Using the Portal for the Discovery of Discipline-Based Electronic Resources

by

Sarah E. Thomas
Carl A. Kroch University Librarian
Cornell University Library

Contemporary society is at once omnivorous and highly selective. People today are coming to expect to have all manner of products and services at their fingertips, and simultaneously, they want them customized for a particular individual. You can order Levi's jeans programmed to your personal body specifications and conceivably, you could choose from any textile in the world to make a totally unique pair of pants.[1] Nick Donatiello, president of Odyssey, a marketing research firm, speaking at the JSTOR American Library Association June 2001 meeting, noted the struggle of the big TV networks to retain market share in a world in which the consumer increasingly prefers to view content he has profiled to watch at his convenience. Since we can now access hundreds of channels through cable and satellite, the consumer wants a tool to filter the diverse content available. New devices such as TiVo enable viewers to create "My channel" through a definition of preferences. The software interprets the viewer's preferences from this profile and independently identifies categories of programs which are consistent with the consumer's taste and previous selections. In the world of books, we are familiar with this feature from Amazon's "Customers who bought this book also bought...."

With these examples illustrating how the commercial world addresses consumer demand, we can examine the same trends in library user preferences for seeking and finding information. The portal is a mechanism for managing the wealth of information resources in a manner that structures information in a customizable way. Looney and Lyman note: "portals gather a variety of useful information resources into a single 'one-stop' Web page, helping the user to avoid being overwhelmed by 'infoglut' on the Web."[2] In a fairly generic example of a portal, Yahoo, one can create a personal portal using the "My Yahoo," feature.

The word "portal" has become very trendy and is used sometimes rather loosely. *The New York Times* cites over 1000 uses in its articles since 1996, and there are hundreds of vertical portals (specializing in a category of information, such as a discipline, or user, such as academics) as well as the broad, comprehensive portals such as AOL. Describing the consumer portal, Traffick, a site which tracks portals, states:

> "Portals offer a wide range of customization options and functionality including: Internet search and navigation; email; customized news, weather, sports, and horoscopes; planners, calendars, and contact managers; bookmark managers to save favorite web sites; real-time chat; message boards; original content on every imaginable topic; shopping; free home pages; "clubs" which function as makeshift intranets; small business services; and much more. Increasingly, major portals are seeing to it that vital content such as news, stock prices, and messages can be accessed with wireless devices and phones." [3]

Academic users of portals value their currency, the ability to locate a large quantity of information with relative ease through the use of a powerful search engine, the immediate

[1] http//www.levi.com/Originalspin/
[2] Michael Looney and Peter Lyman, "Portals in Higher education: What Are They and What is Their Potential?" *EDUCAUSE Review* (July/Aug. 2000): 30.
[3] http://www.traffick.com/story/07-2000-portalfaq.asp#stats

access to full-text and image files, and the adjunct features associated with portals such as news, links to other relevant sites, such as local information, and the capability of customizing and personalizing the portal to reflect individual tastes and requirements.

There are currently a number of efforts underway to translate the features of the portal to the academic and library environment. In doing so, librarians and others supporting research are attempting to unite the attractive aspects of the portal with some of the values of a library, especially those which promote careful selection and organization authenticity, consistency, and permanence.

Within the library, in the past few years, we have experimented with a variety of approaches to increase access to an expanding universe of information and to incorporate the concept of personalization into our tools. The library catalog has, until the late 1990's and the advent of the Web, been the primary locator tool for information acquired by libraries, and libraries have been, for many researchers, the starting and ending point for their information-seeking. The attributes of the catalog and the portal have been described in depth in my paper included in the *Proceedings* of the Bicentennial Conference on Bibliographic Control for the New Millennium[4] and in the Association of Research Libraries' newsletter, *ARL: a Bimonthly Report on Research Library Issues and Actions from ARL, CNI and SPARC.*'[5] As libraries began to provide access to a broader range of electronic resources, they first entered them into their online catalogs, and later created special gateways to networked resources. Many library home pages function as a portal to information of interest to a particular community, offering connections to campus information, local news and weather, as well as the library catalog and other information resources. Some offer the ability to customize the portal, as does North Carolina State University Libraries' "My Library@NCState," which provides library users the opportunity to create pathways to frequently used resources and to learn about new materials of particular relevance to them.

The dimensions of coverage are constantly changing. For many years libraries acquired a significant portion of the world's published output, and managed, between catalogs and printed indexes, to describe or provide access to much of it. Now the scene has changed. According to a study conducted at the University of California at Berkeley, the number of books published worldwide approached 1 million in 2000.[6] *Google*, the search engine, touts that it searches over 1 billion web pages.[7] Concomitantly, more faculty in universities are conducting their research on the Internet and perceiving a declining role for libraries. In a survey conducted by JSTOR in November 2000, faculty anticipated a decreasing dependence on their college or university library, with a decline from 48% to 38% for those who today consider themselves "very dependent" and a further slide from 13% to 21% for those who label themselves "not very dependent." Sixty-five percent considered it important for their library to serve as a gateway or starting point for research, and they predicted in five years this would be even less important, with just over half, or 56%, expecting to begin research at their library. [8] At the University of Washington a survey of faculty indicated that the availability of information technology had made 25% of them more likely to acquire information from a non-library source than they had been in the past. [9]

---

[4] Sarah E. Thomas, "The Catalog as Portal to the Internet," *Proceedings of the Bicentennial Conference on Bibliographic Control for the New Millennium: Confronting the Challenges of Networked Resources and the Web*, Washington, D.C., Library of Congress, Cataloging Distribution service, 2001, 21-37.

[5] Sarah E. Thomas, "Abundance, Attention and Access: of Portals and Catalogs," *ARL: A Bimonthly Report on Research Library Issues and Actions from ARL, CNI, and SPARC*, #212, p 1-3.

[6] http://www.sims.berkeley.edu/research/projects/how-much-info/

[7] http://www.searchenginewatch.com

[8] http://www.jstor.org/about/faculty.survey.ppt

[9] http://www.lib.washington.edu/survey2001WebPages/prelimfacultyresults.htm.

In the face of these trends, libraries are very much in a state of flux. They are exploring new models and altering old ones to create discovery and retrieval tools that will better serve the wired scholar. Several general products are in development, including Endeavor System's ENCompass, RLG's Cultural Materials Initiative, OCLC's Enhanced WorldCat, and the Association of Research Libraries' Scholars Portal. These products have the goal of crossfile searching of trusted data, bringing two desirable features together: convenient searching and the retrieval of high quality, reliable, and relevant resources. Several other specialized services have emerged to accomplish similar goals, but within a disciplinary framework.

Cornell's experience with making electronic resources accessible is a useful case study of the many different approaches that a single institution can take. Its various applications reflect both the maturing of technology and its understanding of user requirements as well as differing perspectives of librarians and computer scientists.

In the beginning, there was the Mann Gateway. Mann Library, which houses collections for agriculture and life sciences, developed the gateway concept in 1994, as it began an early drive to provide access to networked electronic resources. In 1997 the Cornell University Library adopted the Gateway as its single point of entry to electronic resources and expanded it to cover all subjects.[10] Catalogers modified records developed for Cornell's OPAC for use in the Gateway, adding subject class categories which allowed for disciplinary groupings. Owing to a concerted effort to license and link to digital materials of interest to scholars and students, the number of records in the Gateway increased rapidly, rising from a few hundred to a few thousand, and then doubling in 1999-2000 to almost 10,000 networked resources. Aggregator databases remained a challenge for catalogers, who labored to create individual titles for thousands of serials in a compressed timeframe. Wholesale records acquired from vendors were inadequate since they lacked the category codes for Gateway organization. The Library enhanced functionality with a "My Library " feature, which enabled users to bookmark frequently used URLs and provided updates on new monographs added to Cornell's collection. By spring 2001 the Gateway proved unwieldy in some subject areas; the system began to deteriorate in value. Some users found that the sets retrieved; by the rather general categories were too cumbersome to be useful. A Library taskforce had made a decision not to include e-books in the Gateway, since adding several thousand monographic titles would clutter up retrieval still further.[11] Yet, increasingly, users only searched the Gateway, missing out on both traditional and electronic resources.

Following an internal review in September 2001, the Library determined to convert the Gateway into eReference, a compendium of periodical indexes, catalogs, directories, dictionaries and encyclopedias. Although the Gateway had become a highly popular shortcut to electronic resources, the Task Force recommended channeling all searchers through the catalog, where users could discover all materials in the Library's collections, or could limit their searched by subject and format to achieve the effect of the Gateway. This approach simplified workflow in cataloging and benefited searchers in fields where paging through an alphabetical list of hundreds of titles had become unwieldy. In the process, however, the Library forced users who had been bypassing the catalog back into it. Despite this improvement, the catalog in its present form is inadequate because it only retrieves items recorded in the MARC format. Today's information objects have many different types of metadata. The Library is co-developing ENCompass with Endeavor to address its need to enable scholars to search across many files. ENCompass is a digital management tool which features cross-file searching of disparate collections and yields unified search results

[10] Karen S. Calhoun, Zsuzsa Koltay and Edward Weissman, "Library gateway: project design, teams, and cycle time at Cornell University," *Library Resources and Technical Services*, v. 43, no 2, (Apr. 1999), p. 114-22

[11] http://www.library.cornell.edu/staffweb/WGNRABENCEreport.htm

from a single query. ENCompass employs a hierarchical approach to organize data and metadata. Collections, containers, and objects are represented in various levels, with Dublin Core serving as the access to the highest level of intellectual organization, the collection. Other metadata types can be used to describe either containers or objects. Through its support for a variety of metadata and document type descriptions, ENCompass provides the capability of discovering citations, full-text, images, archival material—a full array of scholarly resources. These substantial enhancements to the catalog in the flexibility of managing complex, varied, and distributed data are a valuable enhancement. Universities view such tools as critical assets in gaining control over the many digital collections mushrooming on their campuses.

A vastly different approach to e-resources is being taken in the National Science Foundation's prototype National Science Digital Library. This project now underway at Cornell is a combined effort of the Computer Science Department and the Cornell University Library and is part of a multi-million dollar undertaking by NSF to produce a national library for science, mathematics, and engineering education. With the luxury of starting de novo, rather than bringing the legacy of tradition and millions of records, the Site for Science, as it is known, is striving to cover 10,000 open sources of digital information about science in a prototype under development. Its collections, which currently number less than 200, are developed through active selection by science experts working in a "tightly coupled federation of autonomous digital libraries interacting via agreed standards to function as a single coherent library", are harvested from metadata from OAI (Open Archives Initiative) compliant sites, or are nominated by contributors. The site features a "library," news, and a variety of topical entry points. There are interfaces for students, teachers, and a customizable "My Site." The Site for Science also links to tools, such as mapping devices, tours, exhibits, and images. It brings color, a wide range of materials, and clever subsetting of materials. Two library staff members have participated actively in the development of the prototype, identifying collections for inclusion and serving as metadata evangelists who proselytize to content creators about the value of standard metadata.

The Site for Science seeks to develop a scalable automated library system which integrated an interoperable technical infrastructure, metadata standards, services and organization, and ongoing research in digital library challenges. According to its website:

> " The key project goal is to be comprehensive in the approach to information science architecture, portal design, and production system administration of 'SITE for Science': to embrace every digital resource, for every level of education, in every field of science, mathematics, engineering and technology; to accommodate diverse content, metadata, protocols, formats, authentication and business practices; and to support students and instructors from the most junior to the expert. 'SITE for Science' supports several levels of interoperability: high-quality federations of NSDL members; harvesting metadata from digital repositories; and web crawlers to gather information from scientific web sites. The Cornell 'SITE for Science' team has developed a unique portal interface that allows items to be displayed with access to individual and collection metadata records. Maintaining the connection between collections and items in collections. This feature gives SITE library visitors the advantage of being able to search across all items or collections to find additional library and web-based information about topics and library resources. SITE for Science's scaleable production framework allows for new library services and broad and deep collections of materials to be aggregated and managed in multiple-user interfaces to the same underlying resources."[12]
>
> "Our work will bring the benefits of a major science library to everybody--including students and the general public--at a fraction of the cost of conventional libraries. It

[12] http://nsdlib.nsdl.cornell.edu/nsdl/portal/index.html?page=1

will provide access to good quality scientific information at the place of study, at work, or at home," explains William Arms of Cornell University, the project's Principal Investigator.[13]

One of the premises of the Site for Science is that the labor-intensive efforts of the professional cataloger or indexer must be supplemented and supplanted by automated data gatherers and by contributions from resource creators. The data harvesting programs rely on the application of a common metadata standard that is simple, spare, and capable of enhancement.

Another initiative still in the making is the Association of Research Libraries Scholars Portal. First described by Jerry Campbell, Dean of the University Libraries at the University of Southern California, the Scholars Portal has the objective of "construction of a suite of Web-based services that will connect the higher education community as directly as possible with quality information resources that contribute to the teaching and learning process and that advance research."[14] This project seeks to partner with a commercial software vendor to deliver cross-platform searching capability that will permit a single query to retrieve results from a local catalog, other designated library catalogs, selected Web sites, proprietary electronic resources, finding aids, and the host of library created and managed full-text and image resources. The product will be capable of mapping a search against various types of metadata and will be able to combine results from both public domain and restricted resources in a customizable presentation that reflects local licensing practice and priorities. In addition to facilitating the discovery of materials, the Scholars portal will also aid in its delivery to its academic users, and it will feature tools which support text-processing and citation management. Adjunct services such as electronic reference and instruction will complement the access mechanism. [15] A small group of ARL libraries expects to sign a contract with a vendor this fall and to Beta test the scholars portal at their institutions in spring 2002. The effort intends to leverage the investments of research libraries by creating a unified means of connected dispersed resources considered to be of high quality on the basis of their association with academic or research organizations. Like the Site for Science, the portal query will return federated search results. The Scholars Portal attempts to bring, for the digital age, the value previously conferred by the catalog and the library: information that has been selected for a particular community (academic institutions), high quality, trusted, and mediated. At the same time, it will introduce the convenience, speed, and increased results volume that attract consumers to Internet searching.

OCLC has also been developing its services to incorporate the portal concept. Through its enhanced Worldcat product, it proposes to upgrade its legacy catalog database to include a number of features that have achieved popularity in the Internet. In the CORC (Cooperative Online Resource Catalog) project, OCLC has encouraged the description of electronic resources for contribution to the database, following the traditional model of cataloger contributions to the union catalog. Their vision for the next three years embraces much more, however, with the addition of powerful Internet search engines designed to net other digital resources, and connections to commercial information providers such as Amazon seen as a way of meeting user demand for immediate access. OCLC plans to increase the number of cultural repositories with objects represented in WorldCat, and provide links to digital objects from historical societies, museums, individual collections, and archives. In addition to including metadata for a variety of text, image, audio, and data resources, the repository will expand to cover related materials such as reviews. Cooperative digital reference plus a range of document delivery options will round out its services. Building on its

[13] http://nsdlib.nsdl.cornell.edu/nsdl/portal/index2.html?page=9
[14] http:/www.arl.org/access/scholarsportal
[15] Brian Schottlaender, "Scholars Portal:Of, By, or For?" ALCTS Presidents Program, American Library Association, San Francisco, June 2001.

traditional base of libraries, OCLC will expand both to other non-profit partners and to commercial enterprises to increase significantly the quantity and types of resources over its current offerings.[16] Like the Scholars Portal and ENCompass, it seeks to evolve the traditional "brand" of libraries to cover a broader network of information, but to present this information within a trusted context.

There are many other examples on a smaller scope than these cited above which represent initiatives by librarians and other partners to gather and present information in a useful fashion for a select group of readers. These include simple webpages pointing to resources on a particular subject, such as the one dedicated to Slavic and East European materials created by Cornell's Slavic bibliographer. Another is a collaborative project developed by three university libraries, The University of Wisconsin-Madison Libraries, The Ohio State Libraries, and the University of Minnesota-Twin Cities Libraries, is the Digital Asia Library. The DAL is "a catalog of Asian Internet resources evaluated and selected by area specialists and cataloged by professional librarians."[17] The Staats-und Universitätsbibliothek Göttingen offers web pages for subjects for which it has official collection responsibility. Its *MathGuide*, available in English and German, covers over 1100 scholarly resources described using Dublin core.[18] It permits searching by subject or by format. Sources are evaluated for the quality of their content and clarity of presentation. The site specifies the standards it employs, such as ISO country codes and language codes. New additions are submitted via a web form. It's a relatively straightforward list of sources, rather than a true portal, which can be shaped to an individual's preferences. EULER, or European Libraries and Electronic Resources in Mathematical Sciences, aims to provide integrated access to bibliographic and full-text resources in math. [19] Like the Digital Asian Library, and the *MathGuide*, EULER uses Dublin Core as its metadata standard. Another variation on the portal theme is MIT's COGNET, which calls itself " a unique electronic community for researchers in cognitive and brain sciences, with in-depth current and classic text resources, and a dynamic interactive forum for today's scholars, students, and professionals."[20] It includes traditional bibliographic resources in full-text, commentary, job listings, conference-related information, and personalized workspace. Access to a portion of COGNET's contents is free, but other material is available to subscribers only.

A solid example of a discipline-based portal is Humbul Humanities Hub, which is part of the Resource Discovery Network (RDN) of the United Kingdom. [21] Its goal is to provide effective access to Internet resources which are carefully selected and catalogued by a network of collaborators. The site is interoperable with other disciplinary gateways that are a part of the RDN.

Like the web, these portals are in their infancy, with much maturation required to achieve a quality product that accomplished the goal of connecting the information-seeker with a targeted collection of information resources. The useful portal, as we idealize it today, would embody the following: It would provide one-stop shopping or a single point of entry for a well-defined community. For example, there would be a single portal for scholars working in the field of mathematics. The portal would be a gateway to distributed repositories, and users could customize their own views of the resources. A very high percentage of the repositories would be OAI compliant and they would describe their resources using standard description schemes and Dublin Core metadata. Ideally, there would be no duplication from site to site, and the search engine retrieving sites would also eliminate and duplicate

[16]" Extending the Cooperative:A summary of OCLC's Global Strategy"
http://oclc.org/strategy/strategy_document.pdf
[17] http://digitalasia.library.wisc.edu/about.html
[18] http://www.MathGuide.de/index.html
[19] http://www.emis.de/projects/EULER/objectives.htm
[20] http://cognet.mit.edu/
[21] http://www.humbul.ac.uk/about/index.html

results. To avoid redundant and labor-intensive efforts identifying resources, libraries and others, such as scholarly societies, should partner with each other to build quality sites with access to a broad swath of information resources. To achieve a degree of predictability and consistency, the portal needs a well-defined scope and coverage policy. The argument that we need local specialization seems hollow in the face that researchers consistently start their research outside of the catalog, outside the Library, on the Internet. Using techniques developed by newspapers, we can introduce local information when appropriate, but rely on the collectively built resource to avoid duplication of effort. Gannett Newspapers, a large American newspaper business, has developed an approach that preserves local newspapers while at the same time providing standard coverage of national and international news. They offer their papers a template of global information into which mix local and regional contributions can be made. Libraries must work to achieve similar efficiencies as they move into the description of Internet resources.

Louis Pitschmann, in his excellent work, "Building Sustainable Collections of Free Third Party Web Resources", defines four groups of criteria for selection: context, content, form/interface and technical criteria.[22] Context includes both the provenance of a site as a well as a site's relationship to other resources. "Each site added to a collection must be viewed as an integral part of a larger mosaic. Redundant, superfluous, unrelated, or poorly suited pieces will not enhance the collection; they will only encumber it and ultimately discourage or confuse users."[23] The content needs to reflect the standards of quality familiar to our print collections; items selected will be well-edited, authentic, and have a high degree of accuracy.

Portals should be up-to-date in both coverage and design. Users frequently begin their searches on the open Internet to cast a wide net and pick up the freshest pieces of information. When a portal does not provide access to current information, it loses value. Like a journal that has a long period of time between article submission and appearance or a catalog with a backlog, a portal without currency has diminished relevance. Like a carton of milk, the portal should have a freshness date to indicate its quality. Links, one of the critical features of a portal service, should be tested frequently using a combination of automated and human resources. Broken links should be repaired, and links to sites which have gone extinct should point to the last available copy of that site. One of the greatest advantages of the Web is the ability to connect new resources with retrospective materials and to expand constantly the network of related materials. Even if there are boundaries between proprietary and public information, users should be aware of the limited access information and be presented with options to gain access to it.

One of the implicit statements a library catalog makes is that the publications listed in it will be available, now, and in the future. Web resources challenge this assumption in several salient ways. First, digital technology has evolved rapidly, with technologies retired after only a few years. The portal management must endorse practices such as those recently promulgated in the OCLC/RLG guidelines on digital repositories *Attributes of a Trusted Digital Repository: Meeting the Needs of Research Resources*.[24] Secondly, since there will inevitably be links to URLs outside the direct control of the portal manager, there should be an alternate mechanism for duplicating the content at regular intervals. Brewster Kahle's work on the Internet Archive with the cooperation of the Library of Congress is one scenario that might address this need.[25] Version control would be an issue, but Kahle's concept of the "Way Back" machine might compensate for this. Kahle proposes to create a

[22] Louis A. Pitschmann, *Building Sustainable Collections of Free Third-Party Web Resources*, Washington, DC: Digital Library Federation, Council on Library and Information Resources, 2001, p. 14
[23] Ibid.
[24] http://www.rlg.org/pr/pr2001-attributes.html
[25] http://www.archive.org/

searching tool that can take requesters to the Internet as it was captured at a specific point in its history- the ability to turn back the clock to a moment in time.

The way to move portal development from the Stone Age to the information age is through the application of standards. Currently the Open Archives is promoting the adoption of the use of standard metadata schemes, including unqualified Dublin Core, which, coupled with a harvesting tool, will automate the process of compiling URLs.[26] In the portal design outlined here, there would remain a professional responsibility to evaluate the URL's returned through harvesting against the criteria for selection.

An underlying assumption of this paper is that the portal used in the academic and scientific communities should be global in nature. This not only implies close coordination and collaboration, but also adoption of common standards. There will be a requirement for multilingual interfaces, with different views possible, depending on the perspective the portal user selects.

In summary, the increasing volume of information in all formats creates a need for new approaches in organization. The subject-based portal is one means of providing access to electronic resources. The portal provides a convenient and efficient way for readers to locate information. Libraries and other organizations have begun to adopt the compelling features of the Internet portal and are merging them with traditional value-added services provided by libraries to build a new tool for information and knowledge management. Underlying the development is a drive to ride the wave of technological innovation to accomplish more at a lower unit cost. Presently, this area is highly transitional, and there are many interpretations of the portal concept. Librarians have much to contribute to improvements in access to electronic resources in terms of their expert understanding of organization and of the importance of the need for enduring access, but they will employ new means to accomplish these goals. New standards and practices will emerge. Through experimentation and collaboration they will advance the state of the disciplinary-based portal for the benefit of information-seekers and knowledge creators around the globe.

[26] Clifford A. Lynch, "Metadata Harvesting and the Open Archives Initiative," *ARL: A Bimonthly Report on Research Library Issues and Actions from ARL, CNI, and SPARC*, #217, p.1-9.