labor start-up costs of hardware and software purchasing, programming, training. Computer literacy not required.
Administration—Time Resources	
Instrument available to sample 24-7 and at location convenient to respondent. Shorter time for delivery of instrument to sample, administrator's reception of responses, data entry and analysis/feedback.	Lower risk, fewer uncertainties with technical glitches.
Administration—Materials Resources	
Fewer material resources of paper, ink, mailing. Lower laboratory, equipment, and other material costs. At later stage of project, investment of materials, costs of actual hardware, software, and computer network space may pay off.	At start-up of project, non-electronic instrument has lower material costs.
Analysis	
Direct transmitting (coding and analysis) of quantitative and qualitative data electronically and more accurately. More complete replies to open-ended questions. Potential for customized feedback, including survey results, directly to respondent, based on response content. No meaningful differences in psychometrics (factor structure, reliability) of Web-based versus non-electronic instruments. When given the choice of Web, non-electronic modes, little difference in respondents' mean ratings. Potential to accurately and directly transmit survey results to a Web site.	No meaningful differences in psychometrics (factor structure, reliability) of Web-based versus non-electronic instruments. When given the choice of Web, non-electronic modes, little difference in respondents' mean ratings.

Together



our Web-based services

meet your information access and management needs. EBSCO*host*® provides full text and secondary research databases in a wide variety of subjects, including a new comprehensive TOC database. Extensive linking capabilities integrate our databases and our electronic journals service. Our new A-to-Z service puts your titles in one convenient list so your patrons can quickly see what's available to them and easily link to the content they need.

EBSCO's Web-based services — working together for you.

60 years 1944-2004

EBSCO INFORMATION SERVICES

CUSTOMERFOCUSED **CONTENT**DRIVEN

www.ebsco.com

205.991.6600

Tutorial

Segregating Library Users in a Microsoft Windows Client/Server Environment to Control Access to Public Printers

Perry C. Horner and Dennis Isbell

In response to the escalating costs of offering free printing to library users, the Arizona State University West (ASUW) Library's technical staff devised a low-cost solution to allowing ASU users to log in and gain access to the library's printers while excluding non-ASU users from the printers. The solution was created in a Microsoft Windows client/server environment. A detailed description of the solution and scripts are offered to other libraries in the same environment that might find it useful.

Arizona State University (ASU) West (ASUW) is a new campus in northwest Phoenix. Campus enrollment has grown steadily over the years and now stands at 6500 students, an increase of 12 percent from the year before. That increase has had a particular impact on the ASUW library's practice of offering free unlimited printing to library users from both the university and the community. Since its beginning in 1998, the library has subsidized printing for library users as a service primarily to ASUW students, but also as a courtesy to the community (estimated to make up almost 30 percent of users).

In conjunction with the increase in the number of student and community users, the ASU libraries' dramatic increase in online full-text sources through its subscriptions to online indexing and database services led to a dramatic increase in printing. Very quickly the subsidizing of printing approached an amount that was not sustainable in the long-term. By fall 2000, the library decided it would have to reduce the amount of printing without directly affecting print services to enrolled students. The library estimated that during the fiscal 2001–2002 year, it would spend up to $60,000 to continue free printing to all library users if enrollment and online full-text demand continued at the same rate of increase, and the number of community users stayed constant.

Since the library was developing a Linux-based architecture for its public computing, it was not yet ready to install a charge-for-print system. A Microsoft Windows client/server environment was the architecture in use at the time. The library decided to stop allowing free printing for community users as an interim step before eventually installing a charge-for-print system.

The library's technical staff created and implemented low-cost technical fixes to segregate printing services based on user status and put them into effect in February 2002. Community users were blocked from print services directly from a library workstation, thus immediately cutting the library's public printing costs almost in half. If print restrictions had not been implemented, printing costs were projected to increase to approximately $60,000 in 2001–2002. Instead, costs decreased to $33,473 during 2001–2002 after print restrictions were implemented (table 1). One snapshot of the change was a 40 percent decrease in print jobs between March 2001, when there were no print restrictions, and March 2002, the first full month the library operated with print restrictions (259,000 in March 2001 to 102,521 in March 2002). This drop occurred during a year of both an increase in the number of students being served and an increase in the number of full-text sources available through the library's online subscriptions.

Users have adapted quickly to the new service without any measurable negative impact. ASUW students still received free printing and community users understood the need for eliminating free printing for them, adjusting to the new system with little complaint. The results were significant enough to deem the project successful and reproducible for other institutions. These solutions require minimal development resources and can utilize existing equipment and services to facilitate the process.

The remainder of this paper describes in detail the technical fixes implemented. The author offers these solutions to other libraries with similar problems that are working in a Microsoft Windows client/server environment.

Technical Background of the Project

The Patron Printing project uses a combination of three different server services to facilitate turning on and off a workstation's printing capability. Each workstation by default does not have the local machine print-spooler service running. A valid workstation authentication via a Web interface initiates a server to connect to the requesting workstation and start the local print-spooler service for that workstation. If the user does not perform any activity on the workstation during a fifteen-minute period, the workstation automatically logs itself off and turns off the print-spooler service. The worksta-

Perry C. Horner (p_horner@asu.edu) is Senior Manager of Library Technology Support and Development and **Dennis Isbell** (dennis.isbell@asu.edu) is Associate Librarian and Chair of Reference Services at Arizona State University West Library, Phoenix, Arizona.

tion will then log back on and present the user with a logon screen. This activity also takes place if a user initiates the logout process.

User type and status are logged into a database (figure 2). The date, time, address of workstation, and activity are entered into the database. Using these data, reports can be generated to estimate the number of non-ASUW users who use the library workstations, determine how many users declined the Library Internet Use Policy (LIUP), and examine histories of usage patterns.

The ASUW campus library uses Microsoft Windows NT/2000 client workstations. These clients are configured to automatically logon using the AutoAdminLogon registry setting and a mandatory profile with a generic username. When the profile is served to the client workstation, a logon script executes performing additional tasks such as time synchronization and print-spooler deactivation. The user is normally presented with an open Web browser. The library also employs a combination of ColdFusion and ActiveState Perl scripting languages being served on Microsoft Windows NT/2000 servers. The example code supporting this article can be downloaded from http://wlibinfo.westlib.asu.edu/publish/downloads/printingscripts.zip. With little trouble, one can port the example scripts to either language or even PHP.

Two Methods, Same Results

Two methods are outlined here that can accomplish a managed print solution. The first method requires less investment but limits the architecture to Microsoft products. The second method requires more investment in script programming but allows for non-Microsoft browsers and Web servers. The following details the processes and steps to be

Table 1. Estimate of Costs for Public Printing for One Full Year

FY	Costs	% of Cost Increases	Quantity of Prints
2002–2003	$ 24,557	-27%	--
2001–2002	$ 33,473	-13%	1,097,515
2000–2001	$ 38,679	42%	2,069,607
1999–2000	$ 27,274	32%	1,468,075
1998–1999	$ 20,703	397%	874,973
1997–1998	$ 4,167	--	--

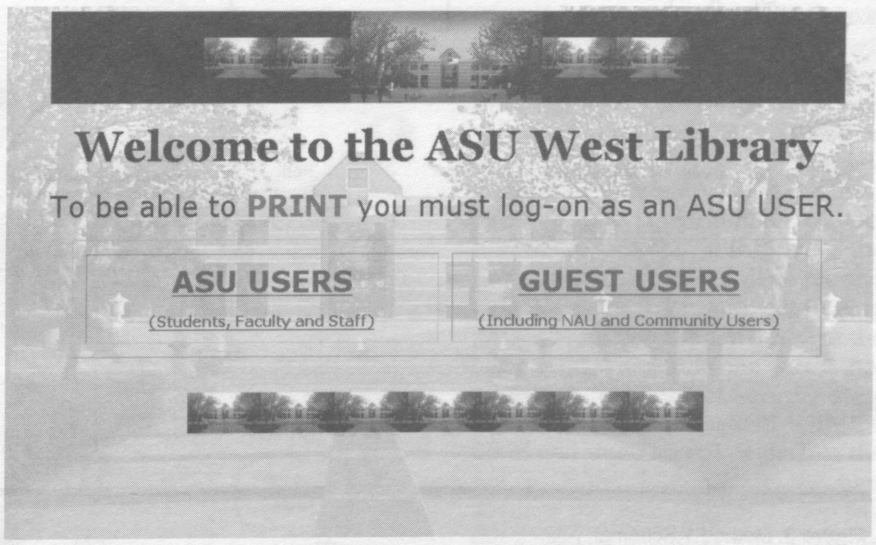


Figure 1. Sample Welcome Web Page

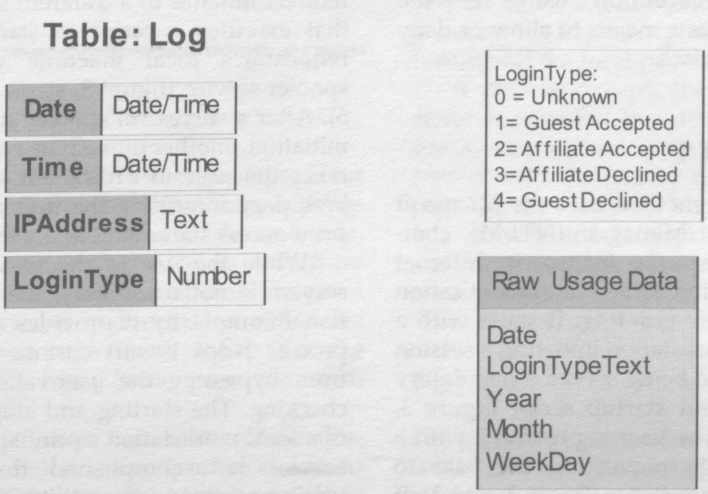


Figure 2. Usage Logging Data Schema

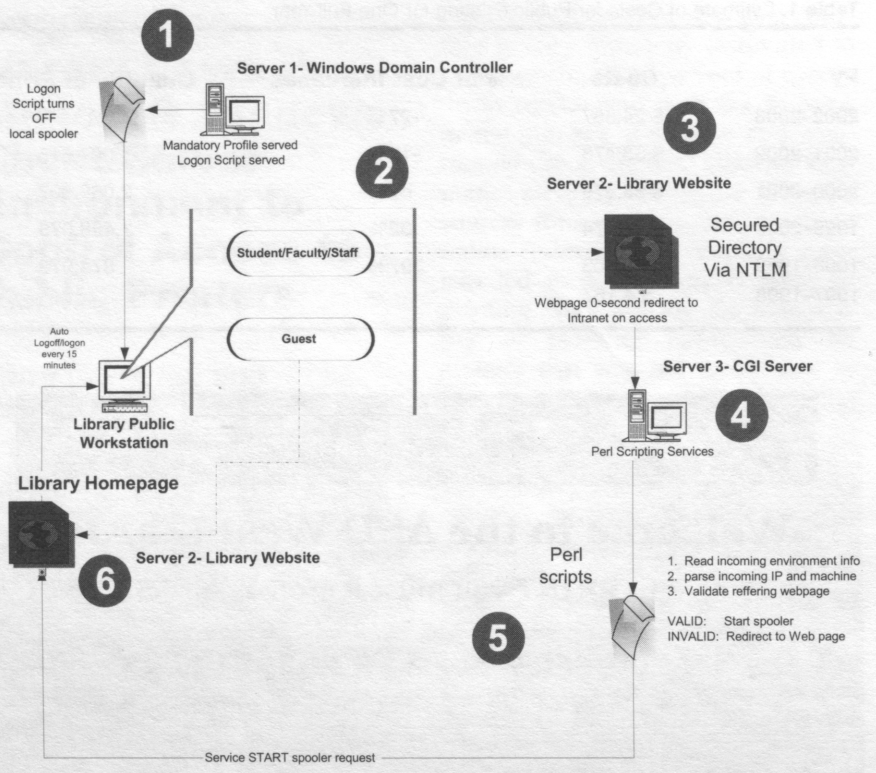


Figure 3. Method 1 Schema

taken to implement a relatively cost-effective solution using a Web browser as a means to allow or deny print services.

Method 1

The first method uses the Microsoft NT Lan Manager (NTLM) challenge/response feature in Internet Information Server for authentication and access granting. It starts with a client workstation initiating a session logon and being served a mandatory profile and startup script (figure 3, step 1). The user is presented with a Web page requesting the user to identify her status (figure 3, step 2). If a user successfully authenticates to the NTLM-secured Web page (figure 3, step 3), an immediate client-page redirect initiates to a different server that executes a script to start the requestor's local machine print-spooler service (figure 3, steps 4 and 5). After a successful spooler startup initiation, another immediate redirect takes the user to a different server Web page informing the user of the print-access status (figure 3, step 6).

While the use of the two Web servers is not mandatory, the additional complexity it provides to the process helps thwart curious users from bypassing the user-validation checking. The starting and stopping of a local workstation's print-spooler service is accomplished through using a command-line utility SC.EXE available in the Microsoft Windows NT and Windows 2000 Resource Kit. This allows scripts the ability to pass information to the SC utility through a single-line command call.

Users failing to perform a successful authentication will not have printer access. A return error page could be used here to redirect the user to a page indicating the user has input an invalid username or password and therefore does not have access to a printer.

The addition of a client workstation screensaver application (LOG OUT.SCR) could be used to logout the workstation after a set amount of time of no activity (keyboard or mouse). This would make the next logon initiate the shutdown of the spooler service and present a fresh Web logon screen.

Method 2

The second method is a bit more complex but allows for a remote directory (authentication) lookup and does not require a pure Microsoft architecture. The client requesting the spooler service can be authenticated against an account server that is separate from the Web server where the access pages are located. This method also includes a means to verify the client workstation requesting the spooler service against a list of allowed workstation IP addresses. A benefit of using this method is the ability to use a non-Microsoft Web browser and Web servers. Method two also uses a custom CGI application called WebAuth that is not covered in this article. WebAuth is a socket server that receives incoming socket data (username, password, and redirecting Web page) via a secure-socket connection. The WebAuth server then performs a Kerberos account-lookup using the supplied username and password. If the request is valid, then the WebAuth server will forward the client browser to the supplied redirecting Web page. With little trouble, users can write their

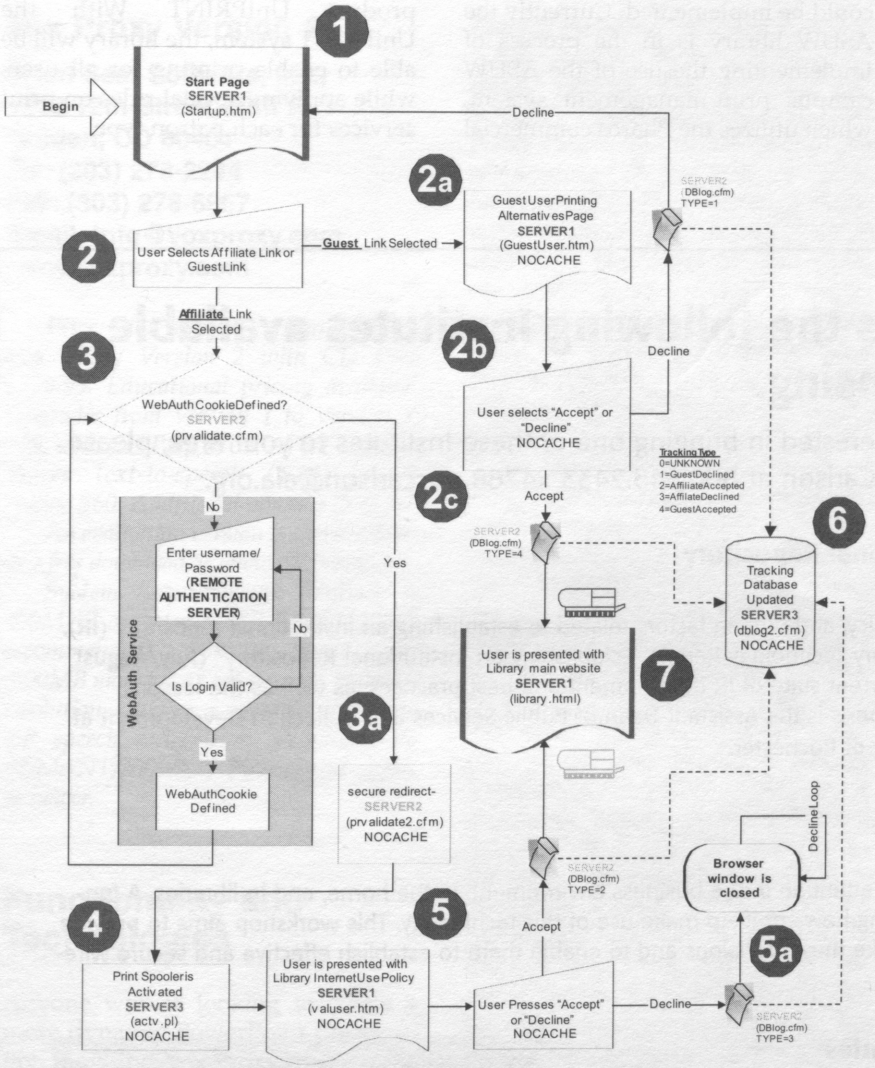


Figure 4. Method 2 Schema

own applications to perform a service similar to this.

With method two, the workstation employs the same automatic logon and profile serving described in method one. The user is also presented with a similar user-identification Web page as with method one (figure 4, step 1). If the user logs on as a guest user, a Web page is accessed describing the print-access restrictions and alternative sources for hardcopy output (figure 4, step 2a). In addition, the user is presented with the LIUP and prompted for accepting or declining the policy (figure 4, steps 2b and 2c). In each case, the action is recorded to a database (figure 4, step 6) and then the client browser is redirected to either the starting page (figure 4, step 1) or the library main home page (figure 4, step 7).

If the user logs on as an ASU affiliate in step 1, a script then checks to see if the user has already authenticated for the print service. This is accomplished by the CGI server checking for a valid cookie on the requestor's workstation. If the session cookie is still valid, then the user is redirected to a different CGI server for spooler activation and does not need to authenticate again (figure 4, step 3a). If the user does not have a valid session cookie, a valid username and password must be entered for the user to be served a new cookie (figure 4, step 3). This step should never happen but is included as a trap for posterity.

After validation, a script is executed that starts up the requestor's workstation local print-spooler service (step 4, figure 4). A long string called a secret code is also passed through this script. The inclusion of a secret code containing at least thirty-two random alphanumeric characters will help minimize the probability that a user will notice and remember the URL and data being passed to initiate the activation of the local print spooler. While this is not necessary, it is another security step. The user is then presented with the LIUP and must accept or decline the policy (figure 4, step 5). If the user declines the policy (figure 4, step 5a), the action is recorded to a database (figure 4, step 6) and the Web-browser window is closed using JavaScript. This forces the next user to start a new logon session, thereby ensuring the logon script will execute turning off the local print spooler.

If the user accepts the LIUP, the action is recorded to a database (figure 4, step 6) and the user is presented with the library main home page (figure 4, step 7). The user now has an activated local print spooler and can access any allowable printer resources.

Future

The methods described here are presented as a low-cost solution to segregating printing privileges in a Microsoft Windows client/server environment. Other methods can be

employed that would provide similar results for other systems. The ASUW library used the described methods as a cost-containing intermediate solution until a long-term solution could be implemented. Currently the ASUW library is in the process of implementing the use of the ASUW campus print-management system, which utilizes the Pharos commercial product UniPRINT. With the UniPRINT system, the library will be able to enable printing for all users while applying special rules on print services for each patron type.



LITA has the following Institutes available for licensing.

If you are interested in bringing one of these Institutes to your area, please contact Rob Carlson at 800.545.2433 x4268 or rcarlson@ala.org.

***NEW!* Establishing an Institutional Repository**

Susan Gibbons, Presenter

This institute covers the technical, policy, and human factors related to establishing an institutional repository (IR). Based in part on the presenter's Library Technology Report "Establishing an Institutional Repository" (July/August 2004), the institute will reflect the current state of IR development and best practices, as well as the lessons learned by early adopters. Susan Gibbons is the Assistant Dean of Public Services and Collection Development at the River Campus Libraries, University of Rochester.

Wireless Networks in Libraries

Marshall Breeding, Presenter

Wireless LANS have captivated much attention in the business environment, in the home, and in libraries. A fundamental challenge lies in determining how—or if—to make use of this technology. This workshop aims to provide attendees with the knowledge to make these decisions and to enable them to establish effective and secure wireless networks in their libraries.

Open Source Software in Libraries

Eric Lease Morgan, Presenter

Participants in this hands-on institute will learn skills enabling them to download and install GNU software and Perl modules, create and download software from a CVS repository for sharing code, evaluate open source software for usefulness and applicability, create and foster a community of developers and users, and conduct usability studies to verify a computer program's functionality. Presenter Eric Lease Morgan is the Head of the Digital Access and Information Architecture Department at the University Libraries of Notre Dame.

Also available:

XML and Libraries, *Ron Gilmour, Presenter*

eBooks: After Hype and Death, What Really Works?, *Mark Beatty, Presenter*

Proxy Web Servers & Authentication, *Peter Murray, Presenter*

Call for more information!

Vox Proxy Version 2

Right Seat Software, Inc.,
1110 12th Street, Unit A
Golden, CO 80401
Tel: (303) 278-2244
Fax: (303) 278-6967
Email: info@voxproxy.com
www.voxproxy.com

Price: Vox Proxy Version 2 $199, Vox Proxy Version 2 with CD Prep $228.95. Educational pricing available. Upgrades from Version 1 to Version 2: $99; with CD Prep $109. AT&T Natural Voices text-to-speech (TTS) speech engine $60. Additional voices $25 each.

An evaluation version is also available as a free download at www.voxproxy.com.

System Requirements: Pentium II 300MHZ or faster, 64MB RAM (128MB recommended), 150MB hard drive (500MB more for AT&T engine), 800x600 resolution or higher, a microphone if using VP speech recognition, Windows 95/98/Me/NT/2000/XP, PowerPoint 2000, or newer.

Functionality and Technologies

Anyone who is looking to create a more dynamic PowerPoint presentation may find Vox Proxy meeting his or her needs. The Vox Proxy version 2 is an add-on scripting software package for Microsoft PowerPoint that allows users to have 3-D animated characters that can talk in their presentations. Based on Microsoft's ActiveX and Agent technology, Vox Proxy runs on PowerPoint 2000, 2002, and 2003 under a Windows environment, but not on Mac and Unix/Linux. The Vox Proxy with CD Prep package consists of Vox Proxy, the Vox Proxy Player, and the CD Prep. The Vox Proxy Player is a free program that plays the Vox Proxy PowerPoint presentations for users who do not have Vox Proxy itself. It is similar in concept to the Adobe Reader for those who do not have Adobe Acrobat, while CD Prep is a utility program that prepares self-starting CDs containing the Vox Proxy Player and presentations. The self-starting CDs can even be played on PowerPoint 97. Vox Proxy comes with TruVoice text-to-speech (TTS) engines in ten languages. The extra-cost AT&T Natural Voices TTS engine is available with a male or a female voice in different languages and accents including United States English, British English, French, Spanish, and German.

Microsoft Agent technology is a set of software services that supports various agents (characters) with interactive personalities under a Windows environment. Vox Proxy contains twenty-five additional characters along with the four original included with Windows. Each character comes with a list of animations that it can perform. Paul, for example, can perform more than 100 different animations such as jump, read, and move. Characters can speak text using a TTS engine that users write for them. The default TTS engine is TruVoice, which is supplied by Microsoft, but the AT&T engine delivers more human-sounding computer speech than the default one. Special tags are also available to customize the intonation of the speech, including emphasis, whisper, monotone, pitch, and speed. Special features are also provided to improve the quality of the speech. For example, a pronunciation dictionary is provided to allow users to control the pronunciations for unusual or foreign words, and speech balloons display the speech on the screen at the same time it is spoken.

To create the Vox Proxy PowerPoint presentations, one must use the Vox Proxy script writer. The Vox Proxy script language is different from scripting languages such as VBScript (Visual Basic) or Perl, since it is more natural-language oriented and thus much simpler for the general user. A Vox-Proxy script is simply a text document containing a list of predefined commands, which control the behaviors of characters, manage PowerPoint control commands such as change slides, and perform other actions such as play multimedia files.

Interface

The installation follows the general Windows applications installation process and by default will create the directory Vox Proxy under the Program Files directory. The product is embedded into Microsoft PowerPoint and is displayed as a pull-down menu named Vox Proxy next to Slide Show (figure 1). There are two ways to write the scripts within the scriptwriter, which provides a standard Windows interface for users to test and diagnose scripts: either directly entering the text representing the script commands and the speech; or using its built-in wizard to select the commands and only enter the speech. Usually it is much easier to use the wizard to create the scripts, as the easy-to-use wizard provides a graphical user interface (GUI) for users to select commands or enter the speech in order to reduce the learning curve. A sample script, for instance, can be as follows: (a) first show the character named Peedy at the center of the screen; (b) then Peedy will say "I am very fond of crackers"; and (c) move to next slide (figure 1).

Strengths

The rating of the package is based on software-selection criteria discussed in Mandelbaum and Walton and Taylor, including hardware and software requirements, software quality, documentation and training, and vendor relations.[1] The average

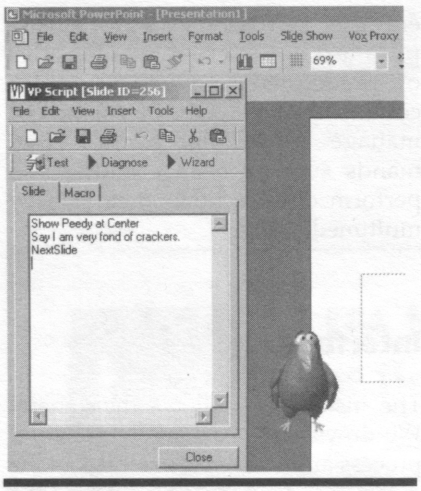


Figure 1. Screen Capture of the Vox Proxy

hardware requirements for the product such as CPU and memory (Pentium II machines) are acceptable, but users who want to use the AT&T Natural Voices engine will need better performing machines (as in Pentium III). The product can only run on Windows, since it uses Microsoft technology, and functions as a plug-in for PowerPoint. However, this is not an issue, judging from the dominant market share of PowerPoint in the presentation software industry. It provides multiple-language TTS engines to facilitate different language presentations, but it is not clear whether corresponding multiple-language GUIs are provided.

The software quality is fairly good in terms of capabilities, stability, ease-of-use, and online help. The product seamlessly integrates with the Microsoft technology and the AT&T Natural Voices engine, and also allows users to customize their presentations in desired ways such as adjusting tones, playing multimedia files, and recording their own speeches. The package contains multiple, ready-to-use characters that can perform adequate animations, although the characters are a little robotic in terms of their appearances and voices. It is recommended to use the AT&T engine—which delivers speech that sounds more human in quality—for serious presentations. All levels of users should feel comfortable writing the script commands without going through significant training because the script syntax is more natural-language oriented. Moreover, the scriptwriter displays the commands and contains multiple, built-in tools such as test and wizard (figure 1). The built-in wizard can be extremely helpful for novice users, while the test tool allows users to test whole scripts as well as a single script command.

The documentation is comprehensive and excellently written, providing adequate information for both novice and skilled users. Various documents are provided along with the CD and also can be viewed on the Web, including FAQs, Bug Alerts, and links to related Web sites. The training tutorials are clear and concise enough for general users to have the software up and running within an hour.

Various short reviews and comments from known and anonymous users are found from secondary sources such as newspapers and a variety of Web sites. It seems promising that the Vox Proxy enables users to present their topics in fun and dynamic ways, since these comments are almost all positive for the product's functionality and the company's support.

Weaknesses

A few improvements are recommended for the Vox Proxy, including providing new features and fixing bugs. The online help contains detailed information for the package, but lacks searching capability for users to quickly locate the desired topic. The package generally follows design guidelines by *Fundamentals of Designing User Interaction* such as installation and interface, but it does miss a few.[2] For example, it does not allow users to choose installation location.

In addition, the scriptwriter and characters tend to always stay on top of application windows, even though you switch to other applications. This could be annoying for users in multitasking jobs. It is recommended that an option for Always on Top be implemented to solve this problem.

Users should note that there is an ambiguity associated with the built-in wizard within the scriptwriter, although the wizard can be extremely helpful for novice users to create their desired scripts. It is possible that a character may perform something that another character is supposed to do, since the script commands created by the wizard sometimes do not have a subject—the owner of the commands. For example, a character may say something that the other character is supposed to say when using the wizard to create a "say" command. Therefore, it is recommended that users check every command to ensure the presence of a subject. The vendor stated that a future release would display the subject to avoid this ambiguity.

Conclusion

To summarize, the Vox Proxy may be a little costly for general users, but educational licenses are available for school or university libraries. It is easy to use and effective to deliver dynamic presentations in multiple languages. The software has many add-in features for PowerPoint that may be beneficial for library staff who frequently deliver presentations.—*Yan Han, University of Arizona Library*

References

1. Jane B. Mandelbaum, *Small Project Automation for Libraries and Information Centers* (Westport: Meckler, 1992); Robert A. Walton and Nancy Taylor, *Directory of Microcomputer Software for Libraries* (Phoenix: Oryx, 1986).
2. Microsoft Corporation, 2002, Fundamentals of Designing User Interaction (Microsoft Corporation, 2002). Accessed Nov. 5, 2003, http://msdn.microsoft.com/library/default.asp?url=/library/en-us/dnwue/html/part1.asp.

eLibrarian

Digi-Net Technologies, Inc.
1034 NW 57th Street
Gainesville, FL 32605
(877) 404-2428
info@digi-net.com

Price: Two price packages available. If licensed, 1 seat@ $4,500 one-time fee, then $500 annually. If hosted, 1 seat @ $250 monthly (paid monthly), and $200 one-time set-up fee.

System requirements: Hardware: 550 MHz Pentium CPU (or equivalent) or faster, 256 MB RAM, SVGA Video Card/Resolution. Co-Browsing: Dual 866 MHz Pentium CPU (or equivalent) or faster—or—1.33 GHz Pentium CPU (or equivalent) or faster, 512 MB RAM. Requires Windows 2000, Windows 2000 Server, Windows NT 4 With SP 4IIS 4 or 5 SSL Certificate, 2 gigs of free space, fully configured DNS. Operating systems: Windows 95, 98, Me, NT 4.0 SP6, 2000, MacOS 8.6 or higher (the Groopz Server will not run on any MacOS prior to MacOS X/Server), Linux 2.0 or higher, Solaris 7 or higher. Browsers: Microsoft Internet Explorer v4.0 or higher, Netscape Communicator v4.0 or higher, America Online (AOL) v4.0 or higher.

By now, most—if not all—information professionals working in the library industry have heard of chat-reference services. This is a service that uses software in order to either bypass or speed up the transactions that e-mail reference services provide. With many software vendors emerging, including a fee-based service from Yahoo, libraries have a wide array of options and bells and whistles to choose from. One particular chat software that has emerged is eLibrarian, a product that is marketed and sold by Digi-Net Technologies. Current users of this software include OhioLink (academic consortium of libraries in Ohio), University of Notre Dame, Skagit Valley College, and Health Sciences Library Consortium in Pennsylvania.

eLibrarian is available in two price packages. One package includes hosting by Digi-Net; the institution runs the operator console at its site. Alternatively, an institution may choose to set up and run the server locally. This review was done with servers hosted by Digi-Net.

Installation for this software is relatively simple. The console installation comprises a self-executable file of about 12MB in size and is available for a variety of operating systems. After the installation, accounts for operators will have to be set up. There is no limit to the number of operator accounts you can create or installations you can perform on your institution's computers. There is, however, a limit to the number of operators who can log on simultaneously. For those institutions that might experience a high number of reference questions, having two or more concurrent operator licenses might be a good choice.

After performing the installation and establishing operator accounts, the system administrator will have to embed a JavaScript into the HTML page that will feature this service. The JavaScript displays an image that indicates whether or not the librarian is currently available. Customization for images and colors prior to the service going live can be made by contacting representatives of Digi-Net.

Like the many other chat-software vendors, eLibrarian offers the ability to push Web pages, push automatic or canned messages, and co-browse. The versatility to handle chat requests originating from a variety of operator platforms was a strength found in eLibrarian. Other strong features included the ability to customize features on the operator's console. Everything from the changing of skins, sounds, notifications, and establishing filters was available for each console installation. If you acquire more than one license, it may be a good idea to limit the authority or access rights that allow for unique customization.

Another strong feature of this software is the ability to query the individual requesting a chat session by asking questions or selecting answers before a chat session has been initiated. Again, this is a feature that is totally customizable and could be tailored to each institutional need. Another unique feature is the ability to smart-route chats depending on where the chat box is located on an institution's page. For example, if a chat box was located on a branch library's page, you can send calls originating from this location to specific operators. Another great feature is the ability to see when a user is on the page your chat box located on. The console will display the type of browser, operating system, time-on-page, and more.

Other features include the ability to route users to exit surveys which are basically HTML forms the institution creates to acquire information on the quality of the chat service. The ability to retrieve chat transcripts from either the main operator or other operators can also be a useful function if you choose to look up questions that needed follow-up. When it gets busy at the reference desk, an operator can also transfer an entire chat session to another operator who may be logged on as a backup. eLibrarian provides the ability to e-mail chat transcripts to users if e-mails are prompted for before a chat session begins. Finally eLibrarian provides an enhanced

Web access for gathering detailed statistics. Statistics will show, among other things, number and length of chat sessions, and page referral.

Some downfalls of this software included notification, co-browsing, and the copying and pasting of text. Notification can be a major concern for libraries. With only a bell ring, or other small sound, a chat request might be missed, especially when the operator might have other windows open or might have the speakers turned off at the reference desk. A Digi-Net software representative said that plans for making chat requests more noticeable were in the works. Co-browsing runs with the same Hipbone technology that is used in other live customer service products. It was found that the time it took to initiate a co-browsing session was noticeable and in some cases froze the browser. With the current version of this software you cannot do the simple right-click, copy, and paste of text. Instead you have to rely on the key presses Ctrl+C and Ctrl+V to do these functions. This can be a hindrance when you are trying to copy text or URL addresses to the end user in a chat session.

Having a product that works on multiple platforms, is customizable, and has other unique features are probably the highlights of this software. Customer support seemed very eager to sell the product, but fell short on communication afterward. When trying to contact tech support or find answers to questions dealing with technical programs, it was often necessary to ask time and time again only to be told that the problem was being looked at. Overall, eLibrarian's strengths did not outweigh the weaknesses for this particular software. According to Russell Garabelis, Digi-Net's Media Relations manager, eLibrarian's software development team is currently working on incorporating an artificial intelligence engine with a knowledge base, voice-over IP, email integration, enhanced co-browsing and enhanced reporting.—*Hector Escobar Jr., University of Notre Dame Libraries*

Correction

During our editorial revision of Laura Cohen's article in the June 2004 issue of *ITAL*, "Issues in URL Management for Digital Collections," on pages 42–49, we inadvertently made changes that introduced the following errors into the text:

1. The text published on page 45,"An example is the starting point URL to set a session variable and retrieve the file for the FirstSearch database WorldCat [URL is shown]," should read as follows:

   An example is the starting point URL for the FirstSearch database WorldCat, http://firstsearch.oclc.org/dbname=WorldCat;done=referer;FSIP, which in one particular instance becomes http://firstsearch.oclc.org/html/webscript.html:%3Asessionid=sp04sw01-48488-ddrpn8iv-6944q4:sessionid=sp04sw01-48488-ddrpn8iv-6944q4 to set a session variable and retrieve the file.

2. In the list of references on page 49, the trailing slash was removed from the URLs in references 9 and 12. In an article that makes the case for inclusion of trailing slashes, this omission was inappropriate.

ITAL regrets these errors.

Index to Advertisers

Art & Tech	107	LAMA	94
EBSCO	127	LITA	cover 2, 118, 132, cover 3
InfoUSA	cover 4	Techbooks	101

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

When your Patrons have Questions,

We have the Answers.

*Reference*USA is the ultimate research tool containing the freshest data available!

Search millions of people, businesses and just about anything you desire.



Our databases contain:

- 14 Million Businesses
- 115 Million U.S. Households
- Canadian Businesses/Residents
- Health Care Module
- And so much more!

Our database products include:

- **Internet Access-ReferenceUSA** - 14 million businesses, 115 million U.S. households, Healthcare Module, & Canadian businesses and residents
- **American Manufacturers Directory and CD-ROM** - Contain every manufacturer in the U.S. with 20 or more employees—more than 168,000 listings!
- **Big Business Directory and CD-ROM** - Include America's largest companies, 193,000 companies that employ 100 or more people. Also includes 703,000 top executives and directors.
- **State Business Directories** - Contain virtually every business within each state. Available in all 50 states plus Washington D.C.
- **American Business Disc** - 14 million businesses on one CD-ROM

"In a library such as ours... *Reference*USA is among the most in-demand databases because of its scope and accuracy."

John Ganly
Assistant Director of the
New York Public Library

5711 S. 86th Circle
P.O. Box 27347
Omaha, NE 68127

ReferenceUSA℠
A division of *info*USA

For More Information, Call Toll-Free: (866) 813-3009
or E-mail: reference@infoUSA.com
Check out our website at: www.libraryUSA.com
Phone: (866) 813-3009 • Fax: (402) 596-7688

Friends of Libraries U.S.A.

23-ITL