Question Master: An Evaluation of a Web-Based Decision-Support System for Use in Reference Environments

John V. Richardson Jr.

Designed for librarians, Question Master (QM) (at http://purl.org/net/Question_Master) is a decision-support system automating some of the more routine, fact-type reference questions encountered in libraries. A series of Web pages guides librarians through a set of clarifying questions before making recommendations of an appropriate electronic or relevant print resource from WorldCat, the OCLC Online Union Catalog. The goal is to improve the accuracy of reference transactions, which in turn should lead to increased end-user satisfaction. Based on usability studies of QM's biographical module, this study found that although the system already was easy to use, its usability could be improved in several ways. Its ability to answer questions was 100 percent, with an accuracy rate of 66 percent compared to Weil's 64 percent accuracy. In addition, QM accuracy was substantially better than most reported studies of real reference environments and certainly better than the Internet results of 20 percent for HotBot and 30 percent for AltaVista.

Writing in 1964, Jesse Shera said: "The popular conception of automation as applied to the work of the reference librarian suggests a mechanical marvel from which accurate and authoritative answers to questions will be disgorged in immediate response to a push of the proper button."[1] In the future, intelligent technology could answer many types of questions from anywhere in the world. For example, both reference librarians in any type of library and end users from school, work, or home could query an intelligent (a.k.a. expert or knowledge-based) front end to their local OPAC to answer reference queries. Such queries could utilize the profession's database of shared cataloging records to answer reference queries: The practical benefits are obvious.

More specifically, the OCLC Online Computer Library Center in Dublin, Ohio, provides access to 36 million OCLC/MARC-formatted bibliographic records, including reference works via WorldCat (the OCLC Online Union Catalog). The reference records are accessible by the 049 field and local holdings by the LCC (050 field) or DDC (082 field) classification number, delimited by the double dagger symbol (‡). Furthermore, the 6xx field, subfield x provides access to the standard form subdivision[2] that further identifies types of reference sources (e.g.,

John V. Richardson Jr. is an Associate Professor in the Department of Library and Information Science of the Graduate School of Education and Information Studies at UCLA; e-mail: jrichard@ucla. edu.

biographical sources, dictionaries, encyclopedias, and indexes).

Considering that more than fifty extant first-generation expert systems for reference service have been reported in the library and information science literature, proof of concept clearly exists.[3] Not only does proof of concept exist, but the author also has articulated the architectural logic of doing reference work; and there are more than 1,500 additional rules for building a second-generation knowledge-based system.[4]

The Research Problem

Stated formally, the problem is that more than 250 million reference questions are asked in U.S. public libraries every year, according to the National Center for Educational Statistics.[5] Many more questions in business and at home go unanswered every year due to information technology barriers that need not exist. Moreover, research suggests that the accuracy of the librarian's response is only about 50 percent: One out of every two questions is answered correctly.[6] Therefore, an intelligent decision-support system (IDSS) could serve librarians well. An IDSS could free up the valuable time of reference librarians so that it could be spent answering the more demanding research-type questions. In the long run, an IDSS would reduce a library's costs of providing access to information. Such a system could be available full-time from any Internet-accessible computer and could recommend the single, best source regardless of language, as well as record the complete transaction.[7]

Research Goals and Objectives

One way to bring the present reality and the future closer together would be to implement a second-generation, question-answering prototype using the World Wide Web. Thus, the threefold, overarching goals of this research project are to support the decision-making process of librarians by automating some of the more routine, fact-type biographical reference questions, to improve the accuracy of reference transactions, and to increase end-user satisfaction. The project's three specific objectives are:

1. to implement a Web-based system that will select the single most appropriate resource, either print based or electronic, regardless of language, in order to answer the end user's query;
2. to evaluate its usability;
3. to test its accuracy (and then compare and contrast its results with earlier studies of human reference librarians and computer-based systems).

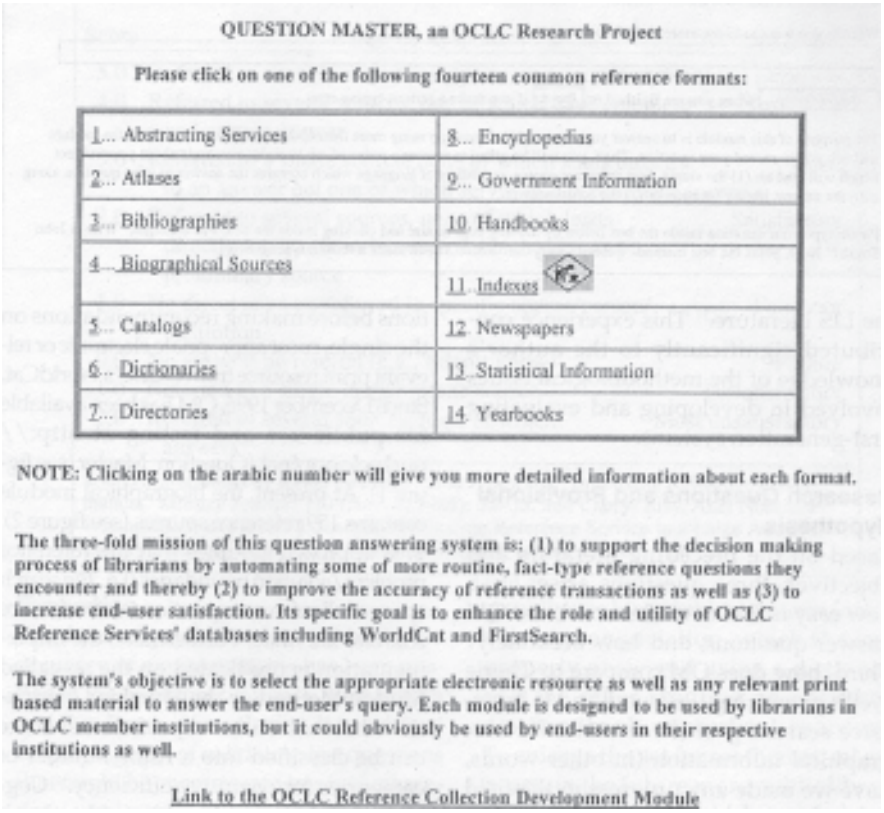
Related Research

Based on an extensive review of the professional literature, the author discovered at least three prior efforts to develop a biographical question-answering system.[8] These are reviewed in chronological order.

In 1967, Cherie B. Weil, a student in the Graduate Library School at the University of Chicago undertook her master's thesis, entitled "Classification and Automatic Retrieval of Biographical Reference Books," under the direction of Professor Victor Yngve.[9] Writing her program in COMIT, a list-processing language, she designed a mainframe batch program at an estimated cost of $900 to make use of 234 biographical reference sources which she characterized on the basis of eight points: living/dead, nationality, gender, occupation, religion, race, memberships, and date. Arguing that "there are not enough reference librarians who have perfect recall of their collections," she tested her system with fourteen test questions which she randomly drew from an advanced reference syllabus and discovered that it could answer eight (66.6%) of those questions accurately.[10] This figure became the unofficial goal to beat.

The next reported system using biographical sources is the Biographical Reference Advisor developed in 1987 at the

FIGURE 1
QM Opening Page



Decker Center for Information Technology at Goucher College (Maryland) by Robert Lewand, professor of mathematics and computer science, and Larry Bielawski, director of the center. They relied on Yvonne Lev and Barbara Simons as their domain experts. Using an IBM personal computer (PC) for their platform, they selected shell software, called First Class, to implement a menu-driven program of 680 nodes on a total of eighty-five decision trees. Biographical sources were characterized by their coverage of contemporary versus historical figures; fourteen different occupations; twelve nationalities; and gender. Development required five months and cost approximately $2,500. Informal evaluation reports indicate that "students rarely use the system," which emphasizes the importance of usability testing.

During the fall quarter 1987, the author introduced an expert system assignment in sections of the required reference services course sequence at UCLA's Graduate School of Library and Information Science. Over the course of the next several years, the author's graduate students developed a variety of modules, including several biographical ones using ESIE, a shareware backward-chaining shell. In addition to positive class evaluations and a multistudent presentation at the 1988 ASIS midyear meeting, several students reported on their experience in the LIS lit-

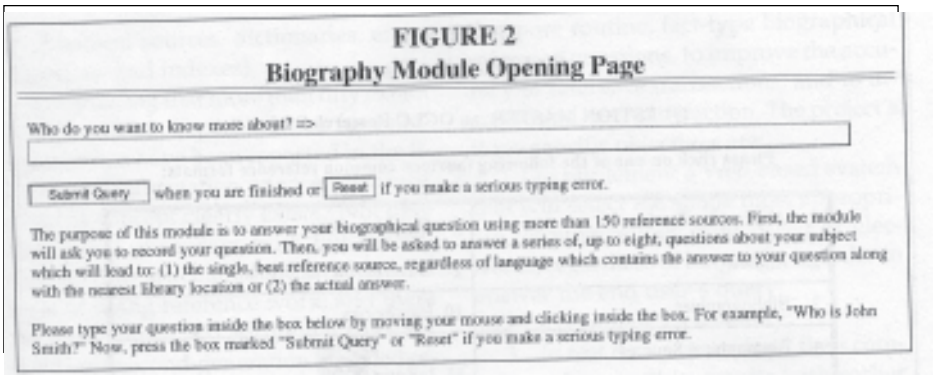


FIGURE 2
Biography Module Opening Page

Who do you want to know more about? =>

Submit Query when you are finished or Reset if you make a serious typing error.

The purpose of this module is to answer your biographical question using more than 150 reference sources. First, the module will ask you to record your question. Then, you will be asked to answer a series of, up to eight, questions about your subject which will lead to: (1) the single, best reference source, regardless of language which contains the answer to your question along with the nearest library location or (2) the actual answer.

Please type your question inside the box below by moving your mouse and clicking inside the box. For example, "Who is John Smith?" Now, press the box marked "Submit Query" or "Reset" if you make a serious typing error.

erature.[11] This experience contributed significantly to the author's knowledge of the methodological issues involved in developing and evaluating first-generation systems.

Research Questions and Provisional Hypothesis

Based on the preceding literature and objectives, three questions arise: First, how easy is QM to use? Second, can QM answer questions, and how accurately? Third, how does QM compare to Cherie Weil's pioneering 1967 work and to brute-force searching of the Internet for biographical information (in other words, have we made any progress in the field of intelligent question answering?). Finally, the author proposes the null hypothesis that there is no statistically significant difference in the accuracy between QM, Weil's system, and what one can find on the Internet. Answers to these questions will tell the profession about the promise of intelligent question-answering systems and give system designers insight into promising approaches.

Methodology

This section covers the construction of QM, the design of its usability testing, and the scoring of its accuracy.

HTML Pages

QM is a series of HTML pages that guide librarians through a set of clarifying questions before making recommendations on the single, most appropriate electronic or relevant print resource from OCLC's WorldCat. Since December 1996, QM has been available for public use and testing at http://purl.oclc.org/net/Question_Master (see figure 1). At present, the biographical module contains 159 reference sources (see figure 2). This approach assumes that the reference problem/solution boundaries (i.e., the search space of all reference questions and reference sources) are finite. Furthermore, the implementation is predicated on the so-called Mudge Method or the Hutchins Heuristic—that each and every reference source can be classified into a finite number of categories for cognitive efficiency.[12] Cognitive effectiveness is achieved by classifying reference questions by format and then by specific source. To determine their utility in the reference environment, the author used a modified distinctive feature analysis to categorize the sources. Hence, the interface poses questions much like the reference librarian should be doing to reach the correct conclusion. As alluded to above, actual implementation of the author's intelligent question-answering system is based on decision rules using a multiple-choice classification process. Without a doubt, the theory is a reductive transformation of the reference librarian's complex, decision-making task; nonetheless, the advantage is that it converts this complex task into a much more manageable one for a computer-mediated environment.

TABLE 1
Taxonomy of Potential Responses

Score	Range of Responses	Service Quality
5.0	Referred to a single-source, complete, and correct answer	Excellent
4.0	Referred to several sources, one of which gave complete and correct answer	Very good
3.0	Referred to a single source, none of which leads directly to an answer but one of which serves as a preliminary source	Good
2.0	Referred to several sources, none of which leads directly to an answer but one of which serves as a preliminary source	Satisfactory
1.0	No direct answer; referred to specific source/person/institution	Fair/poor
0.0	No answer; no referral (e.g., I don't know)	Failure
-1.0	Referred to a single inappropriate source	Unsatisfactory
-2.0	Referred to several sources, none of which answers	Most unsatisfactory

Source: Suggested by Ralph Gers and Lillie J. Seward, "Improving Reference Performance," *Library Journal* 110 (Nov. 1, 1985): 32–35; and Cheryl Elzy, Alan Nourie, F. W. Lancaster, and Kurt M. Joseph, "Evaluating Reference Service in a Large Academic Library," *College & Research Libraries* 52 (Sept. 1991): 454–65.

Evaluation

The detailed procedure for evaluating the accuracy of expert systems has been posited in the literature by John V. Richardson and Rex Reyes.[13] Simply put, it consists of employing a set of validating test questions that might be encountered in a typical academic or large public library and scoring the answers on an eight-point scale (see table 1). In essence, this scheme is based on rewarding efficiency following Cutter's maxim that one does not want to waste the user's time.

For this study, Weil's original 1967 questions were selected initially (see her appendix E, "Results of Accuracy Test") so that a comparison and contrast could be made with her study. Note that her questions are strong on deceased, foreign individuals much as one would encounter in an academic library setting. Unfortunately, four of her original questions (numbers 4, 9, 10, and 11) had to be eliminated because the first one was more strictly genealogical in nature and the next one was no longer a valid contemporary question. The final two did not actually use biographical sources (one involved pronunciation, which she more properly answered using a dictionary, and the other required an encyclopedia, handbook, or yearbook to answer). To enlarge the set for mathematical purposes, in the future more public library-type questions should be added from the biographical module of the OCLC Reference Collection Development Module, which logs users' questions. In this case, the questions could involve more living Americans. The author scanned the log, which represents more than two hundred users, looking for typical questions. Finally, starting in mid-February 1997, the ten questions were used for two brute-force searches of the Web in order to see how well an unmediated search might perform in finding answers. The two searches were conducted using the largest index (31 million pages) created by Digital Research Laboratory's AltaVista Scooter® as well as Inktomi Corporation's HotBot Slurp®, which searches at a deeper level than AltaVista.

Usability

In late March 1997, the OCLC Usability Laboratory (Ulab)[14] recruited four test users to evaluate the quality of QM.[15] Selected by the Ulab, the four users (three women and one man) were typical of the LIS community: white, middle-class individuals with corporate, public and academic library experience. Their positions ranged from library clerk to former head of a large academic reference department, and each had worked with reference questions on a daily basis. These individuals were asked to use QM to find the answers to the set of test questions mentioned above. Each task was to be accomplished in two minutes, which is the average time spent on ready-reference-type questions. While being videotaped for subsequent analysis,[16] users were asked to "'think aloud,' verbalizing what they are thinking and problems they encounter while doing the tasks."[17] After the test, each user completed a questionnaire and was interviewed by the principal investigator and a member of OCLC's Human-Computer Interaction Team.

TABLE 2
Scoring of QM and Weil's Reference Book System

QM No.	Weil No.	QM	Weil	Total Possible
1	1	5.0	2.0	5
2	2	5.0	4.0	5
3	3	3.0	4.0	5
4	5	5.0	2.0	5
5	6	-2.0	2.0	5
6	7	5.0	4.0	5
7	8	3.0	4.0	5
8	9	3.0	4.0	5
9	13	3.0	4.0	5
10	14	3.0	2.0	5
Grand Total		33 (66%)	32 (64%)	50 (100%)
Mean Score		3.3	3.2	

Research Findings

Usability

Based on the Ulab testing, QM scored an average of 4.5 on a five-point scale where 1 was "very difficult" and 5 was "very easy to use."[18] In addition, several unique items were found that could be used to improve QM. In particular, its usability was increased in the following six ways. First, one page was added to define the function of each format; second, the query box was moved to the top of the biographical module; third, one page was rewritten to clarify that the system, at this time, recommends the single, best source regardless of language[19]; fourth, another page was reformatted to indicate more clearly that "brief versus long" refers to the biographical entry in the source rather than the bibliographical record; fifth, several pages (the ethnicity, religion, and occupational pages, specifically) were merged into a single screen following the selection of either living or deceased individuals; and sixth, additional space was added on all pages to make it clear that "unsure" is an option everywhere. Overall, users understood the screens and the terminology, and knew what was going to happen next.

As mentioned earlier, these changes probably account for at least 75 percent of the difficulties any user might have with the system. And even prior to these changes, all the test users thought the system was "easy to use." Now, almost any reference librarian or staff member should be able to use QM with great ease.

Evaluation of Accuracy

Based on the results presented in table 2, QM is able to answer 100 percent of all biographical questions put to it. In several instances, it could provide a single source with the complete and correct answer. However, an equal number of times,

TABLE 3
Question Master Versus Internet Search Engines

QM No.	Internal No.	QM	Alta Vista	HotBot
1	Weil, 1	DSB	1*; 3rd – 4th options/1000	0; 1/6
2	Weil, 2	CWW; SDCB; MDCB	0/155,264	0/29,888
3	Weil, 3	BGMI	0/38,882	0
4	Weil, 5	BLKO	0	0
5	Weil, 6	AO or BDUB	0/70,919	0
6	Weil, 7	MEL	1/7522;†	4/50; 1/4
7	Weil, 8	GDMM	1/10,492;‡	?/173
8	Weil, 9	BGMI	0	0/4
9	Weil, 13	WWWAA	0/30,743	3-4/238
10	Weil, 14	BGMI	0/15,558	0/4
Success Rate			30%	20%

In the Alta Vista and HotBot columns, the figure before the slash indicates the location of the most useful hit within the retrieval set and the second number indicates the overall size of the retrieved set. Figures after the semicolon indicate results of an advanced search.
* = a server error; † = not found; and ‡ = page returned did not contain an address.

it failed rather miserably because it would recommend Biography and Genealogy Master Index when there was incomplete information about a subject, and that individual would not be listed in this source. Overall, QM scored thirty-three out of fifty points (66%), or 3.3 per question, on average. In qualitative terms, that would mean its service quality was somewhat better than good. In one sense, Weil's system performed more consistently in a narrow range by recommending more sources each time, but although these might be judged good sources, usually only one would lead to the correct answer. For a rather large amount of time, the user would be looking for the answer in one of those recommended sources, whereas QM would recommend only one title and the user would know immediately upon checking the source that the answer was not there.

Compared to the Internet searches, QM is superior for several reasons (see table 3). First, many of the Internet searches yielded no results at all, or when they did, they returned exceptionally large, hence useless, retrieval sets. Of course, these large sets could be reduced by using quotation marks or other techniques, although naïve users may not know of or use the advanced searching options. Second, many of the pages retrieved are not available now. Of course, persistent uniform resource locators (PURLs) are one solution to this difficulty.[20] However, when a page does return useful information (20 to 30% of the time), it is highly satisfying and supports the principle of least effort—why shouldn't relevant information be at one's fingertips?

For the Future

Despite QM's exceptionally easy-to-use interface, the author would like to implement a form-based approach to asking the user questions about the query. Second, adding more titles could increase QM's ability to answer additional questions that might be encountered in a real reference environment. Third, deducing additional facets of biographical questions might increase QM's accuracy. And finally, someone should evaluate more broadly the Internet's accuracy using the same criteria as discussed here.

Conclusions

Intelligent question answering is making progress. In one sense, we see technological differences, having moved from an overnight batch environment in 1967 to on-demand answers via the Web. QM is available twenty-four hours a day, seven days a week, whereas Weil's system operated in a batch mode. Nonetheless, there were no significant differences in accuracy since 1967, although QM is more usable, more efficient, and less wasteful of the user's time. However, there is a big difference between mediated and unmediated searching of the Internet. Brute-force searching of the Internet is still not viable, at least for this set of test biographical questions. Intelligent, mediated searching by human or computer is still necessary; however, the ability to reduce human error in the reference transaction seems especially noteworthy. Although QM is "easy to use" and slightly more accurate than what could be done in 1967, there is obvious room for improvement, as mentioned above. Perhaps the reference theory, the so-called Mudge Method or Hutchins' Heuristic, is unsatisfactory, and a better theory of how to answer reference questions is needed. In summary, this is a situation not unlike the now familiar SDC Orbit or Lockheed Dialog online searching systems in the early 1960s—research prototyping. Perhaps in another thirty years, we will have what Jesse Shera stated was the popular conception of automation as applied to reference work. The ideal of accurate and authoritative information from a single computer interface is not that far away.

The author would like to thank UCLA for his year-long sabbatical during the 1996–1997 academic year, as well as OCLC for offering him the position of Visiting Distinguished Scholar. In the course of this research project, the author enjoyed the support and encouragement of many individuals at OCLC, including: Terry R. Noreault, Director, Research and Special Projects; Keith E. Shafer, Senior Research Scientist; Bradley C. Watson, Consulting Systems Analyst; Patrick McClain, Systems Analyst; Vincent Tkac, Senior Programmer/Analyst; Mike Prasse, Head, and Chris Vavro, Analyst, of the OCLC Usability Laboratory, as well as the four test subjects; Susanne Krouse, Administrative Coordinator; and last, but not least, the help desk folks, including Kevin Ball and Bruce Goll. Original funding for the knowledge base in QM was provided by UCLA's Academic Senate Committee on Research and the Council on Library Resources, Grant Number 8027.

Notes

1. Jesse Shera, "Automation and the Reference Librarian," *RQ* 3 (July 1964): 4.
2. Library of Congress, *Subject Cataloging Manual: Subject Headings*, vol. 1 (Washington, D.C.: Cataloging Distribution Service, 1991), section H180, II, 3.
3. John V. Richardson Jr., "A Review of KBS Applications in General Reference Work," in *Knowledge-based Systems for General Reference Work: Applications, Problems, and Progress* (San Diego, Calif.: Academic Pr., 1995), chapter 8.
4. ———, "The Development of a Knowledge Base for an Expert System in Reference Work," in *Knowledge-based Systems for General Reference Work: Applications, Problems, and Progress* (San Diego, Calif.: Academic Pr., 1995), chapter 5.
5. National Center for Education Statistics, *Public Libraries in the United States, 1997* (Washington, DC: GPO, February 1997).
6. Matthew Saxton, "Reference Service Evaluation and Meta-Analysis: Methodological Issues in Summarizing Data from Multiple Studies," *Library Quarterly* 67 (July 1997): 267–89.
7. A more complete list of the actual requirements of an essential system are given in John V. Richardson Jr., "Understanding the Reference Transaction: A System Analysis Perspective," in progress.
8. See table 8.1, entries 1 and 24, as well as page 272, in Richardson, *Knowledge-based Systems for General Reference Work.*

9. Cherie B. Weil, "Automatic Retrieval of Biographical Reference Books," *Journal of Library Automation* 1 (Dec. 1968): 239–49.

10. On page 106 of Weil's study, she indicates that scoring the computer's response was "A = It has the answer or at least part of it; B = Good choice but does not have the answer." She scored eight questions as A and two as B. In her final scoring, she reported 66 percent accuracy whereas the author scored her system as 64 percent accurate. The author interprets these results as having no significant difference after having dropped four of her original fourteen questions.

11. Deborah Henderson, Patti Martin, Lauren Mayer, and Pamela Monaster, "Rules and Tools in Library Schools," *Journal of Education for Library and Information Sciences* 30 (winter 1989): 226–27.

12. John V. Richardson Jr., "Teaching General Reference Work: The Complete Paradigm, 1890–1990," *Library Quarterly* 62 (Jan. 1992): 55–89.

13. John V. Richardson Jr. and Rex Reyes, "Expert Systems for Government Information: A Quantitative Evaluation," *College & Research Libraries* 56 (May 1995): 235–47. In the near future, the author expects that Matthew Saxton will present a methodology (e.g., a seven-point Likert scale modifying Childers's work) in his dissertation tentatively titled: "Evaluation of Reference Service in Public Libraries Using Structural Equation Modeling: The Role of Multivariate Analysis in Testing Theory."

14. Michael J. Prasse and R. Tigner, "The OCLC Usability Lab: Description and Methodology," in *13th National Online Meeting Proceedings—1992 Meeting, New York, May 5–7, 1992*, ed. Martha E. Williams (Medford, N.J.: Learned Information, Inc., 1992), 255–61.

15. Using four test subjects will uncover at least 75 percent of a system's problems, according to Robert A. Virzi, "Refining the Test Phase of Usability Evaluation: How Many Subjects Is Enough?" *Human Factors* 34 (August 1992): 457–68.

16. Michael J. Prasse, "The Video Analysis Method: An Integrated Approach to Usability Assessment," in *Proceedings of the Human Factors Society 34th Annual Meeting: Orlando '90* (Santa Monica, Calif.: The Human Factors Society, 1990), 400–404.

17. *The OCLC Usability Lab: Supporting Efficient Development of Excellent Products* (Dublin, Ohio: OCLC Inc., 1996), 1.

18. The videotape of the user sessions along with Dr. Michael Prasse's Usability Evaluation Summary Report are filed in the OCLC Information Center.

19. For a commercial system, this requirement could be modified to recommend more sources or simply those in a single language, such as English.

20. For more on persistent uniform resource locators (PURLs), point the browser to http://purl.oclc.org.