Harvesting Hyperspace: Developing Technological Solutions to Internet Resource Discovery and Description

Gregory A. McClellan and Thomas P. Turner

Abstract

Many students and faculty make use of Internet search engines to locate information on the World Wide Web. The increasing interest in these resources challenges collection development specialists to find more resources, technical services librarians to describe them, and public services librarians to support them. This paper addresses experiments underway in the Technical Services unit of the Albert R. Mann Library at Cornell University to make use of the underlying World Wide Web indexing technologies in a local setting, to develop new ways for catalogers to approach Internet resources, and to discover cost effective methods for approaching the wealth of resources available on the Internet. We have worked to link this new, more detailed information into the Cornell University Library Gateway, a local database of network-accessible resources. The authors discuss technological solutions for indexing large aggregations of Web resources, electronic serials, Web sites containing multiple file formats and Web sites that make use of frames.

Introduction

Many students and faculty make use of Internet search engines to locate information on the World Wide Web. However, these general resources do not serve our scholars and students well because they lack the care and attention that our local bibliographers and catalogers give to choosing and describing individually selected resources. This paper will address experiments underway at the Albert R. Mann Library at Cornell University to make use of the underlying World Wide Web indexing

Gregory A. McClellan is cataloging librarian for networked information resources, and Thomas P. Turner is metadata librarian, Cornell Unviersity.

technologies in a local setting, to develop new ways for catalogers to approach Internet resources, and to discover cost effective methods for approaching the wealth of resources available on the Internet.

Providing access to locally selected Internet resources involves decisions that impact collection development, public services, and technical services units in very different ways. Bibliographers face the difficulty of discovering relevant materials for our users and gathering enough information about resources to make appropriate choices. Public services staff need to identify which parts of a complex Internet site are appropriate for a user's needs. When Internet sites are added to our on-line catalogs, technical services staff are challenged to find ways to describe these resources, which are often large, complex sites, so that the wealth of materials available will be apparent to our users.

In this environment, librarians must develop new methods for discovering, collecting, and describing. We will discuss current efforts at Mann Library to make use of a locally mounted World Wide Web indexer, based on the Harvest gatherer, to explore large, complex sites. The information is used to create a "metadata tank" because automated indexing provides a finer level of granularity than is available through traditional cataloging. Making use of this technology may help the library take a new approach to dealing with Internet resources. Although we will focus on the technical services perspective, we will also address collection development and public services concerns related to the inclusion of Internet resources in local collections. We will propose a model for the automated generation of metadata for Internet resource description.

In January, 1998, the Cornell University Library administration made money available for a variety of projects through an internal grant program. This program enabled Mann Library Technical Services staff to experiment with the local use of Internet indexing technology to grapple with the difficulties that Internet resources present. By seeking technologically enhanced, rather than strictly human, approaches to these resources, we hope to improve access to these materials as well as the speed with which we are able to process them. Without the additional funding of the Cornell University Library administration, this experiment would not have been possible.

Finding and Describing Resources on the Web

Finding the right resource on the Internet can be difficult. Library patrons can turn either to large-scale Internet search engine services or to local on-line catalogs which provide access to locally selected and cataloged materials. Internet search engine companies create databases of Internet citations by making use of indexing technology to find and organize materials, such as AltaVista and HotBot, (Kimmel 1996) or by having employees select, annotate and/or assign subject captions to Web sites, such as Yahoo! (Lester 1995, Steinberg 1996). These services often result in the creation of large databases without a clearly defined audience.

Libraries are struggling to find ways to approach the Internet and its resources for local, specific audiences. Some libraries have chosen to limit the selection and cataloging of Internet resources. Many libraries have also started to select and catalog Internet resources using the MARC record. However, this task has presented several challenges to libraries in terms of establishing selection criteria for a variety of complex documents, determining cataloging processes, maintaining cataloging records over time, and supporting public use of a vast assortment of materials. Many libraries have approached the selection of Internet-based resources in the same manner as other resources. They have developed selection policies and procedures that analyze the quality, consistency and applicability of each site for the library's collection (Demas, McDonald and Lawrence 1995). Internet-based services are presenting new challenges as well. Walters et al. (1998) describe the complex issues surrounding the selection of aggregations of resources available over the Internet. In addition, although Internet-based and other electronic resources are being added to library collections, print materials will continue to require the attention of technical services and collection development staff. This hybrid collection will require a sharper focus on methods for adding value to the entirety of the library collection (Atkinson 1998).

Defining the Control Zone

The addition of Internet-based resources in the library collection has had an enormous impact on the work of technical services units. New methods for electronic resource description have emerged for cataloging units dealing with electronic resources (Younger 1997, Dillon and Jul 1996). New sets of standards have emerged for working effectively with electronic resources. Most of these new standards require catalogers to broaden their sense of resource description beyond the MARC record and to look at other forms of metadata. Clifford Lynch (1998) describes metadata in the following ways:

> Metadata is literally "data about data," information that qualifies other information. Bibliographic description is a form of metadata, so also is information about intellectual property rights and terms of use, formats of electronic information, reviews, errata, abstracts and summaries, provenance information, and a host of other data. Some metadata can be derived mechanically from objects; other metadata has independent standing as intellectual creation in its own right (p.5).

This wide scope of resource description possibilities involves an assortment of new metadata schemes. Options include MARC, the Dublin Core, PICS, among others, and involve related developments like the Resource Description Framework (Gill 1998, Chepesiak 1999). With the rise in the number of metadata and descriptive options, finding a way to manage those possibilities becomes more important. Libraries need to find methods for building on the strengths of all resource description possibilities rather than trying to impose one method on all resources (Vellucci 1997).

In this complex metadata environment, making use of technological solutions to resource discovery and description should be considered as much of an option as making use of human-dependent methods (such as traditional selecting and cataloging of materials). Looking at the scope of the digital library, Atkinson (1996) calls for the establishment of a "control zone" in which libraries determine which resources are essential for scholarly endeavors. He contrasts this collection with the "open zone," or the network at large. One way to develop this control zone would be to perform traditional selection and cataloging on sets of materials that are defined as essential to scholarly pursuit, such as high-priced items or digital materials created by the library. For materials that may be of interest outside this core set, making use of cheaper, less-perfect methods of automated indexing may be appropriate. Martin Dillon, director of the OCLC Institute, defines four possible methods of description based on the importance of materials: traditional MARC-based cataloging, the use of the Dublin Core, the use of Internet indexing software with some editing and the use of Internet indexing software with no editing. Michael Gorman suggests a similar four-tier scheme: traditional MARC-based cataloging, "enriched" Dublin Core records, basic Dublin Core records and simple full-text searching (Oder 1998). Several initiatives are attempting to develop methods for using traditional as well as new technological solutions, including OCLC's Cooperative On-line Resource Catalog (OCLC 1998) and INFOMINE (Oder 1998).

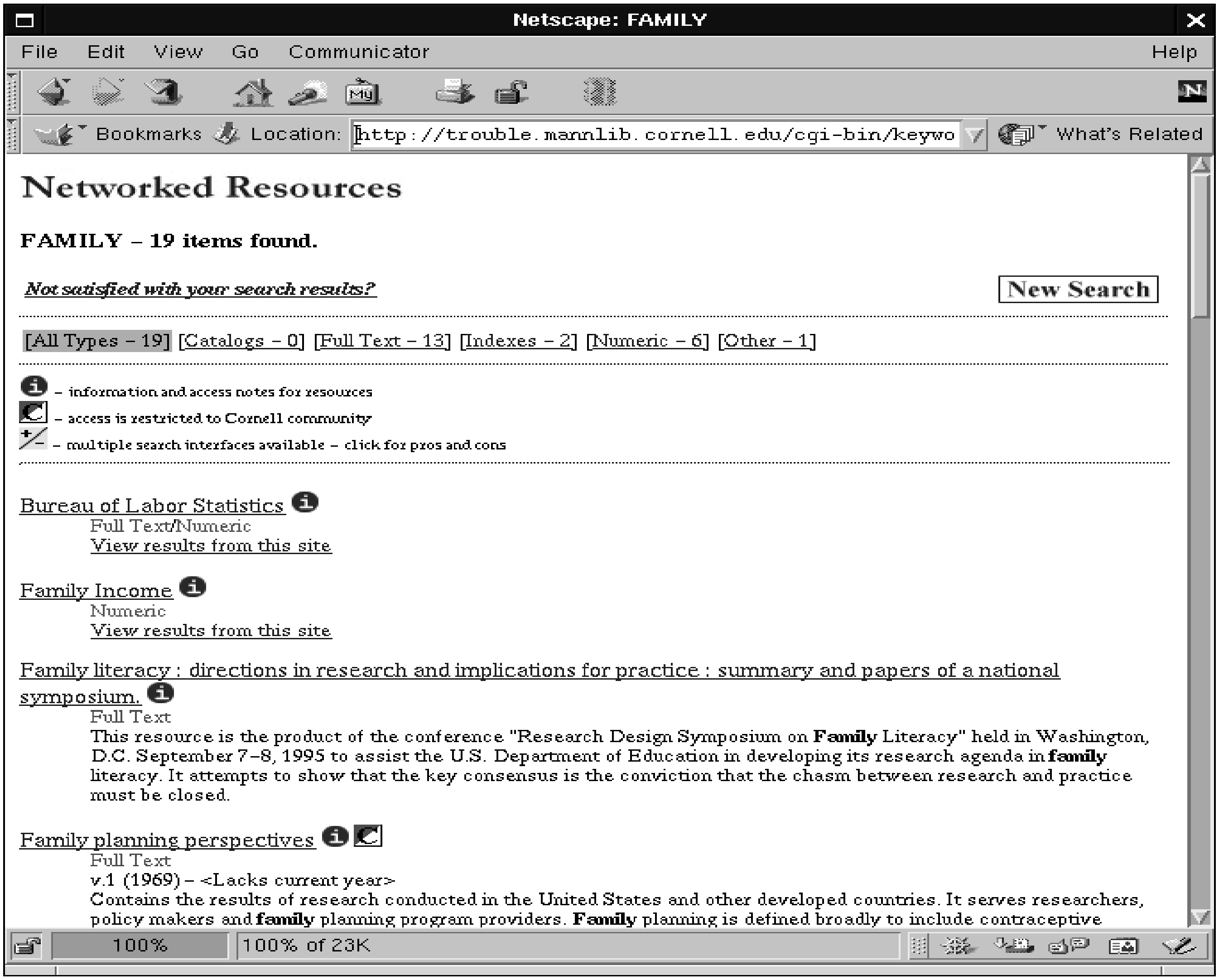
Project Overview

Cornell University's Mann Library has been selecting and cataloging electronic resources for over a decade. Selection criteria, which have evolved over time, are based on standards set by print materials in terms of quality, reliability, and appropriateness for the collection (Demas, McDonald and Lawrence 1995). Current efforts include, among others, selecting free and for-fee electronic serials (Weintraub 1998) and aggregations of resources (Walters et al. 1998). Subject bibliographers identify networked materials within the library's scope. They usually select specific titles for cataloging (such as Ecological Monographs) rather than entire sites (like JSTOR). Once they have determined that selection criteria have been met, acquisitions staff create preliminary records in the on-line catalog and record contact information and price and contract limitations as applicable. Cataloging staff complete the work of describing the items using full-level cataloging.

As part of the cataloging process, a Cornell University Library Gateway record is created. The CUL Gateway (http://campusgw.library.cornell.edu) is a Web-based system which provides a single point of entry to all networked electronic resources selected by the library. This system consists of a searchable database of electronic resource surrogates and is able to dynamically generate a list of resources appropriate to user queries. The records provide a hypertext link to the resource, the title, a description, the genre of the material, summary holdings, general subject categories and similar information. The CUL Gateway also handles many authentication issues related to for-fee networked electronic resources (Garrison and McClellan 1997). In addition, the CUL Gateway contains links to library home pages, library services, and other resources available to library users. For a more detailed discussion of the CUL Gateway and its users, please see the article by Karen Calhoun and Zsuzsa Koltay in this volume (Calhoun and Koltay 1999).

The main technical focus of this project was to integrate an Internet indexer into the already existing architecture of the CUL Gateway. The Harvest-NG indexer (http://www.tardis.ed.ac.uk/harvest/ng/) was chosen, mainly due to its flexibility and extensibility. The task of integrating Harvest-NG into the CUL Gateway was made easier by the fact that both were written using the Perl programming language. After Harvest-NG indexes the requested documents, the system imports the automatically generated metadata into its MySQL database (http://www.mysql.com/). The metadata is then available to be searched via the Gateway's keyword search function.

Each keyword search actually performs two searches: one of the Gateway's human-generated metadata and one of the automatically generated metadata. The results are merged into a single HTML document (figure 1). In this example, you see four results: the first two are the result of the Harvest-NG indexing and the second two are the results from a "traditional" CUL Gateway search. Notice that the traditional results have a cataloger-generated description, whereas the other results have no description. The Harvest-NG indexed titles also have an additional link to the "deeper" results available through the indexing (figure 2). These results present each document's title, URL, and an automatically generated description. The



Networked Resources

FAMILY – 19 items found.

Not satisfied with your search results? New Search

[All Types – 19] [Catalogs – 0] [Full Text – 13] [Indexes – 2] [Numeric – 6] [Other – 1]

i – information and access notes for resources
☾ – access is restricted to Cornell community
+/– – multiple search interfaces available – click for pros and cons

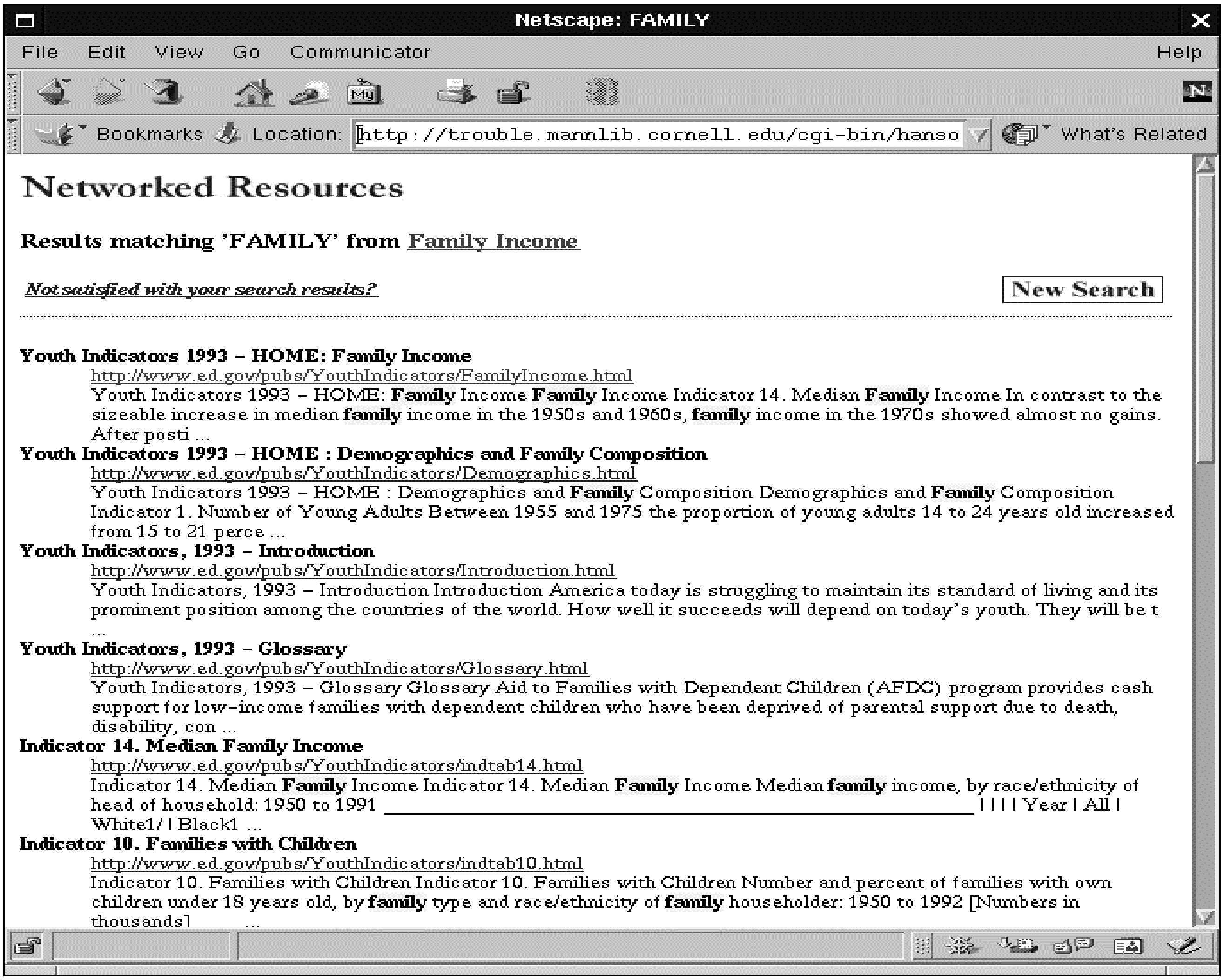
Bureau of Labor Statistics i
Full Text/Numeric
View results from this site

Family Income i
Numeric
View results from this site

Family literacy : directions in research and implications for practice : summary and papers of a national symposium. i
Full Text
This resource is the product of the conference "Research Design Symposium on **Family** Literacy" held in Washington, D.C. September 7–8, 1995 to assist the U.S. Department of Education in developing its research agenda in **family** literacy. It attempts to show that the key consensus is the conviction that the chasm between research and practice must be closed.

Family planning perspectives i ☾
Full Text
v.1 (1969) – <Lacks current year>
Contains the results of research conducted in the United States and other developed countries. It serves researchers, policy makers and **family** planning program providers. **Family** planning is defined broadly to include contraceptive

Figure 1. A sample top-level results page from the experimental version of the Cornell University Library Gateway.


Netscape: FAMILY
File Edit View Go Communicator Help
Bookmarks Location: http://trouble.mannlib.cornell.edu/cgi-bin/hanso
What's Related
New Search

Networked Resources

Results matching 'FAMILY' from Family Income

Not satisfied with your search results?

Youth Indicators 1993 – HOME: Family Income
http://www.ed.gov/pubs/YouthIndicators/FamilyIncome.html
Youth Indicators 1993 – HOME: **Family** Income **Family** Income Indicator 14. Median **Family** Income In contrast to the sizeable increase in median **family** income in the 1950s and 1960s, **family** income in the 1970s showed almost no gains. After posti ...

Youth Indicators 1993 – HOME : Demographics and Family Composition
http://www.ed.gov/pubs/YouthIndicators/Demographics.html
Youth Indicators 1993 – HOME : Demographics and **Family** Composition Demographics and **Family** Composition Indicator 1. Number of Young Adults Between 1955 and 1975 the proportion of young adults 14 to 24 years old increased from 15 to 21 perce ...

Youth Indicators, 1993 – Introduction
http://www.ed.gov/pubs/YouthIndicators/Introduction.html
Youth Indicators, 1993 – Introduction Introduction America today is struggling to maintain its standard of living and its prominent position among the countries of the world. How well it succeeds will depend on today's youth. They will be t ...

Youth Indicators, 1993 – Glossary
http://www.ed.gov/pubs/YouthIndicators/Glossary.html
Youth Indicators, 1993 – Glossary Glossary Aid to Families with Dependent Children (AFDC) program provides cash support for low–income families with dependent children who have been deprived of parental support due to death, disability, con ...

Indicator 14. Median Family Income
http://www.ed.gov/pubs/YouthIndicators/indtab14.html
Indicator 14. Median **Family** Income Indicator 14. Median **Family** Income Median **family** income, by race/ethnicity of head of household: 1950 to 1991 ____ | | | | Year | All | White1/ | Black1 ...

Indicator 10. Families with Children
http://www.ed.gov/pubs/YouthIndicators/indtab10.html
Indicator 10. Families with Children Indicator 10. Families with Children Number and percent of families with own children under 18 years old, by **family** type and race/ethnicity of **family** householder: 1950 to 1992 [Numbers in thousands] ...

Figure 2. A sample of the pages indexed using Harvest-NG.

description is taken from the HTML description META tag if available, otherwise it is generated from the beginning of the document. This page will give the user quicker and easier access to the relevant information on that site.

Eleven Internet resources were indexed to test the ability of the indexing software to cope with both simple and complex Web sites. Most of the resources indexed were selected by staff of Cornell University's Catherwood Library, but were not yet on the CUL Gateway. The Web sites fell into three categories: four serials, four aggregations, and three sites including non-HTML-based document formats such as GIF and PDF. This small test was planned as the first step in the process of determining the effectiveness of automatic metadata generation from Web documents identified by library staff. Not all of the sites had been fully cataloged and, as a result, information about each site varied in detail and human effort. Four of the Web resources had received full cataloging and had a detailed description written of it by cataloging staff. This variation provided an opportunity for staff to determine the optimal amount of information needed about each site. However, all resources chosen for this test were freely available over the Internet. Resources available for a fee will require a somewhat different approach. For a list of the resources indexed see table 1. A twelfth site, an aggregation, was identified for indexing. However, the site, GPO Access: Keeping America Informed (http://www.access.gpo.gov/), does not allow robot-indexing software to analyze its components. As a result, GPO Access could not be included in this test.

Evaluation

This limited-scale project provided an opportunity to test the adequacy of indexing large, complex sites as well as sites containing multiple file formats. The indexing was analyzed for technological success as well as for usefulness by technical services, collection development and public services staff.

Technologically, automated indexing of these sites showed potential for dealing with complex resources. Due to file "summarisers" available through Harvest, Web sites containing WK1 (Lotus spreadsheet) and PDF (Adobe Acrobat) file formats were indexed well. Other formats, such as audio files, were also indexed adequately because links to these files used descriptive text phrases In addition, the automatic indexing found and referenced aspects of the site that catalogers would not have. This detailed analysis could lead to reduced time and effort by selectors and catalogers.

However, several technological difficulties were discovered. Sites that exclude the use of robots to capture information, like GPO Access and many for-fee sites, were inaccessible via this process. This could result in discrepancies in the database in terms of expected levels of granularity or detail. In addition, the use of frames caused significant problems for this automated indexing method. Determining the proper context and document to return was made more difficult by the use of frames because sites using frames were broken into their sub-parts. It is possible, however, to develop automated methods to deal with the difficult structures that frames present to automated indexing software.

In addition to testing technological feasibility, preliminary feedback was gathered from public services, technical services and collection development staff. As a means of concept testing, staff were asked to evaluate

Table 1. List of Titles Indexed Using Harvest-NG

Title	URL
Serials	
Family Income	http://www.ed.gov/pubs/YouthIndicators/FamilyIncome.html
Journal of Extension*	http://www.joe.org/
Poultry Slaughter*	http://usda.mannlib.cornell.edu/reports/nassr/poultry/ppy-bb/
Rice Yearbook*	http://usda.mannlib.cornell.edu/data-sets/crops/89001/
Aggregations	
Bureau of Labor Statistics	http://stats.bls.gov/blshome.html
Demographic Data Viewer	http://plue.sedac.ciesin.org/plue/ddviewer/
Office of Technology Assessment Publications	http://www.wws.princeton.edu/~ota/ns20/pubs_f.html
World Wide HR	http://www.WorldWideHR.hq.dla.mil/
Multiple Document Formats	
U.S. Census Bureau PDF Publications	http://www.census.gov/prod/www/titles.html
United States Labor and Industrial History Audio Archive	http://www.albany.edu/history/LaborAudio/index.html
Wisconsin Herpetological Atlas Project*	http://www.mpm.edu/collect/vertzo/herp/atlas/atlas.html

* Indicates title was fully cataloged as well as indexed using Harvest-NG.

the results from two searches based on: expected results generated by searches, relevance of results to the search conducted, adequacy of information presented about sites and documents, and satisfaction with the extent of the indexing of each site. These questions were asked to determine whether or not we should pursue indexing only sites already cataloged as well as the amount of a site that would be indexed.

Staff members tended to view this project in relation to their experiences using, and helping patrons to use, Internet search engines. Not surprisingly, they preferred summaries produced by catalogers rather than text summaries produced from recording the text at the beginning of a page. Internet search engines usually generate descriptions of resources from the words at the beginning of a page, however, this text often does not provide adequate information to determine whether or not a document is relevant to a search. The summary of the site as a whole was more important to staff than human-generated summaries for each sub-part. Most staff preferred having at least the site-level summary because it allowed them to make a quick determination of relevance. They did want better quality results and more information to determine relevance than they usually receive from Internet search engines. Staff were not bothered by similarities to Internet search engines since they provide a well-known paradigm for information discovery on the Web.

Staff saw the results as a possible improvement over traditional Internet search engines. Staff members were enthusiastic about the possibilities of merging this level of detail with traditional cataloging. However, several staff pointed to indexed pages that were not particularly relevant to their searches suggesting that some clean-up of indexed items would be required. For instance, a search for the word "family" resulted in one site with only a passing use of this word, hence not a particularly useful result. In addition, staff were concerned about providing a context in which different levels of results could be understood by users; most felt this could be done through interface adjustments. Staff were less interested in the results generated from the indexing of serials than they were in the results from the indexing of aggregations and Web sites with multiple documents formats. One public services staff member was concerned that having some, but by no means comprehensive, results for very detailed search terms might lead patrons away from using more appropriate resources like bibliographic databases which would not be automatically indexed in this way. However, this is a problem for the traditional on-line catalog as well as databases of Internet resources.

Conclusion

This project was designed to test of possibility of using automated indexing software to enhance a library's database of Internet resources. This small-scale project is not a production system but could be used as the basis for a much larger initiative. Automated indexing software successfully improved the amount of information that was retrievable about complex Web resources, such as aggregations and electronic serials. Similar initiatives, like OCLC's CORC project (1998), will expand the set of tools that libraries can use to approach the Internet and enhance its usefulness for library users. In addition, more robust embedded metadata, such as the HTML META tag (Turner and Brackbill 1998) and XML (Flynn 1998 Light 1997) will continue to improve automated means of selecting and describing electronic resources available over the Internet.

The authors have identified several issues that require further development and consideration as libraries develop technological approaches to the Web. All of the resources that were tested in this project were available to any user of the Web. Resources requiring passwords, containing firewalls or limiting access in other ways will require very different technological solutions. Most for-fee resource producers will not permit robot-indexing software to scan their sites. In addition, we must provide better ways for users to understand the context of their search results. Interface improvements as well as selective editing of indexer output can help to provide such a context. Lastly, it is necessary to determine the number of levels of a site that should be indexed. At what level do most sites become too disorganized to be useful? Deciding the extent of indexing for each site is at the heart of determining the ultimate usefulness of automated indexing.

References

Atkinson, Ross. 1996. Library functions, scholarly communication, and the foundation of the digital library: laying claim to the control zone. *Library Quarterly* 66, no. 3: 239–65.

———. 1998. Managing traditional materials in an online environment: some definitions and distinctions for a future collection management. *Library resources and technical services* 42, no. 1: 7–20.

Calhoun, Karen and Zsuzsa Koltay. 1999. Designing for WOW!: The optimal information gateway. In *Racing toward tomorrow: Proceedings of the Ninth National Conference of the Association of College and Research Libraries*. Chicago: ACRL.

Chepesiuk, Ron. 1999. Organizing the Internet: the "core" of the challenge. *American libraries* 30, no. 1: 60–63.

Demas, Samuel, Peter McDonald, and Gregory Lawrence. 1995. The Internet and collection development: mainstreaming selection of Internet resources. *Library resources and technical services* 39, no. 3: 275–90.

Dillon, Martin and Erik Jul. 1996. Cataloging Internet resources: the convergence of libraries and Internet resources. *Cataloging & classification quarterly* 22, nos. 3/4: 197–238.

Flynn, Peter. 1998. Frequently asked questions about the Extensible Markup Language. Version 1.3. On-line. Available: http://www.ucc.ie/xml.

Garrison, William V. and Gregory A. McClellan. 1997. Tao of gateway: providing Internet access to licensed databases. *Library hi tech* 15, nos. 1/2: 39–54.

Gill, Tony. 1998. Metadata and the World Wide Web. In Baca, Murtha (ed.). *Introduction to metadata: pathways to digital information*, Los Angeles, Calif.: Getty Information Institute.9–18.

Kimmel, Stacey. 1996. Robot-generated databases on the World Wide Web. *Database* 19, no. 1: 40–43+.

Lester, Dan. 1995. Yahoo!: Profile of a Web database. *Database* 18, no.6: 46–50.

Light, Richard. 1997. *Presenting XML*. Indianapolis, Ind.: Sams Publishing.

Lynch, Clifford. 1998. The Dublin Core Descriptive Metadata Program: strategic implications for libraries and networked information access. *ARL: a bimonthly newsletter of research library issues and actions* no. 196: 5–10. Available: http://www.arl.org/newsltr/196/dublin.html.

Oder, Norman. October 1, 1998. Cataloging the Net: can we do it? *Library journal* 123, no. 16: 47–51.

OCLC. 1998. CORC—Cooperative Online Resource Catalog. Online. Available: http://www.oclc.org/oclc/research/projects/corc/index.htm.

Steinberg, Steve G. 1996. Seek and ye shall find (maybe). *Wired* 4, no.5: 108–114+.

Turner, Thomas P. and Lise Brackbill. 1998. Rising to the top: evaluating the use of the HTML META tag to improve retrieval of World Wide Web documents through Internet search engines. *Library resources and technical services* 42, no. 4: 258–71.

Vellucci, Sherry L. 1997. Options for organizing electronic resources: the coexistence of metadata. *Bulletin of the American Society for Information Science* 24, no. 1: 14–17.

Walters, William H., Samuel G. Demas, Linda Stewart, and Jennifer Weintraub. 1998. Guidelines for collecting aggregations of Web resources. *Information technology and libraries* 17, no. 3: 157–160.

Weintraub, Jennifer. 1998. The development and use of a genre statement for electronic journals in the sciences. *Issues in science and technology librarianship* no. 17. Online. Available: http://www.library.ucsb.edu/istl/98-winter/article5.html.

Younger, Jennifer A. 1997. Resources description in the digital age. *Library trends* 45, no.3: 462–81.