The Commercialized Web: Challenges for Libraries and Democracy

Bettina Fabos

Issue Editor
# Contents

**Introduction**  
*Bettina Fabos*  
519

**Part I: From Information to Commercial Highway**

**Links and Power: The Political Economy of Linking on the Web**  
*Jill Walker*  
524

**On Their Own: Students’ Academic Use of the Commercialized Web**  
*Samuel E. Ebersole*  
530

**Student Searching Behavior and the Web: Use of Academic Resources and Google**  
*Jillian R. Griffiths and Peter Brophy*  
539

**Cyber-Democracy or Cyber-Hegemony? Exploring the Political and Economic Structures of the Internet as an Alternative Source of Information**  
*Julie Frechette*  
555

**Part II: Harnessing the Web for Noncommercial Purposes**

**Current Developments and Future Trends for the OAI Protocol for Metadata Harvesting**  
*Sarah L. Shreeves, Thomas G. Habing, Kat Hagedorn, and Jeffrey A. Young*  
576
<table>
<thead>
<tr>
<th>Title</th>
<th>Authors/Contributors</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lessons Learned with Arc, an OAI-PMH Service Provider</td>
<td>Xiaoming Liu, Kurt Maly, Michael L. Nelson, and Mohammad Zubair</td>
<td>590</td>
</tr>
<tr>
<td>Collaboration Enabling Internet Resource Collection-Building</td>
<td>Steve Mitchell</td>
<td>604</td>
</tr>
<tr>
<td>Software and Technologies</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gateway Standardization: A Quality Assurance Framework for Metadata</td>
<td>Brian Kelly, Amanda Closier, and Debra Hiom</td>
<td>637</td>
</tr>
<tr>
<td>Strategies and Technologies of Sharing in Contributor-Run Archives</td>
<td>Paul Jones</td>
<td>651</td>
</tr>
<tr>
<td>About the Contributors</td>
<td></td>
<td>663</td>
</tr>
<tr>
<td>Index to Volume 53</td>
<td></td>
<td>669</td>
</tr>
</tbody>
</table>
Introduction

Bettina Fabos

The Global Brand of the Year in 2003 title did not go to Coca-Cola, Nike, or Starbucks, some of the most ubiquitous commercial names in our midst. Instead, it went to Google, a highly used but lightly promoted search engine, which beat Apple for the second year in a row. The leading brand consultancy, Interbrand, had surveyed about 4,000 “branding professionals” who determined that the Google brand had made the most impact internationally (Google voted, 2004). To think of Google—the most popular searching tool on the Web—as a brand is important for this issue of Library Trends because it underscores how closely mainstream online information resources are tied to the commercial economy.

The Web has been a commercial medium since 1995, when the government quietly sold the Internet’s backbone (previously controlled by the National Science Foundation) to private enterprise. Ten years ago we saw the beginning of a tremendous push—from the Clinton administration, Bill Gates, and the computer and telecommunication industries in general—to get schools and libraries wired. The push, it turns out, was not necessarily to bring the promised “universe of knowledge” (Clinton’s words) to all young and “lifelong” learners alike. Instead, the push was a careful public relations strategy to build up a user base so that the Web could become a viable commercial advertising medium (Fabos, 2004). Indeed, the rhetoric and accompanying media campaign of the mid-1990s was successful: in just five short years, the Web (as part of the larger Internet) became a mass medium—faster than any communication medium before it.

Before 1998, search engine providers such as AltaVista and Google were some of the most popular destinations on the Web. Beyond syndicating their services to search portals like Yahoo!, however, they generated low
revenue because they were merely the stepping stones to other content-rich pages containing banner ads. That all changed in 1998, when the startup Goto.com began combining the impartial algorithmic searches from search engine providers (usually one of the top five: AltaVista, AlltheWeb, Google, Inktomi, or Teoma) with a database of advertisers, so that many searches, unbeknownst to users, became prioritized according to the highest advertising bidder. Suddenly there was money in search engines. Goto syndicated its services to all the leading search portals, with the rationale that most people search for commercial products anyway. Then the impartial search engine providers themselves began to skew their searches in favor of commercial enterprise. Except for Google, all search engine providers implemented paid inclusion practices: accepting flat fees for including a client’s Web page in every search conducted.

In that year the Yahoo! portal, which had been syndicating Inktomi’s and then Google’s impartial search results, purchased Inktomi outright. Then the leading commercial search provider, Overture (formerly Goto), purchased AltaVista and AlltheWeb. And not long afterwards, Yahoo! purchased Overture, an acquisition that put three of the top search engine providers and the leading advertising index under one portal. And perhaps most significantly, Microsoft (by now regretting not getting into the search business sooner) tried to buy Google in 2003 but ended up building its own search engine provider, which was launched on the MSN site in February 2005 and will be bundled with Microsoft’s next Windows operating system, “Project Longhorn,” in 2006.

Search engines, once solely the online conduit for information, have taken on the contradictory role of conduit for online commerce. These days, even Google, the “ethical” search engine with the company motto “Don’t Be Evil,” is now focusing most of its attention on ad placement, either through its own search pages or through “contextual links” on other content pages (a practice that undermines the very integrity of its own PageRank algorithm). Indeed, the company’s success in this vein is all too evident: Google sold $1 billion of advertising in the last three months of 2004 (Markoff & Ives, 2005). Reflecting on the company’s motto after Google went public in 2004, a New York Times editorial stated: “Such idealistic talk out of Silicon Valley, so seemingly empowering back in 1999, seems embarrassingly naïve now that the party’s ended, at least for the rest of us” (Googling Google, 2004, p. 10). Such is the fate for all of us when Google the search engine became Google the brand.

This issue of Library Trends addresses Web content within the context of Internet commercialization and democracy. These are big ideas and problems, with potentially big solutions, so this issue has cast a wide net, pulling together voices from multiple disciplines, including communication studies, informatics, information management, research programming, computer science, engineering, and library science. I hope this issue highlights the
need for and value of continuing interdisciplinary cooperation and cross-fertilization. We have so much to learn from each other.

The issue is organized into two sections, with one section posing “the problem” of the commercialized Web and the other section posing multiple “solutions.” Part 1, “From Information to Commercial Highway,” examines the political and economic forces that shape and control online content. In “Links and Power: The Political Economy of Linking on the Web,” Jill Walker presents search engines as commercial entities that reinforce the most powerfully funded information—either commercial content or information that is already dominant in the mainstream media. This article provides a base for the next three, which look at the specific consequences of a commercial search environment on student research. In “On Their Own: Students’ Academic Use of the Commercialized Web,” Samuel Ebersole writes about high school students’ use of search engines, concluding that students’ research is inundated with commercial sources and that students do not have sufficient help in negotiating this environment.

In “Student Searching Behavior and the Web: Use of Academic Resources and Google,” Jillian Griffiths and Peter Brophy discuss student searching tendencies at the college level. Among their findings are a heavy student reliance on search engines rather than other academic resources, a general sense that search engines are all inclusive, and a significant difficulty in finding information and resources via search engines, with students trading performance for the path of least cognitive resistance. Finally, Julie Frechette goes one step beyond the world of search engines with an investigation of Web filtering software, which public officials are currently pushing in public libraries and schools. The article “Cyber-Democracy or Cyber-Hegemony? Exploring the Political and Economic Structures of the Internet as an Alternative Source of Information” illustrates how filtering software companies have become the second tier, after search engines, of lucrative Web properties that feed the commercial system. Like search engines, which now act to streamline and control much of the content on the Web, filtering software is suppressing certain kinds of “illicit” content while surreptitiously promoting commercial interests and commercial content.

What these four articles suggest is that, despite the huge amount of trust search engines and filtering tools continue to garner in the public sector, they are far from neutral and relentlessly steer users toward mainstream, mostly commercial content. That would be fine if one was interested in buying shoes. Indeed, the business sector tells us how happy we should be to find shoes that so precisely fit our marketing profile. But students and researchers looking for noncommercial, or at least nonmainstream, content, trying to gather a wide range of information containing as many disparate viewpoints as possible, or trying to access research that is controversial will not be successful, ultimately, in a research environment controlled by commercial interests.
Part 2, “Harnessing the Web for Noncommercial Purposes,” offers a glimpse of the many exciting international projects underway that are providing alternate routes to online content. If part 1 of this issue poses the problem of search engine commercialization, part 2 offers multiple solutions. Numerous computer scientists and digital librarians have been developing open source technology, such as the Open Access Initiative for Metadata Harvesting Protocol (OAI-PMH), iVia, and Data Fountains, that offer (and enhance) a user’s ability to search across multiple (that is, thousands of) subject gateways. These digital repository harvesting services imitate the functions and interface of a search engine, but they can be moulded to search in specific academic areas. In other words, one can create completely noncommercial searching environments that offer the scope and feel of a search engine. Do not be scared off by the unfamiliar acronyms because these developments have profound implications for academic research.

In “Current Developments and Future Trends for the OAI Protocol for Metadata Harvesting,” Sarah Shreeves, Thomas Habing, Kat Hagedorn, and Jeffrey Young report on the latest developments and future directions within the OAI community. Their article provides a succinct history of the OAI Protocol movement, which got its start in 2001 and since then has become widely adopted as an international standard in digital archiving and subject gateway development. Shreeves, Habing, Hagedorn, and Young are at the forefront of the OAI initiative. So are Xiaoming Liu, Kurt Maly, Michael L. Nelson, and Mohammad Zubair, who co-authored “Lessons Learned with Arc, an OAI-PMH Service Provider.” This article discusses the success of Arc, the first end-user OAI-PMH service provider. The searchable repository, which currently indexes about seven million records from several hundred subject gateways, has an immense scale and is a model of future academic Web searching. Besides detailing the Arc system, this article outlines the ongoing research devoted to improving Arc.

Beyond the OAI protocol, Steve Mitchell outlines the essence of iVia technology, a virtual library collection-building software platform. In “Collaboration Enabling Internet Resource Collection–Building Software and Technologies,” Mitchell details a decade of ongoing research at the library of the University of California that aims to help librarians and subject gateway developers more efficiently build metadata collections. Mitchell discusses iVia’s implementation in INFOMINE, a vast subject gateway that holds significant scholarly and educational resources on the Internet. He also mentions the ongoing work on the open source software system called Data Fountains, which expands upon the iVia system in areas of automated data harvesting that are intrinsically tied to the active skills and wisdom of participating librarians. Edward Almasy, co-director of the Internet Scout Project, is also dedicated to facilitating the development of easily searchable subject gateway systems for the academic community and beyond. In
“Tools for Creating Your Own Resource Portal: CWIS and the Scout Portal Toolkit,” Almasy describes a user-friendly means for building high-quality subject gateways. He and his colleagues have developed the Scout Portal Toolkit (SPT) and the Collection Workflow Integration System (CWIS), both complementary technical resources for subject gateway development. Almasy offers detailed descriptions of CWIS in particular, as well as information about hardware and software requirements and support in locating all appropriate software.

The next article describes another kind of toolkit. In their contribution, “Gateway Standardization: A Quality Assurance Framework for Metadata,” Brian Kelly, Amanda Closier, and Debra Hiom discuss the various strategies for streamlining metadata when adding new resources and maintaining subject gateways once they are built. Through trial, error, and careful testing since 2001, they have developed a toolkit that works as a straightforward self-assessment tool for subject gateway developers. Finally, Paul Jones, in “Strategies and Technologies of Sharing in Contributor-Run Archives,” illustrates the important developments in the collaborative subject gateway movement. He discusses contributor-run archives such as the Linux Documentation Project, the Degree Confluence Project, and Etree.org—all technologically inventive portals supported by passionate volunteers who cooperate to build these open source services. No librarians are involved—just experts and public citizens dedicated to sharing their knowledge and/or creative efforts with others.

I hope these writers’ contributions in this issue give us ideas when considering our fate as users of a Web that has become dominated by the powerful commercial economy. The ongoing work toward subject gateway development—all of it developed as free, open source software—provides a small but growing countervailing force to the commercialization of “the universe of knowledge.” Underlying all these efforts is the understanding that, for a democracy to function properly, one needs access to all kinds of information, not just information with a commercial purpose. Also underlying these efforts is the understanding that, in our commercial system, librarians and citizens interested in nurturing a public sphere must work together to control our destiny as information specialists—or somebody else will.

REFERENCES
Links and Power: The Political Economy of Linking on the Web

Jill Walker

ABSTRACT
Search engines like Google interpret links to a Web page as objective, peer-endorsed, and machine-readable signs of value. Links have become the currency of the Web. With this economic value they also have power, affecting accessibility and knowledge on the Web.

Links have always been fundamental to the Web. In the last few years their value has become regulated as search engines and other systems that find and define the structures of the Web increasingly index links and anchor text in addition to keywords and page content. In these projects, links are seen as objective, democratic, and machine-readable signs of value. There has been little or no critical discussion about this aspect of links, though link data is heavily used. This article discusses the implications and the power structures inherent in this relatively undocumented but influential change in the structuring of the World Wide Web and is an attempt to scan the field from a critical, humanist perspective.

TRACKING LINKS
A popular though clearly flawed assumption about the Web is that all its nodes are equally accessible. It is true that the Web has no formalized structure or centralized organization other than the rules of the mark-up and scripting languages we use to write and design Web sites. Even those rules are at times disputed: different browsers obey and interpret them in different ways. However, certain Web sites have always been more accessible than others.

Jill Walker, Department of Humanistic Informatics, University of Bergen, Postboks 7800, 5020 Bergen, Norway
© 2005 The Board of Trustees, University of Illinois
In the first years of the Web, most surfers used human-compiled directories listing sites by topic, portals provided by Internet Service Providers (ISPs) and other commercial actors, and search engines that indexed keywords and text phrases in Web sites. After a while, the extensive commercialization and the growing public awareness that highly ranked search results could be bought reduced the credibility of these sites (Introna & Nissenbaum, 2000). Google drastically changed the search engine game by not simply counting keywords but using links as the primary method of determining the value and thereby the deserved visibility of a Web site. Today most search engines have followed Google’s strategies and calculate the value of links. Indeed, almost any search engine you use today will use one of only three algorithms to power the search. The algorithm will belong to Google, to the Yahoo! Group (which recently gobbled up AlltheWeb, Alta-vista, and Inktomi), or to Teoma (Fabos, in press). So much for diversity.

So what was so new about Google’s algorithm? Google indexes links between Web sites and interpret a link from A to B as an endorsement of B by A. Links can have different values. If A has a lot of links to it, and C has very few, then a link from A to B is worth more than a link from C to B. The value determined in this way is called a page’s PageRank and determines its placement in search results (Brin & Page, 1998; Google, 2004; Marlow, 2001–2002). The PageRank is used in addition to conventional text indexing to generate highly accurate search results. Links can be analyzed more accurately and usefully than traffic or page views and have become both measures of success and dispensers of rank.

Links are increasingly being used in preference to content indexing, not only in search engines but, for instance, to identify communities of Web sites (Flake, Lawrence, Giles, & Coetzee, 2002), or, on a more local scale, to examine social networks and the transferral of memes between webloggers (Marlow, 2001–2002). Google is one of several companies that are developing a map of the Web that identifies connections among individuals, companies, organizations, and Web sites—a map that may prove priceless not only for improving search results but also for personalizing searches and, of course, ads. Sign up for Orkut, the Google-affiliated social networking system, and make all your social relationships machine-readable. Publish a Blogger.com weblog and use your Blogger account to comment on your friends’ weblogs—Google owns Blogger and can access this information. Get a free Gmail email account, with half a gigabyte of storage: “don’t sort; search,” says Google, and now they not only serve you ads based on the content of your emails but they have your personal correspondence in their ever-growing databases.

The extension of search into social networks and personal publication and communication shows that knowledge about the relationships among content is becoming the prime real estate of the Web (Kottke, 2004). Social relationships are simplified in systems like Orkut and Friendster so that
machines can process them. Similarly, links between Web sites are assumed to provide an objective measure of value and to be a sign of peer endorsement. This reductive view of links and its implications should be examined more closely.

**AN ECONOMY OF LINKS**

Links have a direct value on the Web and can be seen as a pseudomonetary unit. A Google search on currency of the Web shows that this is not a novel idea, though it is little theorized. Conventional thinking has assumed that linking from A to B takes value from B and adds value to A. Lawyers have complained that linking to another site’s news items, for instance, may be a copyright violation, and companies have sued against those who link to their site (*Kelly v Arriba Sort Corp.*, 1999). Though more sophisticated, Ted Nelson’s (1982) concepts of transclusion and transcopyright belong to a similar paradigm where content is value and links are mere mechanics, an outside vehicle for the transmittal of content rather than the item of value itself. In its fully implemented state, transcopyright sees a link from A to B as A using something owned by B, which readers should pay for in the form of a micropayment. This makes perfect sense in a traditional, product-oriented economy where content is king. B manufactured a product that A’s readers consumed and should therefore pay for. After Google, it makes no sense at all. The economy of links is not product oriented. It is service oriented, and the service is the link. The link is an action rather than an item, an event rather than a metaphor (Miles, 2001a).

When I link to B, I give B a link. That link translates into a precise (though undisclosed) value in Google’s PageRank and in other indexing systems like Blogdex. The link has a clearer value to B than the content of B’s page has to me or to my readers. I pay B for B’s content with my link.

This instrumental view of links does not exclude its other qualities. Many people creating or following links on the Web link generously, carefully, or haphazardly but without thinking of the economy of links and their value. Some choose to ignore the mercantile qualities of links; many more are unaware of this aspect of links. Even links created solely to increase a page’s placement in search engines may have or acquire other meanings or functions as well. This is the excess of the link, which can also be seen in relation to Bataille’s concept of a general economy, an economy that is not driven by scarcity (Miles, 2001b). Yet even if we are unaware of or refuse to participate in the economy of links, the pervasiveness of link indexing and valuation in search engines and other mapping strategies makes it impossible to entirely avoid this new restricted economy.

Google has not published the most recent algorithms behind their search technology, but the basic system is more or less known (or surmised) by search engine optimizers and manipulators. One striking effect of the PageRank system is link drain. Each Web page passes on a percentage (85..."
percent in prototype, possibly the same now) of its own PageRank score to other pages it links to, and this percentage is shared equally between any pages it links to, whether they are on the same site as the first page or on other sites. In addition, PageRank allows feedback loops between pages, so a link from A to B gives B a higher rank, and a link back returns some of that score to A (Ridings, 2001). Some people believe that linking to certain sites, or participating in link farms, can reduce the anchor site’s PageRank as well (Forum discussion, 2002). Rumors and theories abound at sites for Webmasters and search engine optimizers, based, for instance, on posts by an alleged Google employee calling himself “GoogleGuy,” who repeatedly “advises webmasters to avoid ‘linking to bad neighbourhoods’” (Sobek, 2004, PRO section). From this, and from the dreaded PageRank zero penalty sometimes incurred by sites that have dealt in the black market of linking, the priesthood of the search engine optimization world details theories of BadRank, an unofficial equivalent to PageRank. If you link to a site with high BadRank, this theory goes, your BadRank will increase (Sobek, 2004). Whether or not this is true, it is believed to be true by the devout, and these beliefs are integral to the ways in which we think about and use links.

If links are the currency of the Web, what is the exchange rate? Though links clearly have a value that is internally important on the Web, and that can have external real world implications (in sales for commercial sites and cultural capital and reputation for others), there is as yet no standardized exchange rate between links and real world currencies. Affiliate programs and banner ads could be seen as establishing an exchange market, but these are based on more than the presence of a link. Most banner ads pay only for click-throughs, and affiliate programs run by Amazon and others only pay when a link-follower makes a purchase. Though using Ebay as an exchange booth, as Everquest citizens do to sell their virtual treasure for real dollars, is only a gimmick as yet (Hiler, 2002b), this kind of sale and the sale of links rather than click-throughs in advertising could in time permit us to see links as an independent and real currency. As Castronova has demonstrated of the massive multi-player online game Everquest, non-tangible worlds can have real economies (Castronova, 2001).

Though open exchange of links for real world money is as yet infrequent, there is a black market for links. You can pay dollars or kroner or yen to buy links to your site from link farms, circles of sites with nothing but links. There is also a common law perception of link prostitution or link slutting: shamelessly selling one’s integrity for links. Though these practices are not yet illegal in any nation state, they are in practice outlawed on the Web. If Google discovers such attempts to manipulate a site’s PageRank, the site is penalized by being given a PageRank of zero. Due to Google’s high level of control of access to Web sites, this is equivalent to exile.

The more common form of trade in this economy of links is barter
exchange. Reciprocal linking and link exchange are common practices and are loosely organized as favors or more systematically in Web rings and blogrolling. Link slutting can be a consensual exchange of favors rather than a sale. “Link incest,” or linking inside of a community, is encouraged and often automated in community sites such as LiveJournal, Xanga, and Blogspot. These informal exchanges and the prolific linking in certain looser communities, especially among weblogs, subvert Google’s objective measurement of links (Hiler, 2002a).

Power and Knowledge
Links have value and they give power. Power and knowledge are intimately connected: “There is no power relation without the correlative constitution of a field of knowledge, nor any knowledge that does not presuppose and constitute at the same time power relations” (Foucault, 1977, p. 27). There is no moral high ground here where we can ignore the political economy of links and remain pure and clean, thinking only of the felicity of links, their usability or functionality or beauty. We are participants in this power structure whether we like it or not. We can criticize it, reflect upon it, approve of it, or try to subvert it. We must not ignore it. This standardization of links and their value will shape what the future finds. It defines what can be found. It defines knowledge.

References


On Their Own: Students’ Academic Use of the Commercialized Web

Samuel E. Ebersole

Abstract
This article reviews research conducted in 1998–99 examining students’ perceptions and uses of the World Wide Web for academic purposes. Recent developments in the Web that may be of particular interest to educators and parents of students are considered.

Since the mid-1990s the Internet, and more specifically the World Wide Web, has been eagerly adopted by school districts, administrators, teachers, parents, and students. Recent data from the National Center for Educational Statistics indicates that, in the fall of 2002, 99 percent of public schools and 92 percent of instructional classrooms were wired for Internet access (Kleiner, Lewis, & Greene, 2003). This is even more impressive when you compare 1994 figures, which estimated that 35 percent of schools and 3 percent of classrooms had Internet access. The latest in a long line of technological solutions to our educational woes, the Web, and its evangelists, promise no less than a radical restructuring of the way that students access and acquire information. However, some have raised concerns about the value of the Web as an educational resource. Historians have noted that the use of the Web in a public school setting marks the first time that the end user controls the process of choosing the content to be consumed.

To this end, critics have pointed to the incredible range of content accessible via the Web and its potential for distracting students from the task at hand. Hecht (1997) argued that “having the Internet in the classroom is like equipping each classroom with a television that can be turned on at any time and tuned in to any of 100,000 unrestricted channels, only a tiny
fraction of which are dedicated to educational programming (and even those have commercials)” (p. 15). McNealy (1999) voiced a similar concern when he wrote, “Right now, putting students in front of Internet terminals is no better than putting them in front of TV sets. It may even be worse” (p. 17A).

Public education’s adoption of the Web as a tool for research and as an alternative to traditional resources raises several issues related to the notion of functional equivalence. First, the wide range of content available via the Web allows it to serve numerous “functions” for students. Second, time spent using the Web in school is time not spent in activities that are displaced by Web use. And finally, the value of the Web for academic research is contingent on the quality of the research material contained therein (Bennett, Wilkinson, & Oliver, 1996). Educators’ concern about the unevenness of the quality of information available via the Web is obvious when one reviews the many Web sites devoted to critical thinking skills and Web site evaluation tutorials. The question remains for public schools and the whole of society: With the stakes so high, how can we harness this unwieldy resource so that it serves our educational goals and purposes?

Earlier Research

Research conducted in 1998–99 in ten public schools in a Western state found that students believe the Web to be a valuable resource for educational activities; the study also found, however, that students are often unsuccessful in finding appropriate or useful resources on their own (Ebersole, 1999). Approaching the research from a mass communication perspective, this study applied uses and gratifications theory to the questions surrounding students’ attitudes and opinions about the Web: what purpose it served for them, how they used it, and whether these were related. The study combined quantitative and qualitative research methodologies and several data-gathering approaches with a sample of middle and high school students drawn from ten public schools.

A paper survey was administered first to approximately 800 students. The survey contained 75 items designed to measure students’ (1) affinity for the Web, (2) assessment of the value of the Web for various purposes, (3) skill level for computer and Web use, and (4) uses of, and/or reasons for not using, the Web. The 40 use statements in the survey were generated by students’ anonymous responses to an open-ended question asking them to list several things “that the Web is good for.” These statements, as well as others generated during a pilot study, were presented as 5–item Likert scales that attempted to measure students’ use of the Web at school.

Second, a computer-administered survey requested responses from students as they began to access the Web from the schools’ media centers. This brief instrument asked only four questions: grade level in school, gender, how much the student uses the Web during an average week, and the
student’s purpose for using the Web at this particular time. For the fourth question, the choices presented to the student were factors identified by Principle Components Analysis of the use statements from the paper survey. The seven uses for the Web as presented in the computer-administered survey were “for research and learning,” “to communicate with other people,” “for access to material otherwise unavailable,” “to find something fun or exciting,” “for something to do when I’m bored,” “for sports and game information,” and “for shopping and consumer information.” As an option to the seven use statements presented, the student could select “other” and use a text box to enter a use that better described his or her purpose for using the Web at that particular time. The phrasing of the question, “What is your purpose for using the Web at this time?” was designed to measure gratifications sought and the “behavioral intention” (Palmgreen & Rayburn, 1982) of the student.

The final step in the data collection process was to gather a sample of Web site addresses (URLs) accessed from the media centers’ computers during the survey period. Approximately 123,000 URLs were collected from the computers on which the surveys were installed. The URLs were examined to determine the number of Web sites from the five generic TLDs (Top Level Domains). Also, a random sample of the 123,000 URLs was drawn and selected sites were reviewed and coded by two educators who had been invited to participate in the study. The coders—media specialists employed by a local school district—were asked to visit and explore a Web-based tutorial designed to train users to evaluate Web sites in order to determine their suitability for use as research sources for middle and high school students (Schinker, 1997). Some of the categories used for evaluation were Web address, content, credibility of the author, revision date, and links. A meeting was held with each of the coders to discuss criteria to be applied to the Web sites and to answer questions about the coding process. Once intercoder reliability was established at an adequate level (alpha = .92) the coders reviewed the 500 randomly selected sites and assigned a use category. Next, they rated each site based upon its perceived value as an educational resource.

**Results and Discussion**

The results of the two surveys and the content analysis of sites visited by students suggest that students believe the Web to be an important and valuable educational resource, but they are not consistently successful at finding appropriate and educationally valuable sites. Respondents to the computer-administered survey gave the following reasons for using the Web: “for research and learning” (n = 541, 52 percent); “to communicate with other people” (n = 74, 7 percent); “for access to material otherwise unavailable” (n = 55, 5 percent); “to find something fun or exciting” (n = 85, 8 percent); “for something to do when I’m bored” (n = 56, 5 percent); “for sports and game information” (n = 65, 6 percent); and “for shopping
and consumer information” \((n = 10, 1\% )\). In addition, 165 students (16 percent) chose not to select from the seven options presented. Of these, 94 students elected to write in a response to this question. The write-in responses offered by students to explain their purpose for using the Web were grouped into categories as follows: specific research topics \((n = 20)\), sexually explicit material \((n = 20)\), games and amusements \((n = 14)\), general research and learning \((n = 11)\), combinations of things \((n = 10)\), communication \((n = 5)\), and other unclassified \((n = 14)\). However, an analysis of Web sites visited by students revealed a different story. First, an analysis of the most frequently visited TLDs was conducted. Of the total URLs collected, 77 percent \((n = 94,426)\) were from the .com domain; 5 percent \((n = 6,289)\) were from .net; 5 percent \((n = 5,704)\) were from .org; 4 percent \((n = 4,842)\) were from .edu; 1 percent \((n = 1,640)\) were from .gov; 1 percent \((n = 1,403)\) were from .us; and 7 percent \((n = 8,767)\) were from other or unidentified domain names. These numbers stood in contrast to the distribution of domain names that made up the state of the Web at that time. According to a survey of Web domain names by host count conducted by Network Wizards at the time of the study, the actual make up of the Web was not as heavily skewed toward the commercial domain sites as the student sample would suggest.

The reason this is significant is that when educational media experts ranked a sample of 500 Web sites for “suitability for academic research,” commercial sites received the lowest mean score \((1.59\) on a scale of 1 to 3, with 1 = not suitable, 2 = questionable, and 3 = suitable). Similarly, Web pages from the .gov \((3.0)\), .org \((2.78)\), and .edu \((2.44)\) domain names were rated more favorably by coders but visited much less frequently by students.

Another area where students’ survey responses seemed at odds with the data collected from actual sites visited is the “intended use” or “purpose” for using the Web. As stated earlier, students were asked, “What is your

<table>
<thead>
<tr>
<th>Domain Name</th>
<th>Number of sites (1999)</th>
<th>% of total</th>
<th>Number of sites (2004)</th>
<th>% of total</th>
</tr>
</thead>
<tbody>
<tr>
<td>com (commercial)</td>
<td>12,140,747</td>
<td>41.9</td>
<td>48,688,919</td>
<td>30.3</td>
</tr>
<tr>
<td>net (network)</td>
<td>8,856,687</td>
<td>30.6</td>
<td>100,751,276</td>
<td>62.7</td>
</tr>
<tr>
<td>edu (education)</td>
<td>5,022,815</td>
<td>17.3</td>
<td>7,576,992</td>
<td>4.7</td>
</tr>
<tr>
<td>us (United States)</td>
<td>1,562,391</td>
<td>5.4</td>
<td>1,757,664</td>
<td>1.1</td>
</tr>
<tr>
<td>org (organization)</td>
<td>744,285</td>
<td>2.6</td>
<td>1,332,978</td>
<td>0.8</td>
</tr>
<tr>
<td>gov (government)</td>
<td>651,200</td>
<td>2.2</td>
<td>676,595</td>
<td>0.4</td>
</tr>
<tr>
<td>Total</td>
<td>28,978,125</td>
<td>100.0</td>
<td>160,784,424</td>
<td>100.0</td>
</tr>
</tbody>
</table>

purpose for using the Web at this time?” as they logged into the Web from their schools’ media center computers. Later, educational media experts were asked to assign the same categories to a sample of 500 sites visited by students. The disparity between self-reported uses of the Web and evaluators’ assessments of sites visited is indicated by Figure 1.

**Figure 1. Student Self-Reported Use Compared to Use as Assigned by Media Experts**

Note: Use categories are as follows: 1 = research and learning, 2 = to communicate with other people, 3 = for access to material otherwise unavailable, 4 = to find something fun or exciting, 5 = for something to do when I’m bored, 6 = for sports and game information, 7 = for shopping and consumer information.

---

**Table 2. Mean Suitability for Academic Research of Sites by Leading Domain Name**

<table>
<thead>
<tr>
<th>Domain</th>
<th>(N)</th>
<th>Mean Suitability for Academic Research as Assigned by Coders</th>
</tr>
</thead>
<tbody>
<tr>
<td>.com</td>
<td>(410)</td>
<td>1.59</td>
</tr>
<tr>
<td>.org</td>
<td>(25)</td>
<td>2.78</td>
</tr>
<tr>
<td>.edu</td>
<td>(16)</td>
<td>2.44</td>
</tr>
<tr>
<td>.net</td>
<td>(12)</td>
<td>1.75</td>
</tr>
<tr>
<td>.gov</td>
<td>(9)</td>
<td>3.0</td>
</tr>
<tr>
<td>.us</td>
<td>(5)</td>
<td>2.0</td>
</tr>
<tr>
<td>other</td>
<td>(23)</td>
<td>1.94</td>
</tr>
</tbody>
</table>

*Note: 1 = not suitable, 2 = questionable, 3 = suitable*  
*Source: Ebersole, 2000.*
The disparity between students’ self-reported uses of the Web and evaluators’ assessments of sites visited invites several possible explanations. First, students and educational media experts would be expected to apply different standards and criteria when evaluating the same Web site. In other words, if a group of students had been asked to apply the same use standards to the same 500 sites evaluated by the experts, we would expect there to be significant differences. Also, there are expectations for Web use, often outlined in schools’ Acceptable Use Policy statements, that undoubtedly affect students’ responses, even when anonymity is provided to survey respondents. Students were likely to respond to this question with an answer that they believe to be appropriate when using the Web at school—namely, academic research. However, there may be an equally valid explanation. It could be that students’ initial intentions were sidetracked by several factors, for example, distractions created by “entertaining” Web sites available at the click of a button, failure to readily distinguish between scholarly and commercial content, failure to find relevant material because of poor search strategies, or search engine results that direct users to less appropriate Web sites (as defined by academic research goals). In the following section of this article I will explore ongoing developments in the structure and character of the Web that may be contributing to these impediments to effective use of the Web in a public school setting.

**Recent Developments**

Since its inception, the Web has shown a remarkable pattern of growth, both in raw size and in terms of becoming an increasingly commercial enterprise. The first issue—the Web’s size and diversity—is generally perceived to be one of its greatest attributes. For those looking for information, however, clutter is a very real problem. As Shenk (1997) observed, too much information, what he calls “Data Smog,” can be, literally, too much of a good thing. Recent estimates put the number of Web pages at well over 6 billion, up from approximately 2 billion in 2000. And more importantly, the growth appears to be greatest in the commercial sector. Dot net and dot com Web sites now account for over 90 percent of all sites as measured by TLD host count (Internet Systems Consortium, 2004). And as you may recall from the research reported earlier, Web sites from these domains received the lowest ratings for “suitability for academic research.” Finding Web sites appropriate for the academic enrichment of this target audience can be like finding the proverbial needle in a haystack. In this case the haystack contains many needles, but the size of the haystack is enormous and the needles are remarkably similar in size, color, and texture to the stalks of hay. Or to use another metaphor, even when you are really thirsty, it is easier to sip from a straw than to try to drink from a fire hose.

In this environment of an overabundance of data, the hunt for usable information usually begins with a search engine. Research suggests
that Web users often start with a search engine when looking for specific information, and in a recent survey 56.3 percent of respondents said that they used a search engine at least once a day (iProspect, 2004). As others have already suggested, this reliance on search engines may be instilling a false sense of security, or at least an undeserved confidence, in the search results’ accuracy, relevance, and completeness.

Although it is not evident to the casual surfer, the Web search industry has been contracting in recent years, largely because of mergers and acquisitions. At present three companies dominate the search engine provider market. Google, Teoma, and Yahoo!, which recently absorbed AltaVista, Inktomi, and AlltheWeb, provide the algorithms that return search results on most of the major search portals. For example, the new search portal from Amazon (www.A9.com) is powered by Google; AskJeeves is powered by Teoma. So while there may be an appearance of many options and search engines from which to choose, it is, in fact, a mirage.

But it is the practice of combining algorithmic searches with those from commercial search databases that gives even greater cause for concern. With the notable exception of Google, search portals frequently display results without clear indicators to differentiate the paid results from the unpaid. This practice has resulted in a financial boon for search portals that previously had been unable to capitalize on their success at attracting consumers’ attention. However, for the academic surfer the practice compromises the integrity of the search while, at the same time, biasing the results toward commercial enterprise.

It is not just “pay for listing” or “pay for positioning” schemes that raise questions, however. Critics have suggested that current search engine policies and practices call into question the veracity of their results. For example, in this volume Walker (see also Walker 2002) argues that links to and from Web pages are interpreted by search engine algorithms, which in turn determine search result relevance, and this relevance translates into power that controls access to information. While the search engine providers do their best to prevent disclosure of their search algorithms, countless search engine marketing businesses have sprung up providing the latest “cheats” designed to manipulate the results to favor their clients.

Even without overt manipulation of search results, search engines may be delivering results that reflect inherent biases. According to Introna and Nissenbaum (2000), search engine results “give prominence to popular, wealthy, and powerful sites at the expense of others” (p. 181). They go on to argue that commercial search engines cannot be expected to correct these injustices but rather an alternative must be devised to ensure that the Web is able to exist as a “public good.” This notion of a public good implies that the Web ought to serve the interests of all members of society and all manner of Web content creation and dissemination, not just those that are commercially viable. In this case, alternatives to commercial search
engines ought to be provided to students who use the Web for academic pursuits.

Despite its origin in scientific research and educational pursuits, it did not take long for marketers, advertisers, and public relations practitioners to find the Internet. What they found was an uncharted land that rivaled their wildest dreams. More than a decade later the Web remains the least regulated of all mass media. Although the dot com bust of the early 2000s slowed the commercial expansion of the Web, we are beginning to witness a strong rebound in every area, including online advertising. Today, the one feature that best defines the Web is its unrelenting commercialism. For those who have a vision for the Web that extends beyond the virtual strip mall (for example, Fabos, 2004), this defining attribute must not go uncontested.

It is not just the omnipresent commercialism of the Web that raises concerns, however. The blurring of lines between fact and fiction, between opinion and news, and between credible and incredible reporting also draws into question the usefulness of the Web for young scholars. A high level of sophistication is necessary to understand the hidden economic relationships that often influence content and access to content. For example, students are routinely cautioned about personal postings by Web authors who have strong opinions but weak credentials. But how many are being told about the economic structure that makes a popular blog not only highly relevant to search engines but places targeted ads on the blog intending to reach surfers who match the desired profile? Blogger Steve Rubal (2004) refers to the intersection of public relations and participatory journalism as “Micro Persuasion”—but it may have maximum impact on the veracity of online information. Consider, too, the commercially oriented Web site that provides the equivalent of product placement advertisements. We have become relatively sophisticated and sensitized to product placement in film and television, but when it is buried in the text of an essay or opinion piece, it may be undetectable to the vast majority of unsuspecting readers. Corporate Web sites also routinely publish “white papers” that are favorable to their products and services, but they are without the benefit of objectivity and full disclosure.

CONCLUSION

One thing appears to be clear from this research and other studies conducted with middle and high school students—effective use of the rich resources provided via the Web is complicated by a number of intervening variables. In 1997 a study of sixth and ninth grade science students found that they were often unsuccessful in finding useful academic information. Lyons, Hoffman, Krajcik, and Soloway (1997) observed that “one overall theme is clear from the data: students need a tremendous amount of support to be successful in online inquiry.” Several years later my research
confirmed that middle and high school students are frequently unsuccessful in finding appropriate information either because of poor search strategies or the distractions that abound on the Web (Ebersole, 2000). Today the problem continues as the Web expands and mutates faster than we can equip students with the skills necessary to make sense of this multifaceted resource. And all too often, students searching the Web for information on a particular topic are on their own—sifting through a huge but uneven collection of resources without the aid of editors, research librarians, or content guides.

**References**


Student Searching Behavior and the Web: Use of Academic Resources and Google

Jillian R. Griffiths and Peter Brophy

Abstract
This article reports results of two user studies of search engine use conducted to evaluate the United Kingdom’s national academic sector digital information services and projects. The results presented here focus on student searching behavior and show that commercial Internet search engines dominate students’ information-seeking strategy. Forty-five percent of students use Google as their first port of call when locating information, with the university library catalogue used by 10 percent of the sample. Results of students’ perceptions of ease of use, success, time taken to search, and reasons for stopping a search are also presented.

As part of its commitment to developing the use of electronic resources and infrastructures, including the Internet, as an educational resource, the United Kingdom has expended considerable funds to facilitate the convergence of new learning environments with digital library services and to develop a coherent Information Environment (IE) to support higher education (Ingram & Grout, 2002). The resulting IE is both an enabling infrastructure, designed to facilitate the interoperability of heterogeneous services, and an impressive collection of online resources. While it continues to expand in size, scope, and complexity, formative evaluation has been a key part of the IE. In recent years, a number of government-sponsored projects have sought to investigate and profile the way students use
electronic information services within higher and further education. This article focuses on student Web searching behavior and reports on some of the related studies conducted at the Centre for Research in Library & Information Management (CERLIM) at the Manchester Metropolitan University and at the Centre for Studies in Advanced Learning Technologies (CSALT) at Lancaster University. The results of these studies are significant not only to the IE but also to other subject portal projects and to online library research in general.

**Survey of Existing Search Engine Use Research**

We begin our analysis with an examination of recent research on search engine use. First we analyze research on general Internet users, and then we look at the work focusing on student users. Search engine usage is difficult to measure because search engines—and the Internet in general—are not controlled environments, such as a library home page or a specific information database. As such, it has been difficult to apply the traditional model of recall and precision used in evaluating information retrieval (IR) systems to Internet search engines (SEs).

A further major limitation to search engine use research is that users are adopting different information-seeking strategies than those used in more traditional contexts (Ford, Wilson, Foster, Ellis, & Spink, 2002; Jansen, Spink, & Saracevic, 2000). Jansen also points out that the behavior of Web searchers follows the principle of least effort (Zipf, 1949). This has also been recorded by Marchionini (1992), who stated that “humans will seek the path of least cognitive resistance” (p. 156), and Griffiths (1996), who found that “increasing the cognitive burden placed on the user . . . can affect successful retrieval of information. Where an application required fewer actions from the user, greater success was achieved as there was less possibility for a user to make an error” (p. 203).

An informative review of Web searching studies by Jansen and Pooch (2001) compares the searching characteristics of Web information seekers with those of users of traditional IR systems, but their study separates out Online Public Access Catalogue (OPAC) users from general IR system users. So, for example, they found that OPAC searchers express their information needs in queries of one to two terms, while Web searchers use approximately two terms and IR searchers six to nine terms per query. Searching session length also differed, with Web searchers usually using two queries per session and typically viewing no more than ten documents from the results list, OPAC searchers using two to five queries and viewing fewer than fifty documents, and IR searchers using seven to sixteen queries and viewing ten documents per session. In addition, while 37 percent of IR searchers use Boolean operators, only 8 percent of Web searchers and 1 percent of OPAC searchers use more advanced searches.
Other observations of the average Web searcher (Spink, Wilson, Ellis, & Ford, 1998; Ellis, Ford, & Furner, 1998) point out that ineffective use may be caused by a lack of understanding of how a search engine interprets a query. Few users are aware of whether or not a search service defaults to “and” or “or” and expect a search engine to automatically discriminate between single terms and phrases. Also, devices such as relevance feedback work well if the user ranks ten or more items, when in reality users will only rank one or two items for feedback (Croft, 1995). Koll (1993) found that users provide few clues as to what they want, approaching a search with an attitude of “I’ll know it when I see it,” which creates difficulties in formulation of a query statement.

Larsen (1997) is of the opinion that Internet search systems will evolve to meet the behavior of the average Web searcher. Thus it can be seen that there has been a shift toward the introduction of search features that appear to respond to the ways in which users actually search these systems, for example, search assistance, query formulation, query modification, and navigation. The notion that improved interaction may be key to improving results is attractive in principle but not necessarily true in reality. Nick Lethaby of Verity Incorporated, paraphrased in Andrews (1996), pointed out that users do not want to interact with a system beyond entering in a few keywords.

A separate research project conducted to develop a methodology for the evaluation of Internet Search Engines from a user’s perspective (DE-VISE—Dimensions in Evaluation of Internet Search Engines) also found that interaction was little valued by users as the Interaction dimension had the weakest correlation with users’ overall rating of satisfaction, where Efficiency had the strongest correlation, followed by Effectiveness, Utility, and then Interaction (Johnson, Griffiths, & Hartley, 2001, 2003). It can thus be assumed that most users will not use advanced search features, nor enter complex queries, nor want to interact with search systems. As a consequence, systems such as search engines are now trying to automate query formulation, shifting the burden of formulating precise or extensive terminology from the user to the system.

**Student Studies**

Beyond general studies of search engine users, a number of studies have focused on the student population. Gmor and Lippold (2001) put forward a number of observations from their experiences of student searching behavior on the Web. These findings can be summarized as follows: (1) students use the Web for everything; (2) they may spend hours searching or just a few minutes; (3) searching skills vary and students will often assess themselves as being more skilled than they actually are; and (4) they will give discussion list comments the same academic weight as peer-reviewed journal articles.
Navarro-Prieto, Scaife, and Rogers (1999) sought to develop an empirically based model of Web searching in which twenty-three students were recruited from the School of Cognitive and Computer Science at the University of Sussex. Ten of these participants were computer science students and thirteen were psychology students. Their findings highlight a number of interesting points: (1) while the computer science students are more likely to be able to describe how search engines develop their databases, neither of the two groups has a clear idea of how search engines use the queries to search for information; (2) most participants considered their levels of satisfaction with the results of their search to be “good” or “OK,” and (3) most participants cannot remember their searches and tend to forget those search engines and queries that did not give any successful results.

From their research Navarro-Prieto, Scaife, and Rogers (1999) were able to identify three different general patterns of searching:

1. Top-down strategy, where participants searched in a general area and then narrowed down their search from the links provided until they found what they were looking for.

2. Bottom-up strategy, where participants looked for a specific keyword provided in their instructions and then scrolled through the results until they found the desired information. This strategy was most often used by experienced searchers.

3. Mixed strategies, where participants used both of the above in parallel. However, this last approach was only used by experienced participants.

Twidale, Nichols, Smith, and Trevor (1995), in a study that informed the development of the online journal on digital archiving, *Ariadne*, considered the role of collaborative learning during information searching. Quoting relevant literature, they identified the common searching problems as retrieving zero hits; retrieving hundreds of hits; frequent errors; little strategy variation; and locating few of the relevant records. The only specific searching issue addressed was that of “errors made in searching,” which described how simple typing errors in a sound strategy led to few hits and subsequently led to the strategy being abandoned.

More general observations revealed a number of collaborative interactions between students, which were noted as the following: (1) students will often work in groups (containing 2–4 individuals) around a single workstation, discussing ideas and planning their next actions; (2) groups work on adjacent workstations, discussing what they are doing, comparing results, and sometimes seeming to compete to find the information; (3) individuals work on adjacent workstations, occasionally leaning over to ask their neighbor for help, and (4) individuals work at separate workstations monitoring the activity of others.
Finally, a large-scale, UK-funded study, called the User Behaviour Monitoring and Evaluation Framework, was designed to investigate and profile the use of electronic information services by students within higher and further education in the UK. The framework specifically focuses on the development of a longitudinal profile of the use of electronic information services (EIS) and the development of an understanding of the triggers of and barriers to use (Banwell et al., 2004)). Within this framework, two different research projects (now completed) were created to evaluate service usage trends. The JUSTEIS project (JISC Usage Survey Trends: Trends in Electronic Information Service) surveyed trends in electronic information service usage; the JUBILEE project (JISC User Behaviour in Information Seeking: Longitudinal Evaluation of Electronic Information Services) undertook a longitudinal study of electronic information service use.

JUBILEE and JUSTEIS found that undergraduate students mainly use electronic information systems for academic purposes connected to assessment, although some leisure use was reported, and use of search engines predominated over all other types of electronic information systems. Postgraduate students undertaking a degree by research were observed to have a different pattern of use from that of postgraduate students undertaking a degree on a taught course, and overall some of the postgraduate students used JISC-negotiated services and specialist electronic information systems more than undergraduates. Use of electronic journals by both academic staff and postgraduate students was relatively infrequent. Patterns of use of electronic information systems varied among subject disciplines, and academic staff were found to exert a greater influence over undergraduate and postgraduate use of electronic information systems than library staff. In addition, friends, colleagues, and fellow students were also influential. Different models of information skills provision and support were found in the different institutions and different disciplines participating in these studies. Banwell et al. (2004) suggest that patterns of use of electronic information systems become habitual.

The EDNER and EDNER+ Studies

The search engine usage project we have been involved with since 2000 is called the Evaluation of the Distributed National Electronic Resource (EDNER) Project. Since its successful completion in 2003, we were awarded a one-year extension until July 2004, hence the additional title, EDNER+. The aim of the EDNER studies was to develop understanding of users’ searching behavior within the IE by asking them to assess a selection of IE services according to a range of defined criteria—Quality Attributes. Given the limitations of search engine research and the shift in recent years from the usage of performance indicators to measures of outcome and impact within libraries (Brophy, 2004), we have developed a Quality Attributes approach for this research.
The classic definition of quality as “fitness for a purpose” was developed by Garvin (1987) into an eight dimension, or attribute, model, which can be used as a framework for determining the overall quality of a product or service. This approach has since been adapted for use in libraries and information services by Marchand (1990), Brophy and Coulling (1996), Brophy (1998), and Griffiths and Brophy (2002). Griffiths and Brophy adapted the Quality Attributes by changing the emphasis of one attribute, changing the concept of one attribute, and introducing two additional attributes (Currency and Usability), thus producing a set of ten attributes: Performance, Conformance, Features, Reliability, Durability, Currency, Serviceability, Aesthetics, Perceived Quality, and Usability. A further discussion and presentation of results related to individual attributes is given by Griffiths (2003). The work reported here focuses on results related to discovery and location of resources, resource use, and students’ perceptions of quality.

For the first EDNER study, test searches were designed (one for each of the services to be used by the participants, fifteen in total). These searches were of sufficient complexity to challenge the user without being impossible to answer and were individually tailored for each of the services evaluated. Participants were recruited via Manchester Metropolitan University’s Student Union Job Shop; twenty-seven students from a wide course range took part in the study, and each student was paid for his or her participation. One-third of the sample consisted of students from the Department of Information and Communications who were studying for an information and library management degree, while the remaining two-thirds of the sample were studying a wide variety of subjects (being at various stages of their studies). No restrictions were placed on them having computer experience, Internet experience, or familiarity with search engines. Testing was conducted in a controlled environment based within the Department of Information and Communications. Each participant searched for the fifteen test queries and completed questionnaires for each task undertaken.

The EDNER+ study investigated student use of eighteen services, which were selected from the presentation layer of the IE. Follow-up questions related to the first EDNER study were included. Individual tasks were created for each service, questionnaires were developed and piloted, and methods of analysis were agreed upon. Thirty-eight students were recruited from thirty-four subjects across the university. These students then undertook two days of searching. None of these participants was studying for an Information and Library Management degree. Each participant used all eighteen services and provided feedback on each service via individual questionnaires. Subjects studied included art, sociology, Spanish, primary education, English, law, and computing.

Data gathered during both studies were analyzed in two ways: quantitative data were analysed using SPSS (Statistical Package for the Social
Sciences), and open-response question data were analyzed using qualitative techniques.

Results

The EDNER studies were concerned with two main questions: (1) how do students discover and locate information, and (2) how do services (and aspects of services) rate in a student evaluation, and what criteria are most important to them (results of this work are presented in Griffiths, 2003). The following section presents a selection of the results related to discovery and location of information.

Students’ Use of Search Engines Dominates Their Information-Seeking Strategy

Students were asked to find information on fifteen set tasks, designed to be typical of information seeking in an academic environment, and to complete a questionnaire after each task. Every time they started a new task we asked them where they went first to try to find relevant information. The following presents the most frequently cited starting points as found in the first EDNER study:

- 45 percent of students used Google as their first port of call when locating information
- The second most highly used starting point was the university OPAC, used by 10 percent of the sample
- Next comes Yahoo, used by 9 percent of the students as the first source they tried
- Lycos was used first by 6 percent
- AltaVista, Ask Jeeves, and BUBL were all used as a first resource by 4 percent (each) of the sample of students

Results from the EDNER+ study found that

- 22 out of 38 participants use an SE every day
- 2 use an SE three to six times a week
- 9 use an SE once or twice a week
- 2 use an SE every other week
- 3 use an SE once or twice a month

Of the search engines chosen, 23 used Google, 4 used a combination of Google and Yahoo, 3 used Yahoo, and 5 used a combination of a variety of SEs. Some students exhibited confusion regarding services, listing the library catalogue and the BBC as search engines they had used. It is clear that the majority of participants use a search engine in the first instance. This concurs with the JUBILEE and JUSTEIS results, which found that use of SEs predominates over all other types of EIS. Search engines are liked for their familiarity and because they have provided successful results on previous occasions. Individual search engines were frequently described by students as “my personal favourite,” and phrases such as “tried and
“tested,” “my usual search engine,” and “trusted” were frequently given by the students when asked why they chose this source first.

Google’s popularity was also expressed in many comments about the service, such as: “Google is very straight forward. You put in your word and it searches. It also corrects spellings to rectify your search. Bright, eye-catching—simple. Not confusing”; “Most popular search engine. I always use this for any search”; and “I find the site very helpful. It seems to have whatever I want. I’m happy with it. It is simple but complete.”

Students’ Use of Academic Resources Is Low After search engines, the most frequent starting point was the university library OPAC, followed to a lesser degree by a known academic resource. Thus BUBL, Emerald, Ingenta, and BIDS were all mentioned by participants. There was a very marked difference between information and library management students and those studying for other degrees. The former group was much more likely to prefer these academic resources as a first search tool for similar reasons as the search engine users. Comments such as “Quick and easy to find,” “used to it,” “thought it would have the relevant information,” and “I always use the University electronic journal search first” were typical among these students. Again, ease of use, familiarity, and reliability were key factors in their choice.

Information and library management students used the library OPAC to provide details of, and access to, journals. As might be expected, they knew that they would find such information there. They expected to use “bibliographic databases across different subject disciplines,” and they also more frequently sought out access to sites with “academic information as opposed to commercial.” Some displayed quite a detailed knowledge of the resources available through the library Web site. One student searching for a parenting article “assumed PsychInfo would have an abstract of the article and you can search by either author or keyword”; another, wanting a source on using questionnaires to collect data, “thought there might be something on methodology in the statistics section.”

The information management students used the library home page to find a route to subject categories too. For example, one user seeking information on wildlife tours looked for an “organisation on safaris” via the Tourism link; another chose the Biology link as a possible, though unsuccessful, route to an image of the brain. A third “thought the library home page would have a section for science in general” (it does not) from where one can look for a link to the NASA Web site and then tried the Web of Science before resorting to Google.

This group of students also made more frequent use of services such as BUBL, Emerald, Ingenta, and BIDS. One described BUBL as a “known academic resource with selected/quality sites of interest to academic disciplines” and used it to answer a question on early dynastic Egypt. In con-
trast, another user made the point that, when searching for an article, it was easier to try Google first—“quite good at finding articles”—because otherwise there would be a need to “look at a few different databases, e.g., Emerald, BUBL etc.” Only two students mentioned using the Resource Discovery Network (RDN) (although in no instance was it the first action taken): one seeking information on research methodology, and another as a possible route to an image of the brain, though this was unsuccessful and the student resorted to Yahoo.

The library OPAC was also used by non–information management students to locate information, though to a much lesser degree, and always when looking for journals or articles. “I thought the library pages listed all articles” and “Thought it (an article) was most likely to be in the library catalogue” were two reasons given. One student used the library OPAC because “I knew that the University holds a large source of electronic journals.” However, this action was taken only when a search engine search had failed. Another user searching for an article on “parenting” resorted to the library OPAC because part of it “is medically based so I thought it would be the best place to look.” Other comments indicate some confusion amongst students about the OPAC, describing it as “A search engine for the library, to find books and catalogues” and “With this search engine . . . it is easy and straight forward to use.” It seems that students’ use of resources is now very colored by their experience with search engines, which in turn may lead to expectations that may not be realistic for different types of services.

Among all users, the library OPAC was chosen for its familiarity, its ease of use, its ability to retrieve relevant information, and mostly because there was a clear expectation among some participants that certain types of information resources would be found there. The fact that the most frequent users were information management students might suggest that, when lecturers are aware of and train their students to use the resources that the library provides, their students will become familiar with them and will use them. If this is not done, the status quo approach seems to be resorting to a search engine, with varying degrees of success.

Levels of use of the library OPAC recorded by the EDNER+ study showed that
- 4 out of 38 participants had never used the library OPAC
- 4 only use it occasionally
- 10 use it once or twice a month
- 3 use it every other week
- 10 use it once or twice a week
- 1 uses it 3 to 6 times a week and
- 5 use it every day

One participant failed to report his/her level of use.
Bibliographic database use was recorded as follows:

- 21 out of 38 participants never use bibliographic databases
- 3 use them occasionally
- 6 use them once or twice a month
- 3 use them every other week
- 4 use them once or twice a week and
- 1 student reported that he/she uses them three to six times a week

Of the students who do use bibliographic databases, 3 stated that they use Web of Science, 3 stated that they use Emerald, and 2 listed FAME. All other bibliographic databases were only listed by one participant each: these included SOSIG, Ingenta, Butterworths, Lexis Nexis, and Questia Social Science Library.

Use of Amazon.com for locating information, especially about videos, proved to be popular. Four users looking for the Manchester distributor of an Albert Einstein video went immediately to Amazon to seek this information because “Amazon is a global source for videos” that “sometimes has distribution details and other possible names for the video.”

Perceptions of Use, Success, and Why Students Stop Searching

When participants were asked how easy it was to locate information, the following responses were recorded:

- 50 percent found it easy to locate the required information
- 35 percent found it difficult
- 15 percent had no view either way

Participants’ reasons for finding tasks easy included: “Easy enough to find using the search engines”; “Easier to find formal institutions because they usually have a Web site and these are more often than not listed as recommended sites on the library home page via corresponding subject pages”; “Very easy and direct search taking a small amount of time”; and “It was easy once I went back to Google. Ingenta just messed me about.”

Where participants found a task difficult, the following comments were made: “Why doesn’t someone make a good search engine devoted to articles? It’s hard to find an article without an author”, “It is very difficult to search for something specific”; “It was easy to find an abstract, I just couldn’t find the full article,” and “Got disheartened.”

When participants were asked to locate a Web site to find specific information, 70 percent responded that they were successful, and 30 percent responded that they were unsuccessful. When asked to find information via a specific service, 74 percent responded that they were successful, and 17 percent were unsuccessful (9 percent did not know). Check questions were included to ensure that participants were not overgenerous in their reports of success.

From these results it is clear that, even when users can find information,
it is not always an easy task. This may have serious implications for developers of services as a number of studies (Griffiths, 1996; Johnson, Griffiths, & Hartley, 2001) have shown that users will often trade performance for the path of least cognitive resistance (minimum effort and time).

Students were asked to search for as long (or short) a time as they wanted provided that they spent no longer than 30 minutes on any one service. This upper limit was imposed as a result of other research (Craven & Griffiths, 2002), which found that the average time taken to search for information is between 15 and 19 minutes. The majority of students in this study spent an average of between 1 and 15 minutes searching for information. The DEvISE project (Johnson, Griffiths, & Hartley, 2001) also found that Efficiency correlated most strongly with General Satisfaction, with Effectiveness second, which may suggest that the amount of time and effort required from the user matters more than the relevance of the items found.

Students were also asked why they stopped trying to locate information, with the following reasons given:

- Found information = 70 percent
- Unable to find Web site within time allowed = 15 percent
- Could not find a Web site and gave up = 12 percent
- Technical problems affected search = 3 percent

Participants who were unable to find a Web site within the time allowed usually stated that they had run out of time. Among those who “Couldn’t find a Web site and gave up,” frustration at being unable to complete the task was expressed. “It is frustrating when you can’t find what you are looking for” or “frustration; all sites were irrelevant” were typical remarks. The lack of success was described as “hitting a brick wall” or not “getting anywhere.” Some admitted that they simply did not have any further search strategies, saying they “Don’t know where else to search for it,” “I have searched everywhere I can think of,” or “didn’t know where else to go.”

This frustration was also reflected in some of the comments of those who encountered “Technical problems.” These problems were usually expressed as “slowness.” “Internet was very slow” was the most usual comment. “Taking ages to get to some sites,” “Server could not contact host and very slow for pages to show,” or “Pages would not open” were other complaints. One respondent remarked that he/she “decided to stop, as if I was doing a search for myself I would not have spent that much time.” It may be frustrating for the developers of resources to accept that speed of access may be a criterion on which users will evaluate a service, but studies have shown that this is an important indicator for some users (Johnson, Griffiths, & Hartley, 2001). One respondent gave a very simple reason for stopping—“Teatime!”

**Student Perceptions of Quality** One of the main aims of the IE is to provide a managed quality resource for staff and students in higher and further
education. During discussions with various stakeholders involved with the development of the IE, it became clear that common definitions of what is meant by quality electronic resources could not be assumed. Therefore, participants were asked during testing to indicate what quality meant to them in terms of information available via electronic services (they were not asked to relate their responses to any one particular service). Four criteria were presented to them with which they could either agree or disagree. Participants were also asked to add any additional criteria that were not listed but were important to them. Table 1 presents their responses.

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Reliable</th>
<th>Current</th>
<th>Accurate</th>
<th>Refereed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>52%</td>
<td>81%</td>
<td>89%</td>
<td>26%</td>
</tr>
<tr>
<td>No</td>
<td>48%</td>
<td>19%</td>
<td>11%</td>
<td>74%</td>
</tr>
</tbody>
</table>

Additional criteria listed by students included (1) links to related areas; (2) understanding language used; (3) resources relevant; (4) speed of response; (5) resources useful; (6) resources valuable; (7) clear information; (8) source; (9) accessible; (10) timeliness; (11) presentation; and (12) references.

**Discussion and Conclusions**

These results raise a number of important and interesting issues:

- Students prefer to locate information or resources via a search engine above all other options, and Google is the search engine of choice.
- Students’ use of academic resources is low.
- Students find it difficult to locate information and resources.
- Students may trade quality of results for effort and time spent searching.
- Students’ use of SEs now influences their perception and expectations of other electronic resources.

Students either have little awareness of alternative ways of finding information to the search engine route or have tried other methods and still prefer to use Google—a situation we now refer to as the “Googling phenomenon.” Beyond this, even when students are able to locate information it is not always easy (even when using Google), and with a third of participants failing to find information, user awareness, training, and education need to be improved. While 70 percent of participants felt that they were successful in locating a Web site that provided the required information, only half of these thought that it was easy to locate information. Only the information management students frequently used library resources first to locate
information, though some also used search engines either as a first resort or as a backup. And while there were some indications that other students knew about and used library resources, their use was much less common.

Students prefer particular favorite search engines, though the reasons they give for their preferences are common across all search engines. Some users indicated that, if their first search strategy did not work, they had a string of similar resources to use as a backup: “Lycos, then Google, Yahoo, AltaVista,” “Google, AltaVista, Excite, Northern Light,” or “Goto.com, Yahoo, Lycos” were typical comments. Others tried a search engine first, and if this failed they turned to a different kind of search tool: “Google then Ingenta,” “Google, Biology section of library home page, RDN, Yahoo,” for example.

One reason for some of the problems that students experience when using electronic resources may be that the hierarchical arrangement of current IE subject gateways is confusing to them. Hierarchies are notoriously difficult to navigate horizontally so that, once down a particular branch, students may be unable to navigate successfully to an “unrelated” branch. They are effectively lost. Secondly, without a firm grasp of the overall “shape” of the subject, they may find it difficult to identify the correct branches to follow. It would be remarkable if students in the early years of higher education did have a clear conceptual map of their discipline—this is one of the things they are learning. Thirdly, there may be subject-specific factors at work: for example, the structure of chemistry as a discipline may be easier to follow than that of, say, social science.

Many networked information services provide information to the desktop, which is quality assured in some way, for example, by the institution’s library, by academic publishers, and by the development of the IE itself. As such, users do not always have to concern themselves with exhaustive searches encompassing many resources in order to be satisfied that they have the best information they can get. Indeed, this may be said to be the age of information satisfying—when something is good enough for the purpose rather than seeking to optimize the result (Simon, 1957). Recent studies of the use of electronic resources found that, when users seek information, almost all users will only look at the first page of results (for example, Craven & Griffiths, 2002; Sullivan 1998, 2002). Most users are satisfied that these initial ten or so results are good enough to answer their information need. Users are rarely interested in a comprehensive, high-recall search but rather are satisfied with the retrieval of a few relevant hits.

In addition our research indicates that students are confused as to the meaning of quality when it comes to assessing academic resources. Viewed in the light of the findings of Cmor and Lippold (2001), who stated that students will give the same academic weight to discussion list comments as peer-reviewed journal articles, it would seem that students are poor evalu-
ators of the quality of academic online resources. The original premise of the Perceived Quality attribute is that users make their judgments about a service based on incomplete information and that they will come to this judgment based on its reputation among their colleagues and acquaintances and their preconceptions and instant reactions to it. If the notion of quality conveys so many different meanings to students, it poses something of a challenge to the academic community in encouraging students to understand and use quality-assured electronic resources. It is also apparent that, from a methodological perspective, further work is needed to explore the meaning of Perceived Quality and the interpretation of user responses to this area of enquiry. Fundamentally different understandings of information quality could otherwise lead to questionable conclusions being drawn by researchers and service providers.

Students’ use of SEs now influences their perception and expectations of other electronic resources. While the preference for very simple search engine approaches is prevalent, it is important to note that this does not mean that students are necessarily best served by this approach. Indeed, it may be that students would get better results using specialist subject gateways, but most students do not take this approach. Exclusive use of any commercial SE coupled with a lack of awareness and understanding of peer-reviewed, quality resources is not in the best interest of students or academic staff. As service providers and developers, it is crucial that we learn lessons from those commercial search engines that dominate students’ use and embed those lessons into academic resources that students can find and use easily.

Notes
1. More than twenty years ago the UK began its strong commitment toward harnessing the Internet for higher education. First, the universities’ governmental funding bodies established the Joint Information Systems Committee (JISC) to deal with networking and specialist information services and to advise and support the UK’s higher and further education institutions. During the 1990s the JISC was given the remit to explore and facilitate the convergence of new learning environments with digital library services and to develop a coherent Information Environment to support learning, teaching, and research. Initially known as the Distributed National Electronic Resource (DNER), the JISC Information Environment now provides both infrastructure and services to support the whole of the UK postschool academic sector.
2. The UK’s Resource Discovery Network (RDN) was initially called the Distributed National Electronic Resource (DNER).

References


Cyber-Democracy or Cyber-Hegemony?
Exploring the Political and Economic Structures of the Internet as an Alternative Source of Information

JULIE FRECHETTE

ABSTRACT
Although government regulation of the Internet has been decried as undercutting free speech, the control of Internet content through capitalist gateways—namely, profit-driven software companies—has gone largely uncriticized. The author argues that this discursive trend manufactures consent through a hegemonic force neglecting to confront the invasion of online advertising or marketing strategies directed at children. This study suggests that “inappropriate content” (that is, nudity, pornography, obscenities) constitutes a cultural currency through which concerns and responses to the Internet have been articulated within the mainstream. By examining the rhetorical and financial investments of the telecommunications business sector, the author contends that the rhetorical elements creating “cyber-safety” concerns within the mainstream attempt to reach the consent of parents and educators by asking them to see some Internet content as value laden (sexuality, trigger words, or adult content), while disguising the interests and authority of profitable computer software and hardware industries (advertising and marketing). Although most online “safety measures” neglect to confront the emerging invasion of advertising/marketing directed at children and youth, the author argues that media literacy in cyberspace demands such scrutiny. Unlike measures to block or filter online information, students need an empowerment approach that will enable them to analyze, evaluate, and judge the information they receive.

According to figures provided by the U.S. Census Bureau (2001), more than half of school-age children (6 to 17 years) had access to computers
both in school and at home in the year 2000 (57 percent). With some 17 million children using the Internet in some capacity, including email, the Web, chat rooms, and instant messaging (Silver and Garland, 2004, p. 158), the Census Bureau estimates that 21 percent use the Internet to perform school-related tasks, such as research for assignments or taking courses online.

While these statistics underscore the growth and popularity of the Internet, particularly in schools and educational institutions, concerns have grown about the “safety” of using computer-mediated communication technology. Since the Internet became a mass medium in 1995, parents and schools have approached online content with reservation. As such, politicians, educators, child advocacy groups, and, most importantly, the computer industry, have been vocal advocates for patrolling the Internet and censoring certain kinds of illicit or objectionable content. Beginning in the late 1990s, Federal Trade Commission member Christine Varney summarized the emerging concerns about online safety:

All of us agree that children’s online safety concerns are real and pressing and that we must support the involvement of parents raising children in this new, digital age. We understand that we must all work together—industry, law enforcement, educators, advocates—if American families are to realize the potential of this new medium for enriching the lives of our children and fostering their future success. (Rubin and Lamb, 1997)

Starting in 1997, an Internet/Online Summit was held in Washington, D.C., to enhance the safety and benefits of cyberspace for children and families. Key political figures, such as former vice president Al Gore and former attorney general Janet Reno, joined parents, as well as politicians, law enforcement officials, and educational administrators, to launch a national public education campaign, “America Links Up: An Internet Teach-In,” designed to help Americans understand how to guide kids online (Rubin & Lamb, 1997).

On October 21, 1998, former president Bill Clinton signed into law the “Children’s Online Privacy Protection Act” (COPPA). This measure was enacted by Congress on April 21, 2000, to “prohibit unfair or deceptive acts or practices in connection with the collection, use, or disclosure of personally identifiable information from and about children on the Internet” under the age of thirteen (Grossman, 2000). Along this trajectory, Congress passed the Children’s Internet Protection Act (CIPA) and the Neighborhood Internet Protection Act (NCIPA) in December 2000, which required schools and libraries that receive federal money for Internet connections to adopt Internet safety policies in 2001. The proposed safety measures include usage agreements for proper student use of this medium, audit-tracking devices to supervise student Internet perusal, and software
filtration devices designed to block inappropriate sites in schools (Trotter, 2001).

In 2002 the Bush administration proposed a “National Strategy to Secure Cyber Space,” offering security recommendations for U.S. citizens, businesses, and organizations using computers (Carlson, 2002). Since then the Federal Trade Commission has offered testimony before special committees and the House of Representatives about online pornography through a series of “law enforcement actions against fraud artists whose deceptive or unfair practices involve exposing consumers, including children, to unwanted pornography on the Internet” (Federal Trade Commission, 2004, p. 1).

In addition to these federal initiatives, many states have measures designed to protect children from online predators. In Texas, Attorney General Greg Abbott added more investigators to the Texas Internet Bureau to keep kids safe from those who use online means to prey on children. As Assistant Attorney General Sparks explained, “The Attorney General wants the public to know that he’s tasking people with patrolling the Internet and trying to make it safe for kids; the down side is that more and more children on a daily basis are getting online and on the Internet and as every additional child gets on, that’s one more potential target” (quoted in Ochoa, 2003).

Likewise, educators have expressed concerns about online information overload. According to one school administrator, accessing the Internet in schools is less predictable: “If you used to bring your class to the school library, you pretty much had a sense of what was available for the children to research; now you have no idea . . . they are going to hit sites that are appropriate and sites that are inappropriate” (quoted in Shyles, 2003, p. 176).

Despite a commitment to online “security” in schools, libraries, and homes from so many constituents, few recommendations have materialized into solid strategies or funding initiatives. Almost all of the proposed solutions and policies ignore the more relevant question of how private computer companies, Internet service providers, corporations, and governments stand to gain financially and politically by deciding what kind of information will be “censored” and what kind will be promoted. In fact, it could be argued that the Internet content “crisis” dominating public policy and mainstream media coverage has produced a cultural climate ripe for the commercial exploitation of parents and educators. In this article I argue that such a discursive trend manufactures consent through a hegemonic force that overlooks the invasion of advertising or marketing strategies targeted at young people online. By examining the rhetorical and financial investments of the telecommunications business sector, I contend that the mainstream articulation of “Internet safety” invites parents and educators to regard some Internet content as value-laden (sexuality, obscene language),
while disguising the interests and authority of profit-minded commercial enterprise (advertising and marketing).

What is more, the democratic potential of the Internet as a means to accessing alternative information and perspectives otherwise absent from the mainstream media continues to be threatened by the consolidation of increasingly powerful global media giants, such as Time Warner and Microsoft, which have much to gain from controlling the content Internet users access at home or at school. Consequently, an examination of the political and economic forces on the Internet is necessary for librarians and educators interested in understanding the benefits and limits of the Internet as a means of alternative communication.

**Exploring the Means to Filtering Online Content**

*Parental Guidance*

As a result of this discourse, a number of solutions have been advanced to ward off illicit content appearing on the computer screens of young Internet users, beginning with parental guidance. CyberTipLine grew out of the 1997 Internet/Online Summit and is currently in operation today. Run by the U.S. government and the National Center for Missing and Exploited Children, parents can notify authorities of incidents of online child pornography and child predation. Another derivative of the summit’s “America Links Up” project is the industry-sponsored “GetNetWise” Web site, which was launched in 1999. The “user empowerment” service, which involves a coalition of numerous Internet industry partners and advocacy organizations, offers parental advice, including information about filters to block sexually explicit material, as well as a variety of tools to help parents and caregivers monitor a child’s online activities and find browsers for kid-friendly sites. As one sponsor, AT&T, notes in its promotional material, “Our involvement with GetNetWise reflects our commitment to help users have the best possible online experience” (GetNetWise, 2004).

A more well-known parental guidance initiative, passed in April 2000, was the Children’s Online Privacy Protection Act (COPPA). In accordance with COPPA, the Federal Bureau of Investigation offers “A Parent’s Guide to Internet Safety,” which advises parents to “utilize parental controls provided by your service provider and/or blocking software” and “Monitor your child’s access to all types of live electronic communications (chat rooms, instant messages, Internet Relay Chat, etc.), and monitor your child’s e-mail” (Federal Bureau of Investigation, 2004).

Other parental guidance measures have been created to address online advertising and marketing as well as issues of privacy. Parent advocacy groups, such as Commercial Alert, Consumer Action, the Center for Media Education, and Computer Professionals for Social Responsibility, have taken up the cause of parents concerned about online marketing measures
targeted at children. For example, Commercial Alert has made requests to the Federal Communications Commission and the Federal Trade Commission to require disclosure of embedded advertising in a variety of media and has created a “Parent’s Bill of Rights” seeking to empower parents in the face of an aggressive commercial culture (Commercial Alert, 2003).

Proof-of-Age/Shielding Systems

In addition to parental guidance, many online providers and Webmasters have adopted proof-of-age/shielding systems that use credit card access as another means of content filtering. While COPPA sought to protect children thirteen and under, those located in the fourteen to eighteen year range were not covered by legislation. Providing proof of age before being allowed to access the content of a desired online site emerged as a means to address this gap. This system works in the same way that fraud-screening technology works: merchants collect user information at their Web sites for instant age or identity verification. Once online users submit their name, zip code, date of birth, and age, they are checked through an international electronic database of government-issued identifications. This allows site providers or merchants to determine the consumer’s identity within seconds. Sometimes additional measures, such as online name signature, are required so that user signatures are bound to a public record.

Proprietary Environments

Another reaction to the discourse of online safety has been the advocacy of proprietary environments, where content is screened by editors into specific categories. For example, the leading Internet service provider, America Online (AOL), provides a blocking service that allows users (ostensibly parents) to limit a child’s selected screen name to either a “Kids Only” area, which is recommended for children under twelve, or to a pre-teen/teen environment, with restricted use of chat rooms or newsgroups. According to the site, “Kids Only” is a collection of educational resources and entertainment areas as well as a preselected collection of child-oriented Internet sites, with AOL staff monitoring of message boards and chat rooms. AOL also promotes the company’s “Parental Phone Line” for instructions and advice on choosing and maintaining the settings of this product (the premise here is that the settings are likely to be tampered with by savvy teens and preteens).

In addition to “Kids Only,” AOL has aggressively marketed its AOL@School service, which had been adopted by more than 14,000 schools by 2004 (Williams, 2003). AOL@School offers six online learning portals for grades K–5, middle school, and high school so that students can access Web sites that have been preselected by educators as content and age appropriate. The software needed to access the portals comes with AOL’s “parental controls” designed to “help ensure a safe, secure, age-appropriate experience” that can include school-controlled email, chat, and instant
messaging (AOL, 2004). The popularity of “child safe” proprietary environments has not waned as Web browsers and popular search engines have created their own directories in an attempt to create safe havens for (and develop customer loyalty from) younger online users. YahooOligans’ “Web Guide for Kids” is a collection of predominantly commercial links to online games, music, TV, science, news, jokes, “cool pages,” arts and entertainment, and sports. Like most commercial proprietary environments, YahooOligans is riddled with advertisements and synergistic ties to commercial media products.

Internet Ratings Systems

For those seeking additional regulatory measures, Internet rating systems offer another approach. Unlike the rating system for television content that is uniformly and centrally organized by the television industry, Internet ratings are not assigned consistently by a centralized group of online content providers. The goal is the same, however: industry self-regulation over government regulation. According to ratings system advocates, many of whom work in the software and computer industry, Internet ratings are designed to make it “safe” for schools and parents to let their children access nonpornographic material without government directives. According to Paul Resnick, chairman of the World Wide Web Consortium group at the MIT Laboratory for Computer Science, which includes AT&T Laboratories and Microsoft, the Platform for Internet Content Selection (PICS) was originally created to allow parents, teachers, and librarians to review questionable materials that they would not want their children to come across on the Internet (Resnick, 1997).

Resnick explains, “prior to PICS there was no standard format for labels, so companies that wished to provide access control had to both develop the software and provide the labels. PICS provides a common format for labels, so that any PICS-compliant selection software can process any PICS-compliant label” (Resnick, 1997, p. 107). Yet unlike uniform rating labels, a single site or document may have many labels, provided by different organizations. Consumers choose their selection software and their label sources (called rating services) independently. This separation allows both markets to flourish: companies that prefer to remain value-neutral can offer selection software without providing any labels; values-oriented organizations, without writing software, can create rating services that provide labels. (Resnick, 1997, p. 107)

One of the leading Internet rating systems that uses PICS is SafeSurf, a group that offers ratings along with other tools to help parents and “net citizens” filter online information. One means to achieving its goal is to encourage online content providers to fill out a questionnaire using content descriptors to rate their Web sites. Unlike government- or industry-wide regulatory labeling efforts that may “brand” content, SafeSurf is interested in
maintaining First Amendment rights by offering content providers greater latitude to self-rate their Web material. For example, rather than branding content that includes nudity as pornographic, users can distinguish their inclusion of nudity as scientific, sociocultural, artistic, titillating, graphic, or illegal. Once content providers rate their Web sites or directories, they can download the SafeSurf rated logo of their choice. A SafeSurf staff member verifies the rating and sets up the chosen ratings label. Parents and educators can then use PICS compliant software/browsers to read the settings and to use the ratings to filter content that is not desired. As the SafeSurf group explains, “PICS allows content providers to rate their pages and parents to set passwords and levels for their children. Then, PICS compliant software/browsers will read the settings and use the ratings to filter content that is not desired” (SafeSurf, 2004a).

The Internet Content Rating Association (ICRA) is another international, independent, nonprofit organization that seeks to “empower the public, especially parents, to make informed decisions about electronic media by means of the open and objective labeling of content” (ICRA, 2004). ICRA’s dual aims are to “protect children from potentially harmful material and to protect free speech on the internet.” Like SafeSurf, Web authors complete an online questionnaire describing the content of their site, upon which ICRA generates a content label using PICS computer coding, which the author adds to his/her site. Parents and Internet users can then set their Internet browser to accept or decline access to Web sites based on the labels and user preferences. PICS is now a standard feature included in Internet software and browsers such as Microsoft Explorer.

Third-Party Rating Systems

While ratings systems are designed to allow content providers to voluntarily label the content they create and distribute, third-party rating systems “enable multiple, independent labeling services to associate additional labels with content created and distributed by others. Services may devise their own labeling systems, and the same content may receive different labels from different services” (ICRA, 2004). In other words, online watchdog groups interested in protecting children from online predators or illicit material can offer their own set of restrictive control tools for material that they deem to be objectionable. One such group is WiredSafety, formerly known as CyberAngels, led by Parry Aftab, an experienced international attorney and author of The Parent’s Guide to Protecting Your Children In Cyberspace and A Parent’s Guide to the Internet. Lauded as “one of Internet safety’s most influential players,” (Hill, 2000), Aftab has emerged as a nonprofit leader who has created coalitions with many governmental and nongovernmental agencies, including the FBI’s Innocent Images anti–child pornography and exploitation task force. She was appointed the founding American director of UNESCO’s global Child Safeline project and currently heads
WiredSafety, “the largest online safety, education and help group in the world” (WiredSafety, 2004). With more than 9,000 volunteers worldwide, the group is a coalition of various Internet safety groups, such as Wired-Kids.org, WiredTeens, Teenangels, and CyberMoms and CyberDads, and their affiliate, WiredCops.org, all of whom patrol the Internet for child pornography, child molesters, and cyberstalkers. Additionally, WiredSafety offers a variety of educational and help services for online users. Some of its volunteers access and review family friendly Web sites, filter software products and Internet services, and post their findings on the Web. The group even has a “Cyber911 help line” that offers net users access to help when they need it online. SurfWatch is another online ratings system designed for parental supervision. It too prevents access to Web, gopher, and FTP sites that SurfWatch’s team of “net-surfers” have found objectionable. They maintain an updated list of “not-for-children” Web sites that can be subscribed to electronically.

Commercial Filtering Software and Databases

A more intensive effort to censor “inappropriate” online content has come from commercial filtering software companies (often working in conjunction with powerful Internet content providers and third-party ratings systems). Also known as “censorware,” these filtering products, which include Net Nanny, CyberPatrol, Cyber Sitter and N2H2, range in cost from $25.99 to $80 and are heavily marketed to parents, educational administrators, and libraries. Designed to be installed on home or school computers or to work with network routers or firewall, cache, or proxy devices, these products claim to offer safety measures for youth using computers for online research and recreation. Essentially, most of these programs work by using a combination of filtering and blocking strategies, such as the blocking of Web sites denoted through keywords and databases and the blocking of individual Web sites by specific URLs.

One of the first filtering programs—and most commercially lucrative—is Net Nanny. According to its promotional Web site, Net Nanny® 5 is “the world’s leading parental control software, [and] provides customers with the broadest set of Internet safety tools available today. Our award-winning software gives customers control over what comes into and goes out of their home through their Internet connection, while respecting their personal values and beliefs” (Net Nanny, 2004). Launched in 1998, Net Nanny is a tool allowing parents, teachers, administrators, and librarians to screen incoming and outgoing Internet information, particularly pornographic material. By identifying and blocking various sites and subjects considered inappropriate, the program blocks the Web addresses of known pornographic and illicit sites. Parents can add to the collection of forbidden “code words” used to detect and flag sites. The program works with
all major online providers and in email. It can also prevent children from accessing specific files on a PC’s hard drive, floppy drive, or CD-ROM. Like audit-tracking software programs, Net Nanny keeps a record of a child or student’s Internet perusal, meaning that parents and teachers can check up on the sites that a child has perused.

With all of these features, it is no surprise that Net Nanny’s popularity and financial success has led it to offer additional blocking software such as Net Nanny’s Pop-Up Scrubber, which blocks pop-up ads, Net Nanny’s Ad-Free, which blocks a range of Internet ads, spyware, and profiling cookies, and Net Nanny’s Chat Monitor, which monitors and filters Instant Messaging and other online chat.

Another commercial service, CyberPatrol, works in the same way as Net Nanny by filtering harmful Web sites, newsgroups, and Web-based email. Also commercially successful, CyberPatrol licenses its “CyberLIST” database of site ratings to several additional vendors. Among its ratings categories are violence/profanity, partial nudity, full nudity, sexual acts, gross depictions, intolerance, satanic or cult, drugs and drug culture, militant/extremist, sex education, questionable/illegal and gambling, and alcohol and tobacco. Likewise, Cybersitter blocks sites and subjects deemed unacceptable by Internet users. It offers site lists for automatic blocking and allows parents to have added input in restricting programs, files, and games. According to PC Magazine, Cybersitter offers the strongest filtering and monitoring features, blocking content related to violence, hate, sex, and drugs (Munro, 2004). It also allows parents to choose from thirty-two content categories, such as free email sites, file sharing, wrestling, cults, and gambling, for those interested in added blocking categories. As with other similar products, it lets parents filter and monitor their children’s activities without their knowledge and can record both sides of Instant Messaging sessions.

Joining in the mix of filtering software providers is N2H2 (acquired by Secure Computing in 2003), a company endorsed by eTesting Labs and the Kaiser Foundation as “the most effective and accurate” filtering program and extensive database of objectionable Internet sites (N2H2, 2004). It offers two product lines: Sentian, which is geared toward helping businesses manage their employee Internet access, and Bess, a popular program and database adopted by many schools and endorsed by the American Library Association to help schools and libraries meet CIPA rules for young Internet users.

With so many companies vying to be the best provider of filtering software, it is not surprising that Microsoft would venture into this area by offering its own industry standard Internet filter aimed at regulating youth-directed online content. As part of its monopoly on the Internet browser software Internet Explorer (which accompanies its Windows platform), Microsoft has also implemented a filtering system that can be configured
to block or log all data transfers, including World Wide Web pages, newsgroups, types of messages within any newsgroup, Internet Relay Chat, or Internet hosts known to have objectionable material for children.

**Questioning the Viability of Online “Safety” Initiatives**

Although some of these Internet resources and restrictions make sense for certain schools depending upon the age group and grade level of Internet users, there are some problematic areas within each method that should be cause for concern. The main underlying difficulty raised by these “quasi-solutions” is that they narrowly define what is “inappropriate,” relegating most objections to issues of nudity, sexuality, trigger words, or adult content. This focus neglects to confront the invasion of advertising or marketing strategies directed at children. In many respects, Internet commercialism seems to be a more serious concern, but one would never guess this considering the ad-strewn and content-compromised “solutions” to appropriate Internet content.

First, although child-directed advertising might not be as blatantly offensive, it certainly fosters “values” that, at present, are not considered objectionable to most governmental, parental, and commercial watchdog groups. Although the first tenet of media literacy explains that *all media are constructions*, the problem with advertising and marketing strategies is that they are so much a part of our social landscape and our everyday life that they appear to be natural. Subsequently, the conceptualization of what is inappropriate for children or students only helps to sustain the interests of a commercial system through the omission of advertising; advertising is omitted and thereby deemed appropriate. Just as parents, educators, and anticommercial groups, such as Commercial Alert, have protested the commercial imperatives of satellite-delivered school programs such as Channel One, a company that offers schools free satellite equipment in exchange for a captive audience of students forced to watch its daily, advertisement-driven programming, and the computer equivalent ZapMe!, which tried to turn “the schools and the compulsory schooling laws into a means of gaining access to a captive audience of children in order to extract market research from them and to advertise to them” (Commercial Alert, 2000), we need to be equally circumspect about the amount of advertising and marketing proliferating on “Kids Only” sites and via kid-safe filtering software (Schiffman, 2000).

Moreover, sustaining an Internet-based market economy whereby consumer software programs and proprietary environments become the antidote to inappropriate material is directly at odds with democratic means of dealing with these issues through public discourse, political action, and critical media literacy skills. Most of the products previously analyzed are produced and distributed by profit-making and publicly traded enterprises,
such as the media conglomerates Time Warner, Microsoft, and Yahoo!. Obviously, it is good business to create and sell blocking software products or to offer third-party rating systems that decide—for parents, educators, and librarians—what is in their (both children/students and the company’s) best interest. In a self-fulfilling business transaction, reports of inappropriate content as well as media and political hype about the Internet as an “unsafe environment” lend credence to, or create a functionalist need for, such products. As stated earlier, advertising is overlooked as “inappropriate content” because it is part of everyday consumer culture, unlike pornographic and hate sites, which exist beyond the boundaries of what is deemed “good” for children and teenagers. As Marxist philosopher Antonio Gramsci (1971) has noted, hegemony works within the terrain of everyday life and requires the consent of audiences—or in this case, parents, educators, and librarians. Hence, the commonly employed rhetorical elements that create paranoia about Internet content within the mainstream attempt to reach the consent of parents and educators by inviting them to see some Internet content as value-laden or problematic while camouflaging the interests and authority of a profitable computer software and hardware industry.

Although serious discussion about government regulation goes beyond the purviews of this study, several concerns must be raised regarding commercial software programs. First, the decision to block some sites over others is a very subjective decision. The problem with this kind of regulation is that some groups and individuals might attempt to censor material (under the guise of concerns for “safety”) that threaten their own political and/or religious agenda. Dependence upon commercial Internet service providers and related filtering products limits the democratic principle of the free flow of information and puts commercial enterprise at the helm of online navigation, a troubling fact given that corporate culture can often be extremely conservative and self-serving when it comes to making censorship decisions. In one instance, America Online was charged with using filters to block out several Web sites associated with “liberal” political organizations. One of the top stories featured in *Censored 2001* was AOL’s liberal blacklist, whereby sites for the Democratic National Committee, Ralph Nader’s Green Party, Ross Perot’s Reform Party, the Coalition to Stop Gun Violence, and Safer Guns Now were labeled as “not appropriate for children” (Phillips & Project Censored, 2001, p. 111). Ironically, the youth filters did not prevent access to nudity or to conservative groups, including the National Rifle Association. Designed for America Online by the Learning Company, an educational software company owned by Mattel, such filtering programs confirm suspicions about the process of labeling and omitting Web sites according to political and economic interests.

This kind of censorship raises flags about the capabilities of large media conglomerates to limit access to material deemed politically at odds with commercial interests. Inasmuch as Disney was in a position to rebuke the
distribution of *Fahrenheit 9/11*, Michael Moore’s political documentary produced through Disney’s Miramax film division, large multimedia conglomerates are poised to censor content that is politically or economically damaging to their enterprise.

Second, some of the trigger words used to block Internet sites might be legitimate subjects for research. For example, the often-cited example of an Internet user not being able to access research on breast cancer or sex education (if these words were denoted as trigger words) is indeed troubling. As *PC Magazine* reviewers of Cybersitter 9.0 explain, “Cybersitter errs on the conservative side; by default it may block sites you would deem okay” (Munro, 2004). A telling example of this problem is offered in an article featured in *Electronic School Online*. Author Lars Kongshem writes,

> CYBERSitter yanks offending words from web pages without providing a clue to the reader that the text has been altered. The mangled text that results from this intervention might change the meaning and intent of a sentence dramatically. For example, because “homosexual” is in the list of CYBERSitter’s forbidden words, the sentence, “The Catholic church is opposed to all homosexual marriages” appears to the user as, “The Catholic church is opposed to all marriages.” (Kongshem, 1998)

Likewise, Karen Schneider, a librarian for the Environmental Protection Agency, has led a filtering software assessment project involving more than thirty librarians around the world. She has found that filters “are not reliable and they’re hard to maintain” (cited in Gebeloff, 1999). In one example, recipes using “chicken breast” were blocked due to sensitive word triggers. Rob Gebeloff, author of *Screening Zone: The Trouble with Net Filters and Ratings*, continues to problematize the use of all types of “censorware” programs by pointing out numerous gray areas in judging content. He asks:

> Do you want your kids going to Web sites that discuss birth control? What about AIDS education? Or what about the exploration of Mars? [A recent *New York Times* article pointed out that one filtering program blocked out every Web site with the word “sex” in it, including a site that had the word “marsexploration” in its title]. So clearly, if you’re going to go with filtering, be prepared to make tough calls. (Gebeloff, 1999)

Peacefire—a group critical of filtering software—explains, “We have always felt that filtering software is not only ineffective, but also a violation of the trust between students and staff . . . Unfortunately, most of the censorware companies block anything controversial, not just pornography. I find it very discouraging that this includes information like suicide prevention, safe sex, and gay youth resources” (B. Jenkins, quoted in Kongshem, 1998).

Third, students and computer hackers have already found flaws with such programs and have managed to acquire information from sites that have been blocked. When product evaluators at *Consumer Reports* tested over nine different Web content filters, including AOL’s parental controls,
they discovered that, although AOL offered the best protection, as much as 20 percent of easily located Web sites containing sexually explicit content, violently graphic images, or promotion of drugs, tobacco, crime, or bigotry slipped through the filters. In fact, “Net Nanny displayed parts of more than a dozen sites, often with forbidden words expunged but graphic images intact” (ConsumerReports.Org, 2001).

Fourth, there is an inherent conflict of interest when the main advocates challenging the government’s attempts to protect children from online predation and pornography are the very same groups that seek to profit directly from a “free marketplace” of online smut. In its June 2004 press release, SafeSurf applauded the Supreme Court for its ruling in the Internet pornography case Ashcroft v. ACLU “because the High Court concluded that Internet filtering solutions, such as those originally proposed by SafeSurf over nine years ago, are a better way to proceed than the government restrictions imposed under the Child Online Protection Act” (Jules, 2004). As the chairman of SafeSurf, Ray Sours, exclaimed, “This decision has revealed that the High Court has seen the wisdom in protecting the Internet from governmental censorship and in enabling parental discretion through an intelligent filtering and labeling system. Maybe now, Congress will focus more attention on what has become known as the ‘Safe Surfing’ method of protecting children online” (Jules, 2004, emphasis added). Yet the court’s wisdom is more the result of intense lobbying than constitutional insight. SafeSurf has been lobbying Congress about the constitutionality of the Child Online Protection Act since its implementation, arguing its case before the Congressional Commission on Child Online Protection (COPA) in July 2000, just a few months after COPA’s passage.

Gebeloff addresses this conflict of interest in his critique of net filters and ratings for Money Talks:

I once had a chance to interview Gordon Ross, the fellow who designed Net Nanny. . . . I asked Ross how he, with his background in computer systems, comes up with the list of bad words and unacceptable Web sites that his program blocks. Basically, he told me, it started from a list he put together and then evolved over time to reflect feedback from users. “And we have a disclaimer saying we’re not liable for the list.” (Gebeloff, 1999)

This leads Gebeloff to deduce the ironic disposition of this practice: “We don’t want the government to be our censor, so why should we turn the job over to a computer programmer from British Columbia? The answer, of course, is that we shouldn’t, but that’s what happens when a parent buys filtering software, installs it, and then walks away from their child’s machine” (Gebeloff, 1999).

With laws mandating the use of various forms of censorware to meet government regulations like CIPA, and liability issues at school, the library, or work, it is no surprise that the marketplace of ideas has increasingly chan-
neled its financial resources into for-profit filtering products. Companies easily win over school and library administrators by guaranteeing adherence to government legislation as well as liability protection and parental approval. For $14.95, SafeSurf markets Safe Eyes as an effective tool that “uses the N2H2 website database which has been proven time after time to be the most accurate database available . . . In recent tests, both the U.S. Department of Justice and the Kaiser Family Foundation found N2H2 to be the best” (SafeSurf, 2004b). Official endorsements from prominent governmental, industrial, and educational groups are an added selling point, such as N2H2’s official stamp of approval from the American Library Association for meeting CIPA rules.

As for the pervasiveness of filtering products, a poll conducted as early as 1998 at the Technology + Learning conference revealed that 51 percent of surveyed teachers, technology directors, school board members, and other educators had adopted some form of censorware for all or some students in their district (cited in Kongshem, 1998). Another poll conducted in 2000 by MSNBC.com found that “many users rely on an Internet service provider, or ISP, to do the filtering for them. The big names in this market are America Online, The Microsoft Network, Mayberry USA, Rating-G Online and Getnetwise.com. Filters that are popular with Christians and conservatives include Family.Net, Integrity Online and Hedgebuilders.com” (Nodell, 2000). With no centralized board or groups to review the practices of these filtering companies or ISPs for their effectiveness or appropriateness, it is easy to see how those seeking to meet the needs of their schools, libraries, work, or homes turn to various programs without clear indication of their validity and reliability, especially institutions pressured to have some “safety plan” to meet CIPA legislation or issues of liability.

Accordingly, it is no surprise that filtering producers and marketers stand to gain financially by lobbying for nongovernmental solutions to censorship, as well as a deregulatory media environment allowing telecommunication firms to continue to merge and expand their online assets and streamline Web content. MSNBC’s interest in polling Internet user preferences for filtering is not purely for newsworthiness given its partnership with Microsoft. The same is true for AOL Time Warner. What is more, in addition to cornering the market for libraries, schools, and homes, many of these companies have ventured into the work environment. As MSNBC.com reporter Bobbi Nodell explains, “many filter companies are moving into the corporate market, which is booming because employers are concerned about workers ‘wasting time’ on the job and want to keep them from shopping, checking investments and playing games . . . the corporate market is expected to grow from $60 million in 1999 to $500 million in 2004” (Nodell, 2000).

Confirmation of this trend can be found with Net Nanny. Looksmart, a leading business firm in online search technology, recently acquired Net
Nanny for approximately $5 million in cash and stock in April 2004. Indeed, in their ability to promote and streamline commercial content (while limiting “inappropriate” sites), monitor Internet user habits, profile users for direct marketing purposes, and market products to users, filtering software products can be considered stepchildren of the highly lucrative commercial search engines, which became the most lucrative Web properties in 2003 due to their increasing ability to promote commercial Internet content. As LookSmart CEO Damian Smith stated in 2004:

This acquisition is both strategic and prudent for LookSmart. . . Strategic, because integrating our search technology into Net Nanny provides a stronger product for their users, while also providing LookSmart with a desktop platform from which to launch high margin search and paid listings applications. Prudent, because Net Nanny is expected to produce positive margin contributions for LookSmart in 2004. (LookSmart, 2004)

In other words, this partnership, along with MSN funding, will allow LookSmart to apply its tracking and marketing capabilities to Net Nanny’s software and related proprietary environments. As the company explains to its shareholders, such a partnership “will enhance the leading online filtering software and provide high-quality proprietary search traffic for LookSmart.”

While filtering technology continues to thrive in the Internet’s “free market” system, and as Web content continues to grow exponentially, the profits for filtering technology continue to expand commercially. Net Nanny’s acquisition by LookSmart makes clear that one of the leading “protectors” of illicit online content is poised to become a predator of tracking and marketing to today’s Internet users as it shifts its mission to “high margin search and paid listings applications” (LookSmart, 2004). With substantial profit predictions for filtering companies expanding their business within the corporate market, the goals to protect Internet users, including children, are becoming further marginalized at a time when schools, libraries, and businesses are becoming increasingly dependent upon filtering technology.

To make matters worse, “the Internet’s status as an open forum for ideas” has come under attack since 2002 with a Federal Communications Commission (FCC) ruling that shields cable companies from having to open their networks to smaller competitors and civil liberties and consumer advocacy groups (Wolverton, 2002). As Karen Charman (2002) explains, “without public policies mandating open access,” cable will monopolize broadband width, denying access to other Internet Service Providers in order to capitalize off of hyper-commercialized services that make it easier to buy products. Troy Wolverton (2002) of ZDNet news explains that “lack of competition among cable Internet providers could be a form of censorship . . . even if they don’t completely block Web sites, cable companies
could slow access to them to the point that they become all but impossible to reach . . . while they could speed access to their own sites and those of preferred partners.” Subsequently, if “the Internet content accessed by K-12 youth is patrolled by capitalist institutions, rather than by the government, educational institutions, public libraries or communitarian groups, it will inevitably become more difficult ‘to turn the one-way system of commercial media into a two-way process of discussion, reflection, and action’” (Thoman, 1998, p. 3). As Resnick explains, no matter how well conceived or executed, any labeling or blocking system will tend to stifle noncommercial communication since the time and energy needed to label will inevitably lead to many unlabeled sites: “Because of safety concerns, some people will block access to materials that are unlabeled or whose labels are untrusted. For such people, the Internet will function more like broadcasting, providing access only to sites with sufficient mass-market appeal to merit the cost of labeling” (Resnick, 1997, p. 106). This form of censorship is a serious problem as the possibilities for a decentralized and openly available information network will once again be delimited by a top-down capitalist hierarchy where nondominant, noncommercial, or alternative sources of information will remain peripheral.

Finally, information filtering does not prepare students to learn how to analyze and evaluate information once they are no longer using the Internet within an educational setting. This point has gained momentum as media literacy educators, librarians, and scholars have been grappling with the need for solid media literacy curricula that include a critical and analytical approach to learning with and about online communications technology (Fabos, 2004; Frechette, 2002; Paxson, 2004; Tyner, 1998).

**Testing Content Controls for Cyber-Capitalism**

The hegemonic impulse of online safety profiteers becomes clear when we take a look at some ratings organizations, online proprietary environments, ISPs, and databases recommended by parents, the government, educational institutions, and the industry. First is SafeSurf, a rating organization that claims to be “dedicated to making the Internet safe for your children without censorship.” Through an information database of objectionable sites, a proprietary environment for children, and safety tools for parents, SafeSurf believes they “will enable software and hardware to be developed that will enable more effective use of the Internet for everyone” (SafeSurf, 2004a, emphasis added).

My skepticism about claims that “everyone” benefits through SafeSurf’s methods developed when visiting the SafeSurf home page, where I reviewed their policies, claims, and method to create an environment that is child tested and parent approved. What first drew my attention to their Web site were the various advertisements centered on the page. One ad displayed a large colorful rectangle for Card Service Online, “the leader in online
real time credit card processing,” featuring Mastercard, Visa, Discover, and American Express. Directly under it was an ad for Child Magazine, on sale at the reduced price of $7.95; its pitch: “One year for the price of a bottle.” Beneath this was a bold advertisement link to “Update Microsoft’s Internet Explorer to support SafeSurf Ratings.” Combined, these ads validated my forewarning about the interconnections between powerful computer firms, such as Microsoft, and blocking software products.

My findings led me to presume that more advertising would emerge on the SafeSurf Wave link, which offers Kid’s Wave, a list of “top sites” purportedly “devoted to educating and entertaining children.” On the Kid’s Wave front page, I was informed “There are great places to take your children online.” Below was a grid of partial listings of SafeSurf-approved sites by category. The first category was the “favorite site of the month,” which was Squigly’s Playhouse. By clicking on the cartoon graphic, my hypothesis was reaffirmed: the unfolding visual displayed a large color advertisement for Disneyland with moving graphics and a photo of the Magic Kingdom. The flashing text read “[frame 1: photo and text depicted Disneyland Resort] To really enjoy yourself here; [frame 2: photo of Mickey Mouse described as ‘the Disneyland Trip Wizard’] Pick up your custom schedule here.”

In case the ad was overlooked, each separate clickable Kid’s Wave link for an activity or game was infused with the Disney Resort campaign. For instance, the “Squigly’s Games” page had another large, flashing, color ad for Disney at the top that read, “[frame 1: photo of Mickey Mouse] Are you the Ultimate Disney fan?; [frame 2: photo of Goofey] Click here—enter to win”; on the bottom, a three-frame flashing ad targeted at parents read, “[frame 1] You know what you put on your card; [frame 2] but do you know what he put on your card? [picture of a crowd with a man circled in red]; [frame 3] Find out with your free credit report online.” Other pages, like “Squigly’s Writing Corner” or “Brainteasers,” featured separate Disney ads as well as credit card ads (presumably targeted at parents, but also at a new generation of consumers).

Disney, it seems, is a frequent advertiser on filtering software products. In addition to selling nonsoftware products, such as $40 embroidered golf shirts, Net Nanny’s Internet Web site had an advertisement for Disneyland featured on its front page. Most troubling, however, is that advertising clients are also the sponsors of Net Nanny content. Among its “safe-sites” for kids were “fun” links to Disney, Crayola, and Kids Channel. Under the category “Education” was a Colgate “Kidsworld” link with prominent product advertisements for Colgate toothpaste. Describing its mission in philanthropic terms, Colgate Palmolive Co. purportedly maintains the Internet site “as a service to the Internet community.” A closer look at the page proves otherwise. First, I had to type in my first name and specified password of the day, “toothpaste,” in order to enter the “No Cavities Clubhouse.” There, I was greeted by “Dr. Rabbit” who appeared in his
clubhouse holding a toothbrush and Colgate toothpaste. Although this Web site offered “interesting oral care facts, games, and stories aimed at raising children’s awareness of oral health,” I could not get away from Dr. Rabbit and his Colgate endorsement no matter what activity I clicked on. Moreover, in spite of its “intention” to adhere to the Children’s Advertising Review Unit (CARU) Guidelines for advertising on the Internet and online services, my name and email were still requested so that the “Tooth Fairy” could send me an email message—no doubt carrying her Colgate toothpaste and brush in cyber-flight.

Although not nearly as plastered in advertising as SurfWatch or Net Nanny, CyberPatrol’s Web site unquestionably catered to/partnered with commercial Web sites, including Disney’s Internet empire of kid-targeted Web addresses. A recommended “safe” site was “Toy Story Games,” a game developed by Disney based on its Toy Story movie. Not surprisingly, Disney’s home page was saturated with child and adult-directed advertising. Although the advertising contained here was “2nd level,” meaning that I had to click on the recommended sites before being inundated with ads, the sites contained on the page remained uncontested as child appropriate.

As evidenced within these kid-designated Web sites, the far-reaching clutches of advertisers are rendered invisible in the discourse or underlying rationale of Internet protectionism. While children are deemed to be impressionable when it comes to sex, pornography, adult content, and nefarious language, concerns about manipulative advertising campaigns go largely undetected within “kid-safe” Internet domains.

Conclusion

Media literacy scholar Len Masterman’s explanation of critical autonomy, to “develop in pupils enough self-confidence and critical maturity to be able to apply critical judgments to media texts which they will encounter in their future” (1985, p. 24; emphasis added), does not fit within the logic of commercial filters and the self-regulated corporations attempting to control and streamline Internet content. As Elizabeth Thoman (1998) clarifies, “the media have become so ingrained in our cultural milieu that we should no longer view the task of media education as providing ‘protection’ against unwanted messages.” Hence, a learning model of awareness, analysis, reflection, action, and experience leads to better comprehension, critical thinking, and informed judgments.

Contrary to filtering mechanisms designed to censor or reduce student exposure to “inappropriate” Web sites and online information, a much better approach toward new information technologies is to go beyond teaching students about how to use computers, email, Web browsers, etc. First and foremost, the goals of media literacy must go hand in hand with computer training and online access through the instruction of critical skills by which students learn to discriminate all types of information. While there are
hazards to over-regulation and under-regulation of the Internet, educators and librarians have an important role to play in developing online media literacy initiatives so that students can become discerners of the types of information they need. The goals for taking media literacy to the Internet must go beyond the critical evaluation and use of information to include an analysis and understanding of the impact of political and economic forces that drive and control much of the Internet. Within a “media literacy in cyberspace” model, the issues of ownership, profit, control, and related effects are essential to helping students formulate constructive action ideas that will lead to their own Internet choices and surfing habits (Frechette, 2002). As PICS chairman Paul Resnick (1997) admits, “no labeling system is a full substitute for a thorough and thoughtful evaluation.” In the end, if the power of Internet content labeling, ratings, and restrictions are left to a third party or profit-making companies, then educators, librarians, and parents need to lobby that they serve the public interest rather than private commercial interests.

Note
1. For example, AT&T, Dell Inc., Microsoft, Verizon, America Online Inc., American Library Association, Amazon.com, Center for Democracy & Technology, Comcast, Earthlink Inc., Recording Industry Association of America, Visa USA, Wells Fargo, and Yahoo!

References


Williams, T. (2003). AOL@SCHOOL expands its free online learning service with new alliances


Current Developments and Future Trends for the OAI Protocol for Metadata Harvesting

SARAH L. SHREEVES, THOMAS G. HABING, KAT HAGEDORN, AND JEFFREY A. YOUNG

ABSTRACT
The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) has been widely adopted since its initial release in 2001. Initially developed as a means to federate access to diverse e-print archives through metadata harvesting and aggregation, the protocol has demonstrated its potential usefulness to a broad range of communities. Two years out from the release of the stable production version of the protocol (2.0), there are many interesting developments within the OAI community. Communities of interest have begun to use the protocol to aggregate metadata relative to their needs. The development of a registry of OAI data providers with browsing and searching capabilities as well as accessibility to machine processing is helping to provide a scalable solution to the question of who is providing what via the OAI protocol. Work is progressing on the technical infrastructure for extending the OAI protocol beyond the traditional harvesting structure. However, serious challenges, particularly for service providers, still exist. This article provides an overview of the current OAI environment and speculates on future directions for the protocol and OAI community.

The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) has been widely adopted since its initial release in 2001. Initially developed

Sarah L. Shreeves, Visiting Assistant Professor of Library Administration, IMLS DCC, University of Illinois at Urbana–Champaign, 1301 W. Springfield, Room 52, Urbana, IL 61801, Thomas G. Habing, Research Programmer, Room 155, Grainger Engineering Library, University of Illinois at Urbana–Champaign, 1301 W. Springfield Ave., Urbana, IL 61801, Kat Hagedorn, OAIster/Metadata Harvesting Librarian, University of Michigan Libraries, 920 North University Ave., Ann Arbor, MI 48109, and Jeffrey A. Young, Software Architect, Online Computer Library Center (OCLC), 6565 Frantz Rd., Dublin, OH 43017.

© 2005 The Board of Trustees, University of Illinois
as a means to federate access to diverse e-print archives through metadata harvesting (Lagoze & Van de Sompel, 2003), the protocol has demonstrated its potential usefulness to a broad range of communities. According to the Experimental OAI Registry at the University of Illinois Library at Urbana–Champaign (UIUC) (Experimental OAI Registry at UIUC, n.d.), there are currently over 300 active data providers using the production version (2.0) of the protocol from a wide variety of domains and institution types. Developers of both open source and commercial content management systems (such as D-Space and CONTENTdm) are including OAI data provider services as part of their products. Service providers range from large-scale efforts with a wide scope, such as the National Science Digital Library (n.d.), to small, tightly focused, community-specific services, such as the Sheet Music Consortium (n.d.).

This article provides a brief overview of the OAI environment, two years out from the release of the production version of the protocol. We assume a relatively high level of familiarity with how the protocol works and only give a brief overview. We delve into some of the interesting developments within the OAI world, particularly the use of the protocol within specific communities of interest, the development of a comprehensive registry of OAI data providers, and a resolver for OAI identifiers that extends the protocol beyond its traditional use. We also document some of the current challenges for both data and service providers. We end the article by noting some of the possible future directions for the OAI protocol and community.

Current Developments in OAI Work

The mission of the Open Archives Initiative, the entity responsible for the protocol, is to “develop and promote interoperability standards that aim to facilitate the efficient dissemination of content” (Open Archives Initiative, n.d. a). The Protocol for Metadata Harvesting, a tool developed through the OAI, facilitates interoperability between disparate and diverse collections of metadata through a relatively simple protocol based on common standards (XML, HTTP, and Dublin Core). The OAI world is divided into data providers or repositories, which traditionally make their metadata available through the protocol, and service providers or harvesters, who completely or selectively harvest metadata from data providers, again through the use of the protocol (Lagoze & Van de Sompel, 2001). The OAI protocol requires that data providers expose metadata in at least unqualified Dublin Core; however, the use of other metadata schemas is possible and encouraged. The protocol can provide access to parts of the “invisible Web” that are not easily accessible to search engines (such as resources within databases) (Sherman & Price, 2003) and can provide ways for communities of interest to aggregate resources from geographically diffuse collections. The protocol promotes a structure in which data providers can focus on
building collections and content, and service providers can focus on building services for these collections and content. While the protocol itself says nothing about what happens to metadata once harvested, usually service providers aggregate, index, and build search/retrieval and other value-added services around the harvested metadata. It has been two years now since the production version of the protocol was introduced (Lagoze, Van de Sompel, Nelson, & Warner, 2002a). Below we discuss just some of the current trends and developments within the OAI community.

**Community- and Domain-Specific OAI Services**

As mentioned above, the Open Archives Initiative emerged from and was initially designed to meet the needs of the e-print archives community (Warner, 2003). However, it was recognized fairly early in the protocol’s development that it could be applicable in a broad range of communities, including, but not limited to, libraries, museums, and archives. In fact, the implementation guidelines (Lagoze, Van de Sompel, Nelson, & Warner, 2002b) are deliberately nonspecific so as to provide room for community-specific applications of the protocol (Lagoze & Van de Sompel, 2003).

The initial push for developing OAI service providers was in part due to the Andrew W. Mellon Foundation grants in 2001 (Waters, 2001). The foundation issued seven grants to institutions interested in researching the development of service providers. Three institutions developed publicly accessible services predicated on their research: the AmericanSouth.org project at Emory University; the Digital Gateway to Cultural Heritage Materials at the University of Illinois at Urbana–Champaign (UIUC); and the OAIster project at the University of Michigan. Each service had a different focus. The AmericanSouth.org project focused on aggregating content related to the culture and history of the American South while involving scholars in the process of selection and interpretation (Halbert, 2003). The UIUC project aggregated metadata relating to cultural heritage resources, including finding aids (Shreeves, Kaczmarek, & Cole, 2003), and the OAIster project harvested all possible repositories but kept only those records that pointed to actual digital objects (Hagedorn, 2003).

The different foci were indicative of the future progress of service providers. No one service provider can serve the needs of the entire public, hence user group–specific service providers have become the norm. Many communities have adopted or are in the process of adopting the OAI protocol to help provide federated access to dispersed resources. These communities of interest are significant not only because they have adopted the protocol for a specific domain but also because they have developed additional standards, tools, and metadata schemas to use along with the OAI protocol—much as the originators of the protocol had hoped. Indeed, these domain- and user-specific services may be the best example of what the OAI protocol has to offer.
We highlight three notable community- or domain-specific services in various stages of development below. For a fuller documentation of community-specific service providers and data providers, see the 2003 Digital Library Federation report (Brogan, 2003) and the recent series of profiles of service providers in Library Hi Tech News (McKiernan, 2003a, 2003b, 2004).

**Open Language Archives Community** The mission of the Open Language Archives Community (OLAC) is to create “a worldwide virtual library of language resources” through development of community-based standards for archiving and interoperability and a “network of interoperable repositories” (Open Language Archives Community, n.d. a). OLAC uses the OAI protocol as a means to the latter end. OLAC has extended the protocol to meet the needs for its particular community, specifically through the maintenance of a specialized metadata schema (based loosely on unqualified Dublin Core), data provider tools (including a range of options for organizations without the technical infrastructure to support full-fledged OAI data providers), and service provider tools (Simons & Bird, 2003). Currently OLAC provides access to metadata harvested from twenty-seven data providers through search services hosted at the Linguist List (n.d.) and the Linguistic Data Consortium (n.d.). This integration of search services within important community Web sites increases the visibility and value of OLAC.

**Sheet Music Consortium** The Sheet Music Consortium is a group of four academic libraries—UCLA, Johns Hopkins University, Indiana University, and Duke University—that are building a freely available collection of digitized sheet music. Sheet music presents a particular problem for cataloging because of its various elements: cover art, the sheet music itself, the lyrics, etc. (Davison, Requardt, & Brancolini, 2003). The consortium provides standards for using unqualified Dublin Core to describe sheet music and guidelines for implementation of data provider services. The search service allows the creation of “virtual collections” and allows users to annotate the metadata records (Sheet Music Consortium, n.d.). While work on this service is still in progress, the focus on building a service provider based on a specific type of material makes it well worth watching.

**National Science Digital Library** The National Science Digital Library (NSDL) provides access to the content of collections of science-based learning objects (National Science Digital Library, n.d.). The OAI protocol is the primary means of aggregating the metadata describing this content, although other means are used as well (Lagoze et al., 2002). Funded by the National Science Foundation (NSF), the NSDL has the broadest vision of the service providers described here in that it is attempting to build and aggregate not just a series of digital collections and content but also services to use these resources and the infrastructure to support both. As such, NSF has invested significant resources in the development of content,
services, and infrastructure. The NSDL maintains standards for metadata and guidance for data providers. The NSDL aims for a broad user base (K–12), but its core mission remains to develop this “learning environment and resources network” for science education (Zia, 2001).

Comprehensive OAI Registry of Data Providers

As the OAI community has matured, and especially as the number of OAI repositories and the number of data sets served by those repositories has grown, it has become increasingly difficult for service providers to discover and effectively utilize the myriad repositories. In order to address this difficulty the OAI research group at UIUC has developed a comprehensive, searchable registry of OAI repositories (Experimental OAI Registry at UIUC, n.d.).

Shortcomings of Existing Registries

There were and continue to be several other registries of OAI repositories such as those maintained by the Open Archives Initiative Web site (Open Archives Initiative, n.d. b) and OLAC (Open Language Archives Community, n.d. b). However, nearly all of these suffer from a number of shortcomings. Probably foremost is that the registries typically maintain very sparse records about the individual repositories, usually nothing other than flat lists of base URLs and possibly the repository name. Typically, there is no search mechanism and fairly limited browsing capabilities. An onerous amount of manual snooping using the OAI-PMH verbs directly in a Web browser is usually required by potential service providers before they can assess the utility of a specific repository for their needs.

A second shortcoming of the existing registries is completeness. The registries are usually populated by self-registration or maintained to support the specific needs of a unique community, so few of the registries approach a complete list of all available repositories. “Googling” or following friends or provenance links reveals many new OAI repositories that are not listed in any of the existing registries, even taken as a whole.

Developing the Experimental OAI Registry

In developing OAI service providers for various projects within the UIUC Library, the issues of completeness and discoverability have become more evident. The UIUC research group thus built the Experimental OAI Registry to address these problems. Moreover, based on feedback after the first public announcement of the Registry on the OAI-Implementers listserv, the group realized that the Registry also could be utilized to meet various other needs in the OAI community, such as the need for various output formats to support machine processing of the Registry.

Completeness The UIUC research group addressed the completeness issue by employing three different strategies. The first strategy was a simple inventory of existing registries, both formal and informal, that listed differ-
ent repositories. The second strategy involved following various links that were contained within the OAI responses. The first source of links was the “friends” container (Lagoze, Van de Sompel, Nelson, & Warner, 2002b). This container could be included as one of the optional description elements in an OAI “Identify” response. It allows an OAI repository to list other confederate repositories that may be of interest to a harvester. It is also commonly used by aggregator repositories. The other source of links was the “provenance” container (Lagoze, Van de Sompel, Nelson, & Warner, 2002b). This container could be included as one of the optional “about” elements of an OAI record. The provenance container stores data about the original source of a record that has been aggregated into a different repository. Using “friends” and “provenance,” it was possible to recursively crawl webs of related OAI repositories. The registry maintains this linking information about each repository to produce a network graphic. The third strategy involved using the Google™ SOAP-based Web toolkit (Google, n.d.). Using this toolkit the research group was able to programmatically search the Google Web indexes to find OAI repositories. The group developed a number of search strategies, from using OAI related keywords such as “OAI” or “Open Archives,” to using special Google keywords such as “allinurl:verb=Identify,” which will find Web sites that contain the string “verb=Identify” in their URL. This latter strategy proved the most successful. Once a candidate base URL is discovered, it is tested to determine whether it can respond to the OAI “verb=Identify” request. If it responds, it is assumed to be a valid OAI repository and it is added to the registry. Finally, requests to manually add repositories to the registry are accepted. In the future, self-registration should become an automated procedure.

**Searchable and Browsable** The second major objective was to make it possible to search for OAI repositories using various criteria and browse through different views of the registry without any manual cataloging of the various OAI repositories. To accomplish this the research group developed processes to automatically harvest and index various data from each repository. Essentially, a specialized harvest of each repository is performed. This harvest collects data from the Identify, ListSets, and ListMetadataFormats responses, supplying these data to various tables and fields in a relational database. In addition, sample records from each OAI repository are collected for each combination of set and metadataPrefix supported by the provider. These data are also added to the relational database. Once these data are indexed, including the full-text of each response, various searches and views of the registry are possible.

The primary supported search is for keywords appearing in the various OAI responses, namely Identify, ListSets, and the sample records. A key observation resulting from our search system is that repositories, including rich collection-level metadata either in the optional Identify description
containers or the optional ListSets setDescription containers, will fare better in terms of discoverability. This suggests the desirability of broader use of collection-level metadata by the OAI community.

Amenable to Machine Processing  The third major goal was to expose the registry’s data in ways that were useful for machine processing. The most obvious way to make the registry accessible for machine processing was by making it an OAI repository itself. Thus, basic Dublin Core records about each OAI repository contained in the registry can be harvested via the OAI-PMH. The ERRoL service, described below, is an example of an application that utilizes the OAI-PMH interface to the registry. In the future, additional metadata formats might be harvestable as well, such as the ZeeRex format used by the Search and Retrieval Web/URL Service (SRW/U) protocol (ZeeRex, n.d.). In addition, the registry is also an RDF Site Summary (RSS) news feed provider. Using RSS a person can monitor the registry for new or modified repository records. The RSS feed is available off of the registry Web site (Experimental OAI Registry at UIUC, n.d.). There are also a number of ways to export repository records from the registry. Any list of repositories resulting from a search or a browsable view can be exported using the XML schema of the “friends” description container.

Work is also progressing on a “harvest bag” feature. This would allow a user to accumulate a custom list of repositories, including sets and metadata formats, that they could export in a standard XML schema. This would be similar to the “book bag” feature of other digital library portals, which allows users to save and export lists of bibliographic citations. The vision is that the “harvest bag” list could then be imported into harvesting software to initiate a harvest of the selected sites.

In addition, the research group is working on a SRW/U search service for the registry (SRW, n.d.). This would allow SRW/U clients to search the registry in a manner similar to that provided by the Web forms search interface. The record formats available via the SRW/U interface would be the same as those available via the registry’s OAI provider.

Future Work  While the registry is now fully operational, there remain a number of improvements the group would like to make to increase its usefulness. Following, in no order, are some plans for future enhancements to the registry:

- Enhance the collection-level description of the repositories to enable better search and discover. This might include both manual cataloging and the application of automated classification algorithms to the repository’s records.
- Provide more automated maintenance of the registry, including the ability of OAI data providers to securely add or modify their repository’s records in the registry, including collection-level descriptive data.
• Improve the automated discovery of new repositories, such as automatically running the Google SOAP-based harvester.
• Delegate the creation and maintenance of virtual collections of repositories, including collection-level metadata.
• Improve the view of search results, especially the context of the search hit. The current system does not identify the context of a search hit, which could be the Identify or ListSets responses or the sample records.

Extensible Repository Resource Locators (ERRoLs)

As mentioned above, according to the conventional model of OAI, the world is divided into data providers and service providers. As it happens, though, a few simple tricks with style sheets and HTTP redirects allow an OAI repository to stand alone as an independent Web application. Early examples of this were created by enhancing individual repositories, as discussed elsewhere (Van de Sompel, Young, & Hickey, 2003). Frustration with changing the OAI world one repository at a time, however, led to the development of the ERRoL resolution service (Extensible Repository Resource Locators, n.d.), which automatically extends these same features and more to any OAI repository in the UIUC registry.

ERRoLs are “Cool URLs” (Berners-Lee, 1998) to content and services related to information in an OAI repository. In essence, the ERRoL service is a resolver for oai-identifiers. In its simplest form, the oai-identifier for an item (such as “oai:lcoa1.loc.gov:loc.pnp/cph.3b37282”) can be resolved by appending it to the end of the ERRoL service URL “http://errol.oclc.org/,” as in “http://errol.oclc.org/oai:lcoa1.loc.gov:loc.pnp/cph.3b37282.” The ERRoL service begins the resolution process by parsing the repository identifier (“lcoa1.loc.gov”) from the URL and using it to obtain the official OAI base URL from the UIUC registry. With this, the ERRoL service constructs a standard OAI GetRecord (oai_dc) request to the home repository, which is what the client sees in response.

As a resolution result, however, an XML OAI GetRecord response is of marginal interest at best. Fortunately, appending various extensions to the basic URL form can produce different kinds of results. For example, if we want this same oai_dc record stripped from the OAI GetRecord wrapper, we can append the “oai_dc” metadataPrefix to the URL, as in “http://errol.oclc.org/oai:lcoa1.loc.gov:loc.pnp/cph.3b37282.oai_dc.” This home repository can also supply a “marcxml” record for this same oai-identifier, which can be obtained by appending a “.marcxml” extension, as in “http://errol.oclc.org/oai:lcoa1.loc.gov:loc.pnp/cph.3b37282.marc21.” Any metadataPrefix available for this item can be added as an extension. This ability to strip a record from its OAI GetRecord wrapper becomes particularly interesting when OAI repositories contain XML content, beyond metadata. Here are examples for a repository that can disseminate XHTML...
(metadataPrefix = xhtml), XSL Stylesheets (metadataPrefix = xsl), and XML Schemas (metadataPrefix = xsd) respectively:


Keep in mind that the ERRoL service is stripping these XML documents from OAI GetRecord responses that it retrieves from the home repository. Each shares the same oai-identifier as the oai_dc metadata record that describes it, which, as explained above, can be obtained by changing the extension to “oai_dc.” Having content and metadata in such close proximity makes it easy to build lightweight, interactive, self-descriptive, content-based, automated systems using XSLT and other thin clients.

These examples demonstrate that ERRoLs are a simple mechanism for accessing various manifestations of OAI data, but it cannot be said that they elevate an OAI repository to the level of a human-interactive Web application yet. But just as ERRoLs transformed standard OAI responses into other forms in the examples above, they can just as easily transform them into HTML using the “.html” extension, as in “http://errol.oclc.org/oai:lcoa1.loc.gov:loc.pnp/cph.3b37282.html.” The “.html” extension, as well as others, not only works at the item level with oai-identifiers but also at the repository level with repository-identifiers. In the case of repository-identifier “lcoa1.loc.gov,” URL patterns like “http://errol.oclc.org/lcoa1.loc.gov.html” are possible. Furthermore, standard OAI parameters can be appended to this URL to produce HTML renderings of all the OAI-PMH responses, as in “http://errol.oclc.org/xmlregistry.oclc.org.html?verb=ListRecords&metadataPrefix=oai_dc&set=XSLStylesheets.”

ERRoLs work with any OAI repository that has a unique repository-identifier registered at the UIUC Experimental OAI Registry. In the case of the “.html” extension, the repository displays integrated identity and branding information gleaned from the repository’s “Identify” response, but otherwise the repositories share the same look and feel. It is possible, however, for individual repositories to instruct the ERRoL service to use an alternate style sheet by inserting a <description> element in their “Identify” response. Thus, the GSAFD Thesaurus repository (OCLC, n.d. a) looks and acts differently from the default style shown above. The list of custom style sheets is currently limited to an approved set, but a mechanism is planned that will open this up to arbitrary style sheets.

Other extensions are available at the repository and item levels, and new ones are in the works. It is even possible for individual repositories to specify custom extensions by defining them in “Identify” response <description> elements, although this feature is not fully developed yet.
Having shown the promise of ERRoLs, though, a few words of caution are needed. ERRoLs operate by dynamically interacting with data providers via the OAI-PMH protocol. If these repositories are offline, slow, or less than fully OAI-compliant (which is frequently the case), the ERRoL functions will suffer. Nevertheless, these examples should show that ERRoLs are an interesting alternative to the conventional OAI model.

**Ongoing Challenges for the OAI Community**

We have highlighted a number of developments and ongoing work within the OAI community (and there are many more). But as the number of OAI data providers has grown, two broad areas of concern have arisen, particularly for service providers. These center on the variations and problems with data provider implementations and on the metadata itself. A third concern is the lack of communication among service and data providers. The metadata issues in particular have been well documented (Shreeves, Kaczmarek, & Cole, 2003; Halbert 2003; Hagedorn, 2003; Arms, Dushay, Fulker, & Lagoze, 2003), but we highlight some of the major issues in all areas of concern below.

*Metadata Variation*

While metadata must be created using unqualified Dublin Core (DC) encoding, as well as any other kind of encoding the data provider wishes, the choice of how to use the encoding standard and/or how to fit the encoding to metadata values that already exist varies widely among data providers. One institution’s choice of how to use the DC Type tag can vary greatly from another’s (for example, “HTML” vs. “Preprint”). This can make it difficult to create a search environment in which users feel certain they are receiving what they need. For instance, to normalize data (such as date or type elements) so search limiters can be used requires the development of common values among many disparate ones. The normalization of the subject element—with many different controlled vocabularies (or merely keywords) used by the different data providers—is, for most service providers, prohibitively resource intensive.

*Metadata Formats*

In the same vein, the problem of harvesting a data repository’s additional metadata formats (beyond unqualified Dublin Core) can be a difficult task. For a large service provider with a standard method for processing harvested metadata, including new formats involves adding additional paths to the processing routines. The more formats, the more complex it becomes. Additionally, large service providers may have developed interfaces conforming to the simple Dublin Core standard and not have the ability to integrate more complex and more varied formats. For this, service providers need more all-encompassing game plans and better internal support.
OAI Data Provider Implementation Practices

The OAI protocol is flexible in that there are relatively few required pieces for implementation: valid responses to OAI verbs, the use of oai_dc, a unique and persistent OAI identifier, and a date stamp. The OAI Guidelines for Implementation have a limited technical scope, are intended for a general audience of implementers, and do not describe the consequences of not implementing some of the optional features of the protocol (Lagoze, Van de Sompel, Nelson, & Warner, 2002b). This has meant that many of the features of OAI, such as sets, use of descriptive containers, etc., that are quite helpful for service providers, have been underutilized. Data providers also need to be aware of how their implementation of required items such as date stamps impacts service providers.

Communication Issues

The OAI community is very loosely federated. There are general and technical listservs available through the Open Archives Initiative. However, as some of the issues above illustrate, a serious need for best practices and guidelines exists for both data and service providers. An informal community of service providers has appeared who advise each other on the technicalities of performing harvesting and maintaining their service. While this ad hoc community is welcome, a more formal method of communication between data and service providers is needed.

Future Directions

We have discussed above just some of the current developments in the OAI community. Below we outline some future directions. This list is not meant to be all inclusive but rather a taste of some of the ongoing research and practices in the OAI community.

Best Practices

As indicated above, service providers face serious challenges in both their harvesting and aggregating activities. The development of community-specific best practices and implementation guidelines has been an important part of OLAC and other domain-based service providers. A group of service providers within the Digital Library Federation (DLF) has now begun work on some more general best practices to be used with the DLF and beyond.

Static Repository Gateway

The technical hurdle is still sometimes too great for potential data providers. The Static Repository Gateway, developed at the Los Alamos National Laboratory, is the most recent option for OAI data providers and provides a very low entry point (Van de Sompel, Lagoze, Nelson, & Warner, 2002; Hochstenbach, Jerez, &Van de Sompel, 2003). Essentially, a resource developer can post a single large XML file containing the metadata and
OAI wrappers on its Web server. This file can be accessed through an OAI gateway service. Currently two service providers, UIUC and the University of Michigan, have been working to shepherd potential data providers to one gateway, which has proved very simple for both the service and data providers.

**Mod_oai Project**

The mod_oai project, funded by the Andrew W. Mellon Foundation, is developing a tool that makes content that is accessible from Apache open-source Web servers available through the OAI protocol. This tool will essentially extend the benefits of selective and incremental harvesting available through the OAI protocol to the general Web community (Mod_oai, n.d.).

**OAI-rights**

The OAI-rights committee is working toward a means of incorporating structured rights statements about the resources exposed (that is, the metadata) through the protocol (Lagoze, Van de Sompel, Nelson, & Warner, 2003). The committee does not intend to define a new rights language but only to provide the means of communicating a structured, defined language within the protocol.

**Controlled Vocabularies and OAI**

Controlled vocabularies will become more important as data and service providers try to cope with the chaos that develops from aggregating metadata from diverse sources. Controlled vocabularies will become particularly important within self-archiving systems such as institutional repositories and e-print archives (many of which are also OAI data providers); in many cases there is no cataloger to exert quality and authority control. A lightweight solution to this would be for authority agencies to mount their thesauri as an SRW/U search service, register it with the UIUC registry, and use ERRoLs to provide an HTML interface and URL access to items in the repository (OCLC, n.d. a).

**SRW/U-to-OAI Gateway to the ERRoL Service**

This service will allow institutions to load their data as an SRW/U search service, register it with the UIUC gateway, and automatically get OAI-PMH and ERRoL functionality for free. The OCLC Research Publications OAI repository is the first demonstration of this. This configuration adds searching capability to the mix of ERRoL features (OCLC, n.d. b).

**References**


Lessons Learned with Arc, an OAI-PMH Service Provider

XIAOMING LIU, KURT MALY, MICHAEL L. NELSON, AND MOHAMMAD ZUBAIR

ABSTRACT

Web-based digital libraries have historically been built in isolation utilizing different technologies, protocols, and metadata. These differences hindered the development of digital library services that enable users to discover information from multiple libraries through a single unified interface. The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) is a major, international effort to address technical interoperability among distributed repositories. Arc debuted in 2000 as the first end-user OAI-PMH service provider. Since that time, Arc has grown to include nearly 7,000,000 metadata records. Arc has been deployed in a number of environments and has served as the basis for many other OAI-PMH projects, including Archon, Kepler, NCSTRL, and DP9. In this article we review the history of OAI-PMH and Arc, as well as some of the lessons learned while developing Arc and related OAI-PMH services.

Interoperability is one of the significant research problems in the field of digital libraries (DLs) (Lynch & Garcia-Molina, 1995). The inability to federate, filter, and provide value-added services on remote content limits DLs to covering only local holdings. The Open Archive Initiative (OAI) is a major, international effort to address technical interoperability and facilitate discovery of content among distributed repositories. OAI differs from other interoperability approaches, such as Z39.50 (Lynch, 1997) or SDLIP (Paepcke et al., 2000), through its emphasis on a limited, simple, and easy to implement protocol that layers over an existing repository. The
OAI framework defines two functional roles: data providers (also “repositories”) and service providers (also “harvesters”). Service providers develop value-added services that are based on the metadata collected from data providers. These value-added services could take the form of cross-archive search engines, linking systems, and peer-review systems.

The roots of the OAI lie in a vision to stimulate the growth of open e-print repositories. This concept began to be developed with the Universal Preprint Service (UPS) prototype (Van de Sompel et al., 2000), and was further advanced with the Santa Fe Convention (Van de Sompel & Lagoze, 2000). The UPS prototype was the discussion piece during an invitation-only workshop in Santa Fe, New Mexico, in the fall of 1999. This workshop brought together many of the leaders in the e-print community for the purpose of fostering interoperability between the various author-contributed e-print servers and institutional repositories in use at the time. Contemporary approaches toward interoperability were ad hoc at best. One of the distinguishing factors for the Santa Fe Workshop was the collective experience in building DLs and the associated interoperability problems; earlier interoperability workshops (Scherlis, 1996) were comparatively premature. The immediate result of this workshop was the Santa Fe Convention, an intermediate step toward the metadata harvesting model that would become the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH).

Realizing that the simple metadata harvesting idea had appeal to a broader reach of communities than that engaged in e-print publishing, version 1.0 of the OAI-PMH was released in January 2001. Following an extended period of evaluation and alpha and beta testing, version 2.0 of the OAI-PMH was released as a stable specification in June 2002 (Lagoze, Van de Sompel, Nelson, & Warner, 2002a). The development, history, impact, and secondary effects of OAI-PMH have been discussed in several publications, including Lynch (2001), Nelson (2001), Lagoze and Van de Sompel (2001), Van de Sompel and Lagoze (2002) and Lagoze and Van de Sompel (2003).

**ARC**

Arc (http://arc.cs.odu.edu) is the first end-user federated search service based on the OAI-PMH (Liu, Maly, Zubair, & Nelson, 2001). The Repository Explorer (Suleman, 2001) was released prior to Arc, but its targeted audience is mainly repository developers and maintainers, not end-users. Arc was initially released as an experimental service to investigate issues in metadata harvesting in October 2000. The software developed for the Arc service (http://oaiarc.sourceforge.net/) was released as an open source system under NCSA-style license in September 2002. It has been used in several production and research projects (see Table 1).

Arc was first developed as a proof-of-concept service for OAI-PMH;
however, the development of Arc revealed interesting problems and inspired further research in these domains. In this article we introduce the development and architecture of the Arc system and follow-up research that attempted to improve or optimize the metadata harvesting system and search performance. We will discuss the Archon project for building value-added services to take advantage of rich metadata beyond Dublin Core (DC) (Weibel & Lagoze, 1997); the DP9 service to allow general search engines (Google, Yahoo, etc.) to index OAI-PMH compliant collections; and the recently funded Andrew Mellon Foundation DL Grid project for building a high-performance federated search service. When possible, interesting and general features resulting from these research projects are incorporated back into the publicly available Arc source code distribution.

**Development of Arc**

Arc was initially released as an experimental service to investigate issues in metadata harvesting. It immediately attracted interest because it was the only vehicle to demonstrate the potential and promise of OAI-PMH at that time. As new data providers appeared, they often requested to be added to the Arc system for demonstration purposes; by continuously integrating various new data providers, the software was made stable and fault tolerant. Originally conceived as more of a tour de force, Arc has become a useful tool for helping new data providers to make their collections truly OAI-PMH-compliant by giving them feedback on errors during harvesting.

When applying the Arc software in various environments, we encountered a number of problems such as inconsistent metadata, lack of controlled vocabulary, and XML errors. Based on feedback from other adopters, we have been able to address these problems and have consequently added many new features for customization and installation. The architecture of the Arc system has been refined to easily add or extend new functionalities.

Arc is available for download (http://sourceforge.net/projects/
OAI-PMH uses unqualified Dublin Core as the default metadata set, and most Arc end-user services are implemented on the data provided in the DC metadata. The current supported end-user services include simple search, advanced search, interactive search, annotation service, and browse/navigation over search result. Arc has a Web-based administration interface, which allows users to configure various parameters for harvesting and to check harvester logs to handle various error situations such as erroneous XML replies from data providers.

**ARCHITECTURE OF ARC**

The basic structure of OAI-PMH supports two basic components: the service provider and the data provider. Data providers administer systems that support the OAI-PMH as a means of exposing metadata, and service

![Figure 1](http://example.com/image.png)
providers use metadata harvested via the OAI-PMH as a basis for building value-added services.

The OAI-PMH focuses on the clear interface between data providers and service providers. In Figure 2 we define the Arc model for metadata harvesting that addresses many of these issues. The data provider maintains one repository for digital records. Then a number of service providers work together to conduct metadata harvesting. The harvester is the key service that uses OAI-PMH to maintain the synchronization between data providers and other services, such as centralized federation services, replication services, and citation linking services. In addition, the Arc system includes OAI-PMH proxy, cache, and gateway services to optimize the functioning of the model underlying the OAI-PMH technology (Liu, Brody, Harnad et al., 2002). These services provide an infrastructure that can be used by all other components to achieve interoperability, scalability, and reliability.

Data Providers

A data provider supports the OAI-PMH as a means of exposing metadata. The design of a good data provider presents many challenges, including metadata quality, server availability, and service quality. The quality of data providers has been a significant problem since the genesis of OAI-PMH at the Santa Fe Convention in 2000. Data providers are frequently unreachable on the network or have errors in the data they return in response to OAI-PMH harvesting requests. Details of this phenomenon can be found at Celestial (http://celestial.eprints.org), a service that tracks the stability of data providers over time as well as provides a cache of the data provider’s contents. During the testing of harvesting from OAI-PMH data providers, we were able to overcome particular problems regarding compatibility and adaptability.

Figure 2.
Initially, the use of unqualified DC as a common metadata format in OAI-PMH proved to be very helpful for building a quick prototype, and thus it is continuously used by several service providers. However, richer metadata formats are essential for building a richer service. Dublin Core is a “lingua franca” metadata format only—it is best suited for resource discovery rather than rich semantic description. The OAI encourages the simultaneous exposure of richer, community-specific metadata formats as well. Although a number of data providers have supported richer metadata formats, it is difficult to implement richer services over these metadata without individually studying each format. Consequently, we developed a series of interfaces relying on interactive user refinement to navigate a large corpus of metadata with varying quality, structure, and semantics (Liu, Maly, & Zubair et al., 2002). We found that, while it is possible to search the corpus in this manner, it places a high cognitive load on the user. Further research in this area is required to achieve the “deep semantic interoperability” identified by Lynch and Garcia-Molina (1995).

Harvesters

Similar to a Web crawler, an OAI-PMH harvester traverses the data providers automatically and extracts metadata. However, there are two significant differences between the OAI-PMH harvester and a Web crawler: the harvester recognizes metadata formats, thus allowing use of structured data, and the harvester exploits the incremental harvesting defined by the OAI-PMH, allowing more efficient extraction of information than a regular Web crawler.

For example, consider a Web site of 100 pages that is available through regular Web crawling and an OAI-PMH repository interface. Assume 5 of the pages are updated weekly and that the Web site is harvested weekly both by a Google robot and by an OAI-PMH harvester. Assume that the OAI-PMH interface is configured to distribute 10 documents, batched together, per connection. Assuming that smart Web robots perform conditional HTTP GETs (http status code 304) based on last modified dates, the Google robot will not download more files than it needs to, but it will have to query every individual Web page of the Web site to determine its date of last modification. Figure 3 illustrates this model.

But the OAI-PMH model saves the considerable overhead of establishing TCP/IP and HTTP connections for documents that have not changed. Instead of having to ask each of the 100 files if their modification date has changed, the harvester asks the OAI-PMH interface which files have changed, and the OAI-PMH interface only responds with the files that meet the criteria (see Figure 4).

Given the parameters stated above, Table 2 shows the relative load placed by each method. If the Web site is larger, say 1,000 or 10,000 files, the unnecessary network traffic avoided with OAI-PMH would be even
greater. Even if Web sites updated their content (or added new content) more rapidly, the OAI-PMH approach would still reduce the number of connections by a factor of the number of files batched together in the response (in this example, by a factor of 10). Not only would this reduce the network load for the robots and Web sites, it would also allow for much quicker harvesting of updates and thus more up-to-date Web indices.

Because data providers are different in data volume, partition definition, service implementation quality, and network connection quality, all these factors influence the harvesting procedure. Historical and newly published data harvesting have different requirements. When a service
provider harvests a data provider for the first time, all past data (historical data) needs to be harvested, followed by periodic harvesting to keep the data current. Historical data harvests are high volume and more stable. The harvesting process can run once or, as is usually preferred by large archives, as a sequence of chunk-based harvests to reduce data provider overhead. To harvest newly published data, data size is not a major problem, but the scheduler must be able to harvest new data as soon as possible and guarantee completeness. The OAI-PMH provides flexibility in choosing the harvesting strategy, although optimizing it remains an open question.¹

Scalability Through Hierarchical Harvesting and “Aggregators”

A service provider can also act as a data provider, disseminating metadata harvested from other data providers. This allows for the hierarchical harvesting of content and removes a limitation of having all data providers be at the same “level.” This structure has a great deal of flexibility in how information is filtered and interconnected between data providers and service providers. While hierarchical harvesting was not originally part of the OAI-PMH, there was nothing in the protocol that prohibited it. Arc was the first service provider to introduce hierarchical harvesting, and services that provide hierarchical harvesting are now known by the name of “aggregators.”

Aggregators may normalize, correct, transform, or otherwise change the harvested metadata. Thus, the re-exposed data might not be the same data harvested from the original data providers. Unless the metadata exposed by the aggregator is completely unchanged, the aggregator must issue new identifiers for the OAI-PMH records it makes available for harvesting. The OAI-PMH defines provenance containers to assist in the de-duplication of metadata harvesting from various sources. Guidelines have since been written to assist in the development of OAI-PMH proxies, caches, and aggregators (Lagoze, Van de Sompel, Nelson, & Warner, 2002b).

An illustrative example of an aggregator in action is the NASA Technical Report Server (NTRS) (Nelson, Rocker, & Harrison, 2003). NTRS (http://ntrs.nasa.gov) is a public digital library that provides users access to NASA-authored reports, reprints, and other aerospace related materials. NTRS uses the OAI-PMH to harvest metadata from twelve different data providers. If the developers of other digital libraries wish to include NASA material into their DL, they could harvest directly from the twelve data providers from which NTRS harvests. Or, because NTRS is also an

<table>
<thead>
<tr>
<th></th>
<th>Initial Harvest</th>
<th>Weekly Updates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google Robot</td>
<td>100 connections</td>
<td>100 connections</td>
</tr>
<tr>
<td>OAI-PMH Harvester</td>
<td>10 connections</td>
<td>1 connection</td>
</tr>
</tbody>
</table>

Table 2. Network and Server Load by Harvesting Type
OAI-PMH aggregator, they could simply harvest the metadata via OAI-PMH directly from NTRS. NTRS effectively becomes a one-stop shop for NASA and aerospace materials. This hierarchical construct could be repeated many times, allowing complex harvesting systems to be constructed. In the example above, harvesting directly from NTRS not only removes the burden of harvesting from the twelve sites directly, but it also addresses the problem of discovering new repositories. If NTRS adds a new aerospace repository to its collection, the sites harvesting NTRS will automatically gain this new collection on their next harvest.

Registration and “Friends”

Discovery of new and existing data providers is an important and difficult topic for the OAI community. The OAI-PMH does not define an explicit registration service: repository awareness either happens out of band from the OAI-PMH (for example, a repository administrator manually submits the repository base URL to a service provider), or a harvester learns of other repositories through an optional “friends” container in the “Identify” response. This allows a repository to “link” other repositories of which the administrator is aware. This is similar to a Web crawler learning of new Web servers by crawling and analyzing Web pages.

It might seem tempting at first to try to include a registration service as part of OAI-PMH. But, ultimately, this would not allow large-scale deployment of OAI-PMH data providers. In the same way that you can install a Web server and not have to register it with any authority, data providers can be set up in the same way. There are Web sites that do try to keep track of the publicly available data providers, but these are definitely not complete, just as any list of public Web servers is not complete. Discovery of data providers remains a difficult issue within the OAI community, but we feel that “friends” and aggregators will provide the necessary mechanisms for addressing this issue.

Proxy, Cache, and Gateway

The current and emerging applications based on metadata harvesting require a scalable and reliable infrastructure to support them. OAI-PMH proxies, caches, and gateways are tools to optimize the functioning of the data provider/service provider model underlying the OAI-PMH. An OAI-PMH proxy dynamically forwards OAI requests to data providers. For example, it can dynamically fix common XML encoding errors and translate between different OAI-PMH versions. An OAI-PMH cache caches metadata and can filter and refine them before exposing them to service providers. It also serves as a simple cache that reduces the load on source data providers and improves server availability. An OAI-PMH gateway can convert the OAI-PMH to other protocols and applications. For example, the gateway could convert between different protocols (for example, SOAP) and OAI-
PMH. The goal is to achieve interoperability, scalability, and reliability of OAI-PMH services.

**End-User Services**

Applying the OAI-PMH harvester to a compatible data provider can lead to many harvesting possibilities—the most obvious being federated search services and repository synchronization. Based on OAI-PMH, there are two approaches to building a federated digital library that allow users to access content in all the libraries through a single interface: centralized and replicated. In the centralized approach, a federation service harvests metadata from the OAI-PMH-enabled libraries and provides a unified interface to search all the collections. This approach has been adopted by Arc and other OAI-PMH service providers.

However, a centralized search service is not a suitable approach if the primary objective is to use native library interfaces. A centralized approach also suffers from the organizational logistics of maintaining a centralized federation service and having a single point of failure. The replicated approach addresses these problems. This approach can be viewed as mirrored OAI-PMH-compliant repositories, where every participant has its own federation service. The consistency between these services is maintained using OAI-PMH. As a federation service is locally available, it becomes easy to push other participants’ metadata into the native library. In addition, this approach supports several levels of redundancy, thereby improving the availability of the whole system. For example, a failure of a system at one repository would not severely impact users at other repositories. In fact, users at the affected repositories would continue to search and discover reports from other repositories, though they may not be able to see reports that are added to the system at other repositories during the downtime.

The OAI-PMH also provides an interface that exposes the “hidden” information to general Web search engines. Other services such as cross-archive citation linking are also emerging.

**Related Projects and Improvements**

The construction of the Arc system motivates studies on registration service, repository synchronization problems, metadata quality, and scalability. The architecture of Arc is generally designed so that different modules can be plugged in and tested, and, when appropriate, they are incorporated into the Arc system and open-source software.

As of 2004 we have been involved in several projects toward improving Arc in various aspects. Archon is a project we launched in response to the problem of how to build rich service based on complicated metadata; DP9 is an effort to make OAI-PMH repositories open to general Web crawlers; and the DL Grid project addresses the scalability problem of a large meta-
data harvesting system by using DL Grid technology. We will discuss each of these in more detail below.

**Archon**

Funded by the National Science Foundation (NSF), Archon is geared toward building a federated digital library focused on physics with support of rich metadata (Anan et al., 2003). By design, OAI-PMH does not provide support for services beyond basic harvesting of metadata records. Due to the different quality of services and metadata implemented by various providers, the challenge exists to provide rich services besides simple keyword searches. With the Archon project, we have tried to address providing value-added services like citation and reference processing, equation searching, and data normalization to the process of dynamic harvesting. In order to improve several aspects of the Arc system, Archon focuses on:

- harvesting and parsing richer metadata formats and full text besides DC
- automatic citation linking across heterogeneous digital archives (APS, arXiv, and other physics collections)
- normalizing and automating the generation of missing metadata fields (for example, subject)
- equation searching—displaying and searching equations that are embedded in metadata fields such as title or abstract in formats such as LaTeX or MathML. Equations in these formats are not easy for users to browse or view.
- post-processing after search—the varying quality of metadata makes it difficult to build a unified search interface. Post-processing provides an alternative way to take advantage of richer metadata sets without complex search interface.

The development of the Archon system makes it clear that a rich search environment can be developed in a metadata harvesting system. However, it also reveals a lack of standards in this environment, such as metadata formats, controlled vocabulary, citation and reference information, and standard equation expression, which places the burden on service providers to understand all proprietary formats.

**DP9**

DP9 is an open-source gateway service that allows general search engines (for example, Google, Inktomi, etc.) to index OAI-PMH-compliant archives (Liu, Maly, Zubair, & Nelson, 2002). DP9 does this by providing a persistent URL for repository records and converting this to an OAI-PMH request to the appropriate repository when the URL is requested. Search engines that do not support the OAI-PMH can thus index the “deep Web” contained within OAI-PMH-compliant repositories.

Currently, indexing OAI collections via an Internet search engine is
difficult because Web crawlers cannot access the full content of an archive, are unaware of OAI-PMH, and cannot handle XML content very well. DP9 solves these problems by defining persistent URLs for all OAI-PMH records and dynamically creating a series of HTML pages according to a crawler’s requests. DP9 provides an entry page and, if a Web crawler finds this entry page, the crawler can follow the links on this page and index all records in a data provider. DP9 also supports a simple name resolution service: once an OAI Identifier is given, DP9 responds with an HTML page, a raw XML file, or forwards the request to the appropriate data provider.

Various caching mechanisms are also implemented in the DP9 service to make it more crawler friendly. However, due to the limitation of Web crawling technology, there is no guarantee that a document can be indexed by any Web crawler. Further research, such as the mod_oai project (http://www.modoai.org/), is underway to ensure better integration between OAI-PMH and Web-crawling technology.

Digital Library Grid

When dealing with a large number of data providers and documents, we discovered that it is necessary to parallelize each individual module of the Arc system, such as metadata harvesting, indexing, and searching. This leads to the Mellon-funded Digital Library Grid project (http://saturn.seven.research.odu.edu/grid/index_new). The objective of the Digital Library Grid project is to develop a high-performance federated search service that exploits the resources of a grid. It will make available a large amount of information that is distributed amongst heterogeneous digital libraries. In this project, we are developing the software tools to

- adapt Arc and Lucene indexing software (http://jakarta.apache.org/lucene/) to the grid
- deploy a cluster for parallel, high-performance searching based on Lucene
- develop software support to move indices and metadata between low- and high-latency nodes

In this project we propose to distribute the cost of publishing to collection builders (data providers), distribute the cost of harvesting and indexing to existing grid nodes, and only leave the cost of maintaining the federated search service to one institution (service provider), thus making it more sustainable. Since grid nodes by definition have unused capacity, no new hardware needs to be acquired, and we can, in essence, piggyback the onus of maintaining the infrastructure on the efforts to maintain the grid. The second advantage of this approach is availability of the service. The current Arc is running on a single processor without any redundancy. In the new approach, we plan to use hardware redundancy by exploiting the grid technology. For searching, we plan to exploit parallelism by partitioning
the indices amongst a cluster of PCs. A user query will be executed in parallel across these partitions, resulting in high performance. For supporting parallel indexing and searching, we will extend the open source Apache Jakarta Lucene search engine.

**Conclusion**

The Open Archives Initiative Protocol for Metadata Harvesting has transformed the way Web-based digital libraries are built. Originally conceived as a way of federating e-print repositories, OAI-PMH is now used by a variety of academic, government, and even commercial publishing interests. It has proved to be useful even beyond “document-like objects.” For example, Van de Sompel, Young, and Hickey (2003) discuss how OAI-PMH is used for thesauri, Web-usage logs, and an OpenURL registry.

The Arc service provider traces its roots to the original UPS prototype system that was featured at the initial workshop in Santa Fe. Arc debuted as a public service in 2001 and has been in continuous operation since then. It now indexes nearly 7,000,000 records harvested from several hundred repositories. Arc is available as an open source and has been deployed by a number of different institutions. It has also served as a platform for many other OAI-PMH projects, including Archon, Kepler, and DP9. Arc was the first service provider to introduce “hierarchical harvesting,” which formed the basis for what are known as OAI-PMH aggregators. Because of Arc’s immense scale, it has informed the community on a number of issues related to synchronization, scheduling, caching, and replication. Arc is currently being used in an ongoing project to merge OAI-PMH digital libraries with grid computing.

**Note**

1. Further discussion regarding the synchronization of harvesters and repositories can be found in Liu, Maly, Zubair, & Nelson, 2003.

**References**


Collaboration Enabling Internet Resource Collection-Building Software and Technologies

STEVE MITCHELL

ABSTRACT

Over the last decade the Library of the University of California, Riverside and its collaborators have developed a number of systems, service designs, and projects that utilize innovative technologies to foster better Internet finding tools in libraries and more cooperative and efficient effort in Internet link and metadata collection building. The open-source software and projects discussed represent appropriate technologies and sustainable strategies that we believe will help Internet portals, digital libraries, virtual libraries, library catalogs-with-portal-like-capabilities (IPDVLCs), and related collection-building efforts in academia to better scale and more accurately anticipate and meet the needs of scholarly and educational users.

Our work and its intent is best introduced by providing an overview of the projects, services, and software that we have been working on for the last several years: iVia, INFOMINE, and Data Fountains. iVia will be described in depth from the standpoints of its overall system, content and uses supported, end-user features, content development and management features for institutional collaborators, features for individual expert content builders, and incentives for collaborative collection building.

iVia

iVia (http://infomine.ucr.edu/iVia/) is a portal or virtual library collection-building software platform (Mitchell et. al., 2003). It was designed to support multiple institutions and projects in collaborative collection-building efforts. The system (or components) is used by INFOMINE and
the National Science Digital Library (NSDL) of the National Science Foundation, among others. The software, written primarily in C++, is licensed as open source and is available to all. iVia features a very large number of custom-configurable user interfaces and information retrieval options to support the institutional identity management (that is, branding) and user finding needs of diverse, collaborating organizations. Institutional collaborators will also be able to avail themselves of multiple metadata creation options, including support for multiple “production lines” and levels of editorial control. Resource- and labor-saving machine assistance is featured and used to semi- and/or fully-automate a number of tasks in both Internet resource identification and metadata generation. The

Figure 1.
former is made possible through new work in focused crawling and the latter through innovations in automated classification (which include the assignment of Library of Congress Subject Headings [LCSH] and Library of Congress Classifications [LCC]). iVia support has come from the Library of the University of California at Riverside, the U.S. Institute of Museum and Library Services (IMLS), NSDL, and the Fund for the Improvement of Post-Secondary Education of the U.S Department of Education (FIPSE).

INFOMINE

The INFOMINE (http://infomine.ucr.edu) virtual library service was conceived from inception as a multi-institutional, collaborative effort and has served the academic community since 1994. It has the mission of identifying, describing, and therefore making visible and useful to the academic community the significant scholarly and educational resources on the Internet. More than 230,000 resources populate the collection. These represent all major academic research disciplines and are the product of the collaborative efforts of librarians, faculty, and graduate students at the University of California (Riverside, Los Angeles, Santa Cruz, and Irvine campuses), Wake Forest University, and California State University (Fresno and Sacramento campuses).

INFOMINE draws upon a hybrid collection design that consists of metadata created by (1) subject experts (at INFOMINE and at collaborating institutions); (2) machine processes or machine processes with expert refinement; and (3) external collaborating institutions that share data streams of records, which are imported through OAI-PMH or other means, translated as needed, and then added to the INFOMINE collection (for example, MARC records of the University of California Shared Cataloging Project and Dublin Core records from some collections within the NSDL). INFOMINE represents a rich collection of records with rich metadata. For example, the number of subject and keyword terms applied in expert-created records that describe resource themes are much more numerous than in standard library catalogs. INFOMINE is used for both end-user searching and collection development on the part of other Internet portals, digital libraries, virtual libraries, and library catalogs-with-portal-like-capabilities (IPDVLCs). It uses iVia software as its system platform. INFOMINE support has come from the Library of the University of California at Riverside and the collaborating libraries mentioned above, as well as from IMLS, NSDL, and FIPSE.

Data Fountains

Data Fountains (http://infomine.ucr.edu/Data_Fountains/) is an open-source software system and a service for automated or semi-automated Internet resource discovery and metadata generation. Based in the
iVia system, it expands beyond iVia considerably by creating an array of independent, though federated, collection-building systems for collaborating projects with the goal of generating the basic “ore” (links to important Internet resources and associated metadata records and rich full text) for these projects. It also improves upon core crawling and classification techniques. Each collaborating project and/or subject community works with and fine tunes its own Data Fountain, that is, its own set of focused crawler(s) and classifier(s). The records and full text derived are exported to and utilized within the collaborator’s own native interface, backend system, and databases. In iVia these crawlers and classifiers are shared, as is the backend. Expert-machine interaction, which relies upon the subject domain expertise and the wisdom and conventions in collection building of participating librarians, is emphasized more in Data Fountains than currently in iVia and should result in more accurate content. That is, semi-automated approaches are more fully designed into and featured in the system and are critical to improving its performance. Given that Data Fountains is currently under development, much of the following instead addresses iVia, its close relative. Data Fountains work is supported by IMLS and the Library of the University of California at Riverside. Please contact us if you are interested in implementing Data Fountains in your project.
We designed the technology behind the INFOMINE, iVia, and Data Fountains projects to enable and facilitate cooperative service building and effort. That is, we wanted the technology providing the foundation for these systems to be collaborative and participatory and to gain significant increases in accuracy and resource savings through this. While the system strongly supports fully automated and fully manual processes for collection building, the technology also supports semi-automated processes emphasizing interactive subject domain expertise. We see our work as building machine-assisted IPDVL community-ware. We are developing and bringing to the library community new, machine learning–based technologies that are

- Enabling: These technologies provide systems that scale better in the Internet environment and save expert labor and other resources. They
enable collaborative efforts of many types at the same time that they are supportive of multiple modes of collection building and user access. These technologies also enable us to reduce redundant effort by better distributing collection and metadata development efforts among similar projects.

- **Participatory:** Collaborating institutions, as co-designers, participate in developing and customizing the software to fit their needs (for example, in interface, data views/landscapes, record creation, and retrieval). Collaborators work in codesigning systems that emphasize identifying, enhancing, and/or developing synergies among collaborating projects. This is done as well by identifying promising expert/machine processes/interactions that will augment and improve the performance of both. Experts actively participate in improving machine processes and vice versa.

- **Supportive of Librarian Community Expertise, Values, and Effort:** These technologies help amplify and facilitate the transfer of academic librarian subject expertise, public domain orientation, objectivity, service orientation, and other scholarly and educational community values and capabilities into efficient and effective Internet-based information. Tools such as iVia allow us to build very useful collections that are based on and express our considerable wealth of knowledge in subject domains, fully featured interfaces, sophisticated (that is, precise) user access, and rich, well-organized metadata. While Google-level accuracy and approaches suffice for many information-finding needs, they do not generally serve the in-depth finding needs of academics. Google may partially “disintermediate” the role of the expert librarian in some areas, but, in the long term, this will not extend to areas where superior information quality, sophisticated access, and accurate provenance verification are critical to major research and fact-finding efforts. It is incumbent upon the library community to work with this technology, to adapt it to its needs, and to come to own it just as physical collections usually own the facilities in which they are located. This is what our projects are about: bringing public domain community-ware and machine-learning technology in resource discovery and metadata generation, among other areas, into the library.

**Focus on iVia—An Open-Source Software Platform for Collaborative Internet Collection Building**

**Hardware**

The following hardware supports the INFOMINE application of iVia:

- Public search interface server: end-user and content-builder (including expert-guided crawler) interfaces are supported
- Public search interface server backup
• Database server (both the metadata and full-text databases are here)
• Database server backup
• Crawler/classifier processes server (for example, vlcrawler, Nalanda iVia focused crawler)
• OAI import/export server
• Additional mass storage equipment: 2 terabytes of storage including a RAID array (1 terabyte of storage) accessible via Network File System (NFS) (networked storage)

A standard machine would be an AMD XP 3200+ CPU, 1.5 GiB of high-speed RAM, and an 80GB disk storage.

Software

iVia software is licensed as open source (GNU GPL and LGPL). Open-source software is free software intended to be of use to and be further developed and refined by its users. In iVia’s case this would be users in the library and Internet Portal community. The open-source approach enables institutions to pool resources and inexpensively develop and refine software that meets their needs. In fact, in addition to the software we have developed, our system is based on many very successful and well-known open-source packages, including the Linux operating system (including Debian, RedHat, and Suse variants), MySQL and Berkeley DB databases management packages, and Apache Web server software.

iVia code is in C++, this being one of the most powerful, flexible, and standardized of programming languages. Some of our interface code is in Java. Currently the iVia program size is close to 10 Mb (>230k lines).

Standards

iVia is based in standards. Metadata standards include Dublin Core and MARC (we use Dublin Core but can translate from/to MARC). Subject schema standards include Library of Congress Subject Headings (LCSH) and Library of Congress Classifications (LCC), these long being standards in the U.S. academic library. Using these will eventually allow iVia, as finding tool software, seamless subject access (no translations involved) to both the Internet and print records of knowledge. For data transfer among collaborators, iVia uses the Open Archives Initiative (OAI-PMH) approach as well as standard delimited formats (SDF). OAI-PMH is used as well internally to transfer/harvest records from our crawling and classification databases and our user databases.

Fields Supported

Forty-seven fields are supported in our database. Of most direct value to users are URL, title, alternative title, creator (author), subject—LCSH, subject—LCC, keywords, description, selected full-text (1–3 pages of rich text), MyI (a field that helps institutions create custom data views), and lo-
cal URL (often of value for collaborators in accessing fee-based material). Other fields of note and their functions are general subject categories (for example, biological, agricultural, and medical sciences); created at; created by; modified by; last modified by; access restrictions; restricted to; publisher; audience levels; resource types; language; coverage begin; and coverage end.

Content Managed

Format types represented through iVia include HTML resources and, shortly, PDF, Postscript, and others. Metadata as well as representative, rich full text is generated or harvested from the resource being described and makes up the content of our databases. This data represents free and fee-based resources and includes resource types as varied as digital libraries, other virtual libraries and portals, e-journals, e-books, e-print archives, databases, hypertext fiction, maps, and more. Content retrieval is robust and quick. Berkeley DB indexing capabilities are used to augment performance through MySQL.

iVia Uses

Major applications of iVia to date have included INFOMINE, one of the first Web-based services offered by a library. INFOMINE (an Internet resources virtual library–type finding tool) has been supported by iVia in serving academic researcher and student end users both nationally and at specific institutions (for example, the University of California at Riverside and Wake Forest University). Collection development for others has been another major function, with many other academic virtual libraries using iVia/INFOMINE as a resource discovery service for their own collection-building efforts. iVia/INFOMINE is also used by librarians in creating Web-based subject guides or pathfinders in various subjects (this is facilitated through using our “canned search” generator and MyI field), as well as by faculty creating Web resource modules on their course pages in support of curriculum units.

While INFOMINE has been the major application of iVia so far, with most aspects of iVia as described in this article being applied in INFOMINE, we have been working with the National Science Digital Library (NSDL) to develop an NSDL iVia. Among the major goals of this project are the integration of our Web crawlers and classification software into NSDL’s core system for purposes of open Internet resource discovery and related classification (that is, resource identification and metadata generation). Just as crucial here will be the use of this software to generate metadata for existent, “deep Web” collections (for example, article databases or e-print collections or other databases where access is through a search front-end) in many different document formats other than HTML.
iVia User Features

Through the INFOMINE application, iVia has demonstrated sophisticated and flexible user features geared toward varying levels of searching expertise. Most searchers will use defaults that are transparent to them as they use the basic search (http://infomine.ucr.edu/). Librarians, information specialists, and researchers may choose to use the many user configurable features found in Advanced Search and Browse (http://infomine.ucr.edu/cgi-bin/search). Advanced Search and Browse features are present in each individual collection (for example, http://infomine.ucr.edu/cgi-bin/search?category=bioag).

In more detail, iVia’s search and browse features include the following: multiple subject and resource type collections or categories, including Biological, Agricultural and Medical Sciences; Business and Economics; Cultural and Ethnic Diversity; E-journals; Government Information; Maps and GIS; Physical Sciences, Engineering, Computer Science, and Math; Social Sciences and Humanities; and Visual and Performing Arts.

The availability of standardized, fielded metadata, as well as rich full-text, enables advanced searching capabilities including Boolean (for example, and, or, not) and Proximity operators (for example, near 1–20); exact searching using quotes or stem searching using asterisk; nested searching using parentheses; and various types of limit searching. One can limit to expert or expert plus robot-originated records (the latter being those that have been automatically identified and described), or combine general subject categories (for example, BioAgMed or E-journals), any combination of fields (for example, title, keywords, subjects, and/or description, and so on), resource type (for example, article databases, electronic journals, or e-print collections), and/or type of access to resource (such as free, fee, or a mix).

In iVia, search interfaces are presented on the bottom of each results page if search modification is desired. In the event of zero result searches, spelling is checked and possible spelling alternatives are suggested. Finally, in full display, most indexing terms are presented as links, which can be clicked on to narrow or broaden a user’s search.

Browse indexes are available for both all subject categories and individual subject categories. Specific browse indexes are available for titles, creators (including authors), subjects—LCSH, subjects—LCC, keywords (these often include minor subjects and lay-person terminology), resource types (for example, standards, style manuals) and Whats New! (that is, recent expert additions to the collection).

Records are displayed in three formats: title only, regular (title, description, and origin of record as either expert or robot created), and long (accessed by clicking on “More Info” in the full display). The latter includes a great number of fields of interest to users or collection builders including URL, title, description, broad subject categories, creators, subject—LCSH,
subject—LCC, keywords, access, audience level (academic, K–12, or lifelong learner), institutional owner (which collaborator contributed the record if expert in origin), URL checker information, and INFOMINE collection information (mostly for record keeping: who added, who modified, record number, record origin). Results pages can be displayed in groups of 30, 50, or 100. They can be ordered alphabetically by title or by relevance to the query as judged by how many query terms were hits, how many were hits in major or minor fields (for example, title being more highly weighted than keyword, which is more highly weighted than full text), and whether terms in a specified phrase were found in exact or approximate adjacency.

iVia Content Development and Management—Features, Tools, and Machine Assistance for Institutional Collaborators and Expert Content Builders

iVia emphasizes numerous innovations for improving and making more efficient collection development and management efforts for both individual or multiple collaborating projects. These translate into significant labor and resource savings in building collections. These innovations can be best understood from the standpoints of institutional collaborators and individual experts creating new content, as detailed below.

Support for Institutional Collaborators

Institutional identity management or branding is important for iVia collaborators. Access to collaborative resources needs to reflect, within reason, the established ongoing Web presence and interface of the collaborating institution. To this end iVia provides multiple interfaces and methods of accessing data in collections it supports. The user interfaces and desired data views of collaborator project sites are supported. For example, the interface that the user is accessing from can be detected by iVia, which activates searching and other interface capabilities that meet existent profiles set up for this by the collaborating institution. Access is also enabled for selected external collections that rely on metasearching.

Custom Data Views and Access Supported iVia provides pre-constructed interface modules that can be quickly assembled and customized by collaborators in building interfaces to iVia data. These interface modules reflect the themes and presentation of the collaborating project while still taking full advantage of unique iVia retrieval and other user features. The suite of programs that facilitate this is known as “Theme-ing.” Special fields, such as MyI (which allows institutions to create custom data views), support Theme-ing and custom interface access. For example, retrieval filters can be created by participating institutions to channel user searches through selected subsets of iVia data (for example, perhaps only the records for fee-based resources in the collection that have been subscribed to by the
particular institution). This is done by identifying and tagging, in the MyI field, those records that the institution wants its users to view. Parallel fields are also supported for similar reasons. For example, some collaborators want short descriptions and others long. Hence, there are two, parallel, description fields. Users coming from the institution desiring short descriptions will see only these.

**Metasearching Access** iVia also enables access to its content through the interfaces of selected, completely external finding tools, which rely on general methods of metasearching. For example, the Ex Libris online public access library catalog system provides access to INFOMINE content, as does the California Digital Library Searchlight system. The nice thing about metasearching is that large numbers of diverse collections from multiple projects can be searched simultaneously. However, significant downsides exist because of the need to include generally very simplified, lowest common denominator searching of only the shared fields among the databases searched, which can be very few; this eliminates search access to unique, useful fields. Another problem is the limited ability to eliminate duplicate, overlapping results returned from the databases searched.

**Multiple Modes of Content-Building Supported** Even if collaborating institutions have been building Internet resource collections for some time and have established ways or styles of doing things, iVia takes this into account by providing multiple means for new collaborators to ramp up and begin creating content in ways with which they are comfortable. To this end iVia supports from one to three levels of editorial review as well as a pending record database that holds records in the process of being built and reviewed prior to their being approved and moved to the main working database. Some collaborators use just one level of review, that of the editor of the subject file (for example, the BioAgMed file in INFOMINE). Others have developed a well-defined division of labor whereby catalogers review the subject content of records created by public service librarians or metadata specialists prior to review by the editor of the subject file.

Similarly, in support of various divisions of labor and optimum utilization of staff with varying skill sets, each content builder can be assigned a different level of access to iVia content-building features. Managing editors of a subject file have full permission of many kinds, including batch deletes and batch changes, to the content of the whole database. Metadata specialists, on the other hand, may only be allowed to add content to the pending record database, with their records going through multiple levels of review before being added, by the subject file editor, to the working database.

**Hybrid Collections of Heterogeneous Metadata—Support for Multiple Incoming Data Streams and Types of Records** Just as one of the main benefits of collaboration in mutual content building is sharing the collection development load among participants, iVia also makes it possible to utilize the work of other collection-building projects that choose to not be an integral part
of the project. To do this, iVia has a hybrid collection design that supports diverse, heterogeneous record types and record origins (Mason, Mitchell, Mooney, Reasoner, & Rodriguez, 2000).

As manifested in the INFOMINE application of iVia, the system builds content by ingesting and threading together a number of diverse data streams. The first of these is, of course, the records created within the iVia system by experts. Sources for these currently include content builders from the University of California at Riverside, UCLA, and individuals from other UCs; Wake Forest University; and California State University at Fresno and Sacramento. There are about 20,000 of these expert-built records internally created for and through INFOMINE’s iVia system. INFOMINE’s iVia also imports and, as needed, translates from collaborating external data streams. For example, MARC records for Internet resources cataloged by the UC Shared Cataloging Project (SCP) are imported, translated to Dublin Core, and utilized (about 25,000 records in INFOMINE are of this origin). Through collaborators at UC Santa Cruz, Lexis Nexis serial titles are imported (accounting for close to 6,000 records). INFOMINE’s iVia also uses OAI-PMH to import records from selected NSDL-associated collections (about 10,000). In INFOMINE, there is a total of close to 60,000 expert-created records either of internal origin from closely allied institutions or that have been created externally by sharing institutions and imported. All of these expert-driven data streams form a first tier of records in the architecture of iVia.

The second-tier collection supported by iVia consists of records that have been created automatically by crawler/classifier robots. There are also records that are of robot origin but that have been refined, augmented, and vetted by experts. This is an example of semi-automation with experts receiving machine assistance in resource discovery and metadata development. Currently, there are three crawler/classifiers (to be described below) that have created over 170,000 records. As in Google, these records, while far from MARC perfect, remain very useful and have been created relatively inexpensively. In the architecture of iVia they form a large second-tier collection that is used to support the first-tier collection of expert-built records. Complemented by the 60,000 expert-created records, INFOMINE’s total collection size is around 230,000 records and growing rapidly.

Importantly, the content of iVia records ranges from just metadata to metadata augmented by selected, rich full text that has been robotically harvested from the resource itself. Judicious use of full text is of great help to user retrieval by drastically increasing the amount of material that can be searched and therefore the granularity or detail in searching that can be supported. Full text also helps correct for controlled subject vocabularies that are often too removed from common parlance and/or too general or specialized to adequately serve a wide variety of user audiences.

The collection designs discussed above have been very successful. They
have been able to reflect and provide intelligent organization and access to content from many different sources and of many different types. In a world of multitudes of important collections and approaches to metadata, the iVia hybrid collection approach has been very useful for end-user access.

Support for Expert Content Builders

Just as iVia provides means for facilitating and aggregating the mutual efforts of multiple institutions, it also provides a great amount of time saving, machine assistance, and other means of expediting the work of expert collection builders. Machine assistance is provided in new resource discovery (that is, collection development), metadata generation (that is, indexing), and in a great number of smaller collection-building tasks.

Machine Assistance through Automated and Semi-Automated Resource Discovery

Automated and semi-automated resource discovery (that is, collection development) is a major boost in collection building and saving the time of experts in finding relevant new resources. iVia uses several Web crawlers to scour the Web (or selected parts of it) to identify scholarly and educational resources of interest (Chakrabarti, 2003). The crawling technology can run fully automatically, but it has been built to include important roles for experts in guidance, refinement, and truing. For example, experts work with the crawlers to monitor and adjust resource acceptance weighting thresholds or the criteria by which a crawler will identify a resource as relevant. Screening for duplicates or resources already in the database is a perennial challenge. This is done through automated means as well as through experts monitoring lists of potential duplicates found through either exact or fuzzy matches of title and URL information. For irrelevant sites that keep re-occurring in crawls, iVia content-builder community blacklists are maintained that prohibit future crawler visits.

For custom, finite crawls, we have built crawlers that are fully expert guided in the sense that well-defined crawling targets are provided by experts and crawling occurs in a very directed manner. iVia’s “Expert Guided Crawler with Drill Down/Drill Out” takes expert-provided individual or multiple URLs and crawls them. Experts specify the number of levels down into a site that should be crawled (most sites being organized hierarchically) as well as the distance of other sites linked to from the expert-provided site that should be pursued (for example, options are one to two jumps from the original URL). This semi-automated crawler gives the expert the ability to “mine” for new resources/links in a very precise way. A single page or site can be crawled, or a community of closely linked sites can be crawled. Likewise, we are building a focused crawler that will take a topic that is very well defined by experts and concentrate on just that topic. This is a semi-automated focused crawler that will be dependent on feedback and truing from participating experts for best results.

Just as experts interact with and improve crawler processes and ac-
accuracy, the interaction can be reversed with crawlers suggesting the most promising of sites as needing expert attention from content builders. That is, the most highly weighted sites that are automatically included in the crawler collection are flagged for expert review and refinement. Similarly, iVia database and record usage statistics are kept so that the most used or visited records of crawler origin can be flagged for expert attention, whereby the automatically created metadata present can be improved. Such a record is then moved from the second-tier, robot-created collection to the first-tier, expert-created collection. These are both important collection development tools and provide useful assists for experts.

**Machine Assistance through Automated Record/Metadata Generation or Import** Automated and semi-automated metadata generation provides expert content builders with a great advantage (Chakrabarti, 2003; Frank & Paynter, 2004). Collection size and depth is greatly improved through records created in these ways. Specifically, iVia’s second-tier collection of records, those that have been created fully automatically, provides a great boost for the utility and value of the collection as a whole to users and greatly augments and complements expert content-building work. At the same time, the existence of automatically created records provides great assists for expert record-building activities when they are viewed as “foundation records” or records that have been partially built (from a librarian standpoint) and that can be improved upon through some expert effort. Working with these automatically created records as foundation records and improving them saves expert time compared with creating records from scratch. Foundation records can be seen as the basic “ore” that can be easily refined for more demanding or discerning uses where more rigorous (though more expensive) metadata may be the norm.

Expert content builders are also aided, as mentioned above, by iVia’s ability to import and share records with other collections though OAI-PMH and standard delimited formats. This also contributes to boosting collection size, depth, and value for the end user.

**Specific Machine Assistance to Experts in Record Building** Numerous small machine assists are supplied by iVia to make expert record building more efficient. In the aggregate, these are crucial and save much expert time. For example, iVia supports

- Duplicate checking: prior to building an expert record, the iVia checker finds both exact and fuzzy matches within the URL and title fields for experts to review. Also identified and deleted, by checking exact lengthy character strings, are mirror sites.
- Record cloning: multiple records can be built representing closely related sites, authors, or organizations. Similarly, multiple records on the same or related subjects can be cloned and the subject and keyword indexing, among other metadata, saved and re-utilized.
Batch editing: just as multiple records can be imported or exported in batches, their metadata can be edited and changed globally in batches. This saves much time in cases, for example, where a convention on naming a resource type has changed.

• URL Canonization: variants of URLs are canonized to proper form when this is needed.

• URL change notification: always a challenge is keeping up with changing URLs. To do this iVia has developed a “URL Checker and Pursuit” utility that flags problem URLs, notes the nature of the problem, notes potential locations indicated by forwarding messages, and (after three consecutive failures of a URL over a period of three weeks) flags the editor of the subject file with the record with the problem URL and suggests possible working URLs.

• Pull down menus of various controlled vocabularies: these would include resource types, keywords, and broad subject disciplines.

• User corrections/suggestions/new content: these are encouraged and funneled to content builders. This has been a major source for identifying possible new content and correcting errors.

• Online and point-of-need guidance: help is provided via manuals, style guides, and pop-up screens with pointers.

• Collection development assistance: this is supplied to other collections through iVia’s email-based “New Resources Alert Service” and through the Whats New! index.

Under the Hood

The techniques, approaches, and algorithms that make machine assistance to experts in collection building and, more generally, iVia possible are described in more depth at the iVia site, http://infomine.ucr.edu/iVia.

A Collaboration-Inducing System

There are a number of catalysts that should stimulate increasing collaboration with iVia and its participants. The foremost is that, working together, a powerful, far-reaching, and high-quality finding tool and both internally developed and allied, externally developed collections, with proven value to researchers and students, will continue to grow and thrive. Working together, collaborators reduce redundant efforts by sharing and distributing collection development tasks and by unifying system building and support activities. Collaborators participate in a state-of-the-art system incorporating resource-saving machine assistance in numerous tasks.

Furthermore, the iVia system is in the public domain, free, and open to custom development. At the same time, iVia and the collections it provides access to can be utilized through custom interfaces and data views that meld well with the Web presence of the collaborating institution. Additionally, as one of the first library-based Web services, iVia/INFOMINE developers
have a great deal of experience in meeting scholarly Internet user finding needs. Finally, the collections that populate iVia through INFOMINE are significant, well-organized, and useful. INFOMINE is among the largest librarian-built collections of its type.

**Summary**

iVia is a powerful and flexible, collaboration-enabling, open-source, Internet collection-building, and finding tool system. It is of use in building Internet collections of metadata and full-text data representing resources from the Web as exemplified through INFOMINE, one of the earliest and more significant of academic virtual libraries. The metadata generated includes library standard subject schema. iVia supports single or multiple subject focuses as well as both single or multiple institutional efforts. It is intended as community-ware and has proven itself to be of value in multi-institutional collaborations such as INFOMINE, NSDL, iVia, and, shortly, Data Fountains. User retrieval options are numerous for both fielded and full-text data and support both beginning and advanced searchers. iVia supports custom branding, interfaces, and data views for those accessing its collections. Numerous modes of content building are possible featuring varying levels of editorial review, styles of indexing, and divisions of labor. iVia is noteworthy because it saves resources and labor by integrating fully automated, semi-automated, and fully manual modes of record building. Resource discovery through various iVia Web crawlers and metadata generation through iVia classifiers (and other means) results in collections that require fewer resources and less expert labor to reach significant size. iVia emphasizes collaboration and empowers the librarian expert through the use of machine assistance.

**References**


EDWARD ALMASY

ABSTRACT
Creating a full-featured resource portal on the Web is no small task, and it can be even more of a challenge without a team of Web designers and programmers. In the fall of 2000 the University of Wisconsin–Madison’s Internet Scout Project (Scout) received funding from the Mellon Foundation to build an open-source software package intended to enable collection developers to share their collection’s metadata via the Web. In October of 2002 Scout began a new effort, funded by the National Science Foundation (NSF) as part of the National Science Digital Library (NSDL) initiative, to build upon prior work and create a software package that would help STEM (Science/Technology/Engineering/Math) content authors and collection developers share their work online and integrate it into NSDL. The software packages resulting from these two projects, the Scout Portal Toolkit (SPT) and the Collection Workflow Integration System (CWIS), are very inexpensive to maintain and operate and easy for nontechnical staff to download, set up, and populate with metadata. Conforming to international standards for metadata, data harvesting, and Web technology makes SPT and CWIS useful for and usable by a wide variety of projects and organizations, allowing and encouraging collaboration and record sharing among projects.

In today’s Internet, with information overload prevalent even within a single discipline, scholars and researchers struggle to find the precise material they need in the tangled web of online information. The major search engines do not offer great precision or any guarantee of authority;
the best sites in a given field are spread around the nooks and crannies of the Internet and need to be located and then individually searched for relevant information; and even topical electronic mailing lists can require substantial effort to monitor and sift through to extract useful information. In addition, most scholars and researchers lack the extra time needed to roam the Web trying to stay abreast of all the new resources and tools that, ironically, could save them time by making the task of locating useful information easier.

In some disciplines this problem is being addressed by organizations that take a leadership role by building Web sites called “subject gateways” or “discipline-based resource portals.” These Web sites usually focus on a specific topic or scholarly discipline, and they often provide information in a variety of forms and from many sources. For example, a discipline-based portal may feature the following:

- a browsable directory of online resources, described and arranged by subject
- a search facility that includes only resources related to the field and that allows searching by title, author, subject, etc.
- current news stories related to the field
- forums for discussing specific discipline-related issues
- facilities for scholars to comment and share information about specific resources

By bringing together various collections and access points into one integrated Web site, a resource portal can bring coherence to the body of online information available in a given field of study, providing scholars and researchers with a facility that will save them substantial time and increase their awareness of other work in their field.

Given all of the above, a discipline-based resource portal sounds like a fine thing to put online, but building a high-quality portal with even a portion of these facilities can be a daunting undertaking. Although the benefits of setting up a resource portal are clear, many organizations with a strong focus on a particular discipline do not have ready access to extensive technical resources, and even those organizations that do are likely to have those resources already committed to existing projects or working to support the organization’s day-to-day operations.

The Scout Portal Toolkit (SPT) and the Collection Workflow Integration System (CWIS), open-source software packages developed by the Internet Scout Project under grants from the Andrew W. Mellon Foundation and the National Science Foundation (NSF), respectively, were created to address this problem. They allow a group or organization (or even an ambitious individual) to share a specific knowledge base via a full-featured portal on the Web, with little or no investment in technical resources or infrastructure. In fact, many groups and organizations already have avail-
able the minimal resources needed to put a resource portal online using SPT or CWIS.

The Internet Scout Project (Scout), based in the Computer Sciences Department at the University of Wisconsin–Madison, was started in 1994 to develop better tools and services to find, filter, and present online information and metadata. Scout’s flagship publication is the Scout Report, an electronic periodical that is published weekly and identifies and describes the best new online resources of interest to educators, students, and researchers. In 1996 the content from the Scout Report and related Scout publications began to be collected into a searchable and browsable online database labeled the Scout Archives. As the Archives grew for the next six years, two things gradually became apparent: (1) that the original Scout Archives infrastructure had outgrown its content and user base, and (2) that there were many other groups and organizations who had the subject knowledge and desire to assemble and share collections of online resources but did not have the needed technical expertise. In 2000, Scout received funding from the Mellon Foundation to develop a software package to meet these needs, and the Scout Portal Toolkit was born.

CWIS (pronounced “see-wis”) was developed after SPT; it was chartered by the NSF to build on prior work to provide software that would enable STEM (Science/Technology/Engineering/Math) content creators and collection developers to quickly and easily put their work online and integrate it into the National Science Digital Library (NSDL). Because new technology developed for CWIS has also been integrated into SPT, both SPT and CWIS have similar capabilities, differing primarily in that CWIS includes a user interface, default nsdl_dc metadata schema, and a number of other features designed specifically to help with integration into NSDL, while SPT comes with a user interface that is more easily customized and a less complex default metadata schema intended to be less intimidating to the metadata neophyte. For practical purposes we will refer just to CWIS for the remainder of the article, but all features and capabilities discussed as part of CWIS can be assumed to also be available in SPT, unless otherwise noted.

Defining and Cataloging

Setting up CWIS is intended to be (and usually is) a simple process. (Detailed information about hardware and software requirements, as well as where to obtain the software, can be found at the end of this article.) Once CWIS has been installed, the first task at hand to make the resource portal useful is the entry of resource metadata records.

This section will explain what types of data can be entered into CWIS and what tools are built into CWIS to allow for metadata addition and manipulation. The software is distributed with a set of sample data that is loaded during installation, so that administrators and resource editors can
see how the portal works with data in place. When it is no longer needed, administrators may easily delete this sample data and enter new data into the portal.

**Defining a Schema**

Out of the box, CWIS comes with a default metadata schema that includes the full Dublin Core Element Set (Dublin Core Metadata Initiative, 2003a) with several extensions taken from the qualifier sets defined by the DCMI Education (Dublin Core Metadata Initiative, 2004) and DCMI Administrative Metadata (Dublin Core Metadata Initiative, 2003b). This schema may be more than sufficient for many project and user needs. However, some groups will wish to modify these fields or add additional fields to better fit their specific workflow and needs. By using the Metadata Field Editor, a user with administrative privileges can add new fields. The nine basic data types supporting the creation of these new metadata fields are

- **Text**—A free-form field that may contain any textual data
- **Paragraph**—A free-form field that may contain any textual data. To ensure proper display and entry, a **Paragraph** field differs from a **Text** field in that the **Paragraph** field is expected to normally hold several lines of text.
- **Number**—A numeric field that may have limits imposed and can be compared to other values when setting up searching criteria
- **Date**—A field containing a date value or date range. Date fields can contain whatever level of precision is appropriate and support several additional attributes, such as a prepended “c” to indicate that the value entered represents a copyright date.
- **Flag**—A Boolean field containing a true or false value. Labels may be assigned to True and False for clarity.
- **Controlled Name**—A field containing an entry from a list of values that is maintained separately to ensure consistent naming. **Publisher, Creator, Contributor,** and **Subject** are examples of controlled name fields that are part of the default schema.
- **Option**—A field containing one or more of a number of attributes. **Resource Type, Language, Audience,** and **Format** are examples of option fields that are part of the default schema.
- **Tree**—A field containing one or more entries taken from a hierarchy of values that is maintained separately. A **Tree** field might be appropriate to contain values from a standard classification taxonomy.
- **Image**—A field containing an image uploaded to the portal

In addition to basic field attributes like type and name, the Metadata Field Editor allows an administrator to set default values and a number of type-specific field attributes, such as minimum and maximum value for numeric fields and on and off labels for flag fields. The Metadata Field Editor also
allows tailoring the performance of the search engine (discussed below) by indicating which fields to consider for keyword searches and how to weight the fields when ranking search results.

Once metadata fields are defined, the primary method of entering resource metadata into CWIS is the Metadata Tool.

**Entering Metadata**

The Metadata Tool is designed to speed resource cataloging and support the accurate and consistent assignment and recording of metadata required to build and maintain a useful discipline-based resource portal. It allows resource editors to add, edit, duplicate, and delete records. The Metadata Tool also provides special features that aid in resource management. Drop-down menus are available for option, controlled name, tree, and flag fields, and these menus speed entry of commonly used values and help keep metadata vocabulary consistent. Of course drop-down menus are only viable for fields with a modest number of choices so, for controlled name or tree fields with more than a few hundred entries, the Metadata Tool also provides searching and browsing interfaces for selecting those fields. For example, resources in the (Internet Scout Project’s) Scout Archives, which uses CWIS, are cataloged using a subject hierarchy (adapted from Library of Congress Subject Headings) with more than 20,000 entries. As it is impractical to display that many selections on a drop down menu, CWIS automatically switches to a searching/browsing interface for assigning entries to that field.

The Metadata Tool also provides fields that support workflow management and editorial review. For example, the default CWIS schema includes a Release Flag flag field, which defaults to *Not Okay for Viewing*. When a user is browsing or searching, this field is checked for each resource and, if it is set to *Not Okay*, the resource is not displayed, which allows editorial review of material before it becomes available to the general public. If editorial review is not desired, the default value for this field can be set to *Okay for Viewing*, via the Metadata Field Editor, and new records will be available immediately.

**Workflow Management**

To make use of the Metadata Field Editor, the Metadata Tool, or any of the personalized portal services, users must create an account on the portal with a login name and password. Once logged in, access to various features is controlled by eight permission flags, which may be assigned from any account with System Administrator access. These permission flags are:

- Personal Resource Administrator
- Master Resource Administrator
- Release Flag Administrator
- Controlled Name Administrator
Figure 1.
Classification Administrator
News Administrator
Forum Administrator
System Administrator

Users may be granted any of these flags independent of the others, allowing, for example, the portal administrator to designate certain individuals as responsible for maintaining the controlled name lists or classification hierarchies, while other individuals handle more administrative matters like monitoring activity in the portal discussion forums or posting news to the front page of the portal. This means that resources editors, and others involved with the portal, could potentially be spread out geographically and still effectively work together to contribute or edit resource records in a coordinated fashion.

SEARCHING AND BROWSING

Once the proper metadata fields have been configured and populated with initial resource metadata records, collection developers may be ready to share their efforts with end users. Of course, those users need some way of locating the specific information on the portal that best meets their needs, and CWIS provides several routes toward achieving that end.

Browsing Resources

The simplest and most familiar method for locating information on a Web site is browsing. CWIS supports hierarchical browsing interfaces, based on classifications (tree fields) assigned to the resources.

Since classification hierarchies may be either wide (a large number of entries at the top level) or deep (a large number of levels) or both, the CWIS browsing interface is dynamically generated, based on the structure of the classification tree. This prevents the browsing interface from becoming unwieldy when a given section of the tree is very broad, while still minimizing the need for users to search through multiple pages to find the classification they seek. Because resources may be classified with more than one taxonomy, CWIS generates a separate interface for each field marked as available for browsing and remembers which of these interfaces each user most recently selected to help users orient themselves more easily each time they log into the portal.

In most collections the resources will not be evenly distributed throughout the taxonomies used to classify and subsequently browse through the resources. To provide users with some idea of the distribution of entries through the classification tree, CWIS displays the number of resources present under any branch of the tree. This can prove particularly useful to collection developers in assessing where their collection’s strengths lie and where additional effort may be needed to round out the collection. Branches or leaves containing no resources are not displayed in the brows-
ing interface, which allows collection developers to enter or import a large taxonomy for use when cataloging resources without unnecessarily cluttering the selection presented to end users.

**Keyword Searching**

Because searching is often faster and more effective than browsing (provided the user has concrete insight on at least some aspects of the desired items), many people prefer to search rather than browse to locate data on the Web. CWIS provides two separate search mechanisms, both based on Scout’s OSMASE search engine.

*Keyword searching,* the method familiar to most Web users, is very similar to the approach presented by Google™ and other general Web search engines. Users enter terms that are related or may appear in the entries that they are looking for, and those terms are used to determine which resources best fit the search. CWIS supports most of the conventions offered by sites like Google, such as phrase searching (enclosing several words in quotation marks to indicate that the user is looking for the words in that specific sequence) and term exclusion (prepending a minus sign to a word to indicate that the user only wants results that do not include that term). Keyword search results in CWIS are ordered by their relevance to the terms entered. How various metadata fields are weighted to determine this relevance (for example, terms appearing in *Title* are more significant than those appearing in *Description*), as well as which fields are considered for keyword searching, can be adjusted via the Metadata Field Editor, allowing portal administrators to tweak search performance to best fit their portal content and audience.

**Fielded Searching**

To better take advantage of the precision offered by the metadata assigned to resources, CWIS also supports fielded searching. With a fielded search, users enter terms in a fashion similar to a keyword search, but along with those terms users can specify the fields in which to look for the terms. For nontextual fields, fielded searching also allows users to specify constraints that can be used to narrow the search results. For example, when searching through a collection of digitized rare books stored online, a user might be able to specify only nonfiction books that were published between 1885 and 1890.

Because of the potential complexity of a fielded search, CWIS provides the ability for each user to save a set of search parameters and recall them at a later date to run the search again.

**User Agents**

Combining this ability to save fielded searches with what has become known in Internet jargon as “push technology,” CWIS also offers a feature sometimes referred to as “user agents.” This capability allows users to set up
and save a fielded search that returns items they may find of interest and then have that fielded search automatically performed nightly or weekly by the portal, with any new results found being assembled into a report that is sent to the user via email. These reports can be invaluable to users by keeping them abreast in a timely fashion of new resources that may become available, and they benefit the portal developer by actively maintaining awareness of the portal among the user community.

For resource metadata administrators, CWIS also offers the option to run these searches on an hourly basis, which may facilitate editorial review or other workflow processes set up among a group of collection developers. For example, rather than requiring catalogers to notify editors when new records have been entered, the editors can be notified automatically of new entries awaiting their review by user agents they have set up on the portal.
Rating and Recommending

An active user community and its contributions are key components of most successful Web portals. To leverage user community participation, CWIS offers three features: resource ratings, resource recommendations, and resource comments.

Rating Resources

Resource rating allows users to indicate their opinion on the usefulness of an individual resource and to generate a cumulative rating for the resource based on these opinions. The cumulative ratings are beneficial both to other users, who can use them when determining which resources are most likely to meet their needs, and to the collection developers, who can use the ratings to monitor what portions of the collection users are finding most beneficial. Cumulative rating values are displayed graphically when browsing, searching, or viewing the full resource record, and users may interactively rate resources from the browsing and search result interfaces.

Getting Recommendations

Resource ratings provide information about the perceived usefulness of a resource to other users and the collection developers, but they can also represent a body of information about the needs and preferences of the users who assigned the ratings. To take advantage of this information, CWIS includes a recommender system.

One recommender system with which many people are familiar is that provided by Amazon.com™, where a user rates a number of books and then, based on those ratings, Amazon recommends other books that they may find of interest. The facility provided by CWIS operates in a similar fashion, although it is a content-based recommender system rather than a collaborative recommender system such as Amazon’s. A content-based system, which bases recommendations on item attributes, was chosen over a collaborative system (Breese, Heckerman, & Kadie, 1998) because collaborative systems, which base recommendations solely on preferences expressed by groups of users, typically require a very large number of ratings before they begin to offer useful recommendations.

Notes and Comments

As an adjunct to the resource ratings, SPT also provides users the ability to post comments on resources, which are then displayed along with the individual resource record. Again, this can benefit both other users and the collection developers by providing more detail about why users may have found a particular resource useful.

To help insure the integrity of these facilities, only registered portal users may rate resources or post comments, and an interface is provided to allow administrators to quickly review (and edit or delete) recent com-
ments, without having to step through every resource record to learn what has been posted.

**Displaying and Sharing**

Entering data into the portal and helping users find the portion of that data that meets their needs have both been discussed, but there still remains the problem of presenting that data to the end user in an effective fashion.

Effective presentation can vary widely depending on the subject matter and intended audience. Fortunately, CWIS provides several mechanisms that can be used to tailor a portal to meet these specific needs. These mechanisms do require some technical expertise, but they do not have to be employed to build a useful discipline-based portal. However, if the technical expertise is available, in some situations these mechanisms can be used to dramatically improve the portal experience for the end user.

**Multiple Interfaces**

Some portals may need to serve disparate user communities, presenting a different face or offering different features depending on the user. For example, a portal focused on educational resources about paleontology may want to serve both grade school children and high school students, but a Web site design that can catch and hold the attention of an eight-year-old will likely be judged by a sixteen-year-old as too childish or condescending, and site designs well-suited for either of those groups will likely not be optimal for use by their teachers.

To address this type of situation, CWIS supports multiple user interfaces, assignable on a per-user basis. In practical terms, this means a portal can have two or more user interfaces that differ significantly from one another in appearance and functionality while still using the exact same underlying CWIS installation, configuration, and metadata.

All CWIS pages for a given interface are built from a common CSS-based page template, to which page-specific code is added. If specific code for a page is not found in a new interface, then page-specific code from the CWIS default interface is used. This means that it is possible to dramatically alter the appearance of a portal site by adding a new interface containing just a new common page template and then making changes to that template. This approach also allows additional changes or additions to be made on a per-page basis to change the appearance or add functionality, without having to provide a new set of pages for the entire portal.

Some limited user interface customization ability is available in CWIS and SPT without modifying or adding HTML. For SPT the colors and logo graphics in the default interface may be set by the portal administrator via a Web-based customization tool. For CWIS the default interface comes with
Figure 4.
a half dozen additional “themes” that provide alternate color schemes and header graphics.

**Sharing Metadata**

Sometimes when presenting data, the intended recipient is a computer rather than a human being. To address this need, CWIS supports exporting data in three formats: RSS, OAI, and tab-delimited text. The OAI (Open Archives Initiative, 2004a) format is an XML-based protocol for harvesting metadata. Developed to be a low-barrier (that is, easily implemented) method for sharing metadata, the protocol has been adopted very rapidly by the online metadata community. CWIS supports version 2.0 (Open Archives Initiative, 2004b) of the OAI protocol, including qualifiers. RSS (Backend. Userland.com., 2003) is a well-established XML (World Wide Web Consortium, 2004) format for syndicating online content, typically conveying article titles or headlines. The first version of RSS was developed and released by Netscape in 1999 (Libby, 1999), and the format has since been adopted as a de facto standard (Syndic8.com, 2004) among weblogs and other Web sites where syndicated headlines are desired. CWIS supports RSS version 2.0 (which is backward compatible with RSS version 0.92). Although the Metadata Tool will likely be the primary method of entering new data into the portal, sometimes—particularly during initial setup—an administrator may want to import records into the portal in bulk. To allow this, CWIS supports importing tab-delimited resource records in a flexible format, with the first record in the imported file defining the order and meaning of fields in subsequent records. As with data entered via the Metadata Tool, CWIS can adapt to some degree to field content in imported records; for example, dates or date ranges may appear in almost any common format and will be interpreted and stored correctly for use within the portal.

The tab-delimited export format matches the data import format described earlier and should be compatible with many common applications. Each format is targeting a different audience. RSS will most commonly be used to share resource titles and information to be displayed on other Web sites, such as those implemented with uPortal (uPortal, n.d.). OAI will most commonly be used when sharing data with other groups that are working with online metadata, such as NSF’s National Science Digital Library (NSDL) (National Science Digital Library, n.d.) initiative. The tab-delimited format should be of use when collection developers want to manipulate data with other, non-Web-based applications.

**Getting Started with CWIS and SPT**

CWIS and SPT are designed to be easy to install and configure, in most cases taking less than ten minutes to get up and running when installed in the recommended environment.
Requirements

CWIS and SPT have been developed to run on a Linux-based Web server that supports PHP 4.1.0 (or later) (PHP Group, 2004) and a database server running MySQL 3.23 (or later) (MySQL, 2004). PHP must have been installed with MySQL support. If graphics manipulation is desired, PHP must include the GD library. Although Linux is the target platform, there are sites currently in operation running CWIS or SPT on Solaris and OS X. Running the software on some variants of Microsoft Windows is possible but not recommended.

As far as hardware requirements, CWIS and SPT will run on almost anything that will support PHP and MySQL. If the portal will include a large number of resources (thousands or tens of thousands), collection developers will likely want to be running CWIS on faster PC hardware because the search engine and recommender system can both be CPU-intensive.

Where to Download

CWIS and the Scout Portal Toolkit are available for download from the Internet Scout Project site on the following pages:

http://scout.wisc.edu/Projects/CWIS/download.html
http://scout.wisc.edu/Projects/SPT/download.html

Two files are available on each page: the software package itself and an installation script. The integrity of the files can be verified by checking their MD5 checksums against the values posted at the bottom of the page.

Further Customization

The dynamic interface support provided by CWIS is intended primarily to allow customization via HTML, but there are times when more extensive changes or additions are warranted. To support this, CWIS offers programming hooks for customization, where additional code may be linked in a way that will affect the operation of existing CWIS functionality. Examples of this might include additional filtering of search results or on-the-fly processing of resource metadata prior to display.

Of course, new versions of CWIS and SPT are likely to be released with additional functionality or enhanced performance. When an existing installation is upgraded to a new version, interface or programming changes are preserved wherever possible.

SPT and CWIS in Use

Because some sites are very heavily modified and others are not publicly accessible (and, of course, the software is free for download, and registration, while strongly encouraged, is not required), accurately determining the number of active SPT and CWIS installations in the field is not possible. However, the best estimates as of this writing (July 2004) put the total number of active production installations at somewhere between 45 and
60 sites and the number of SPT- or CWIS-based sites under development but not yet available to end users somewhere between 200 and 250. These numbers are expected to dramatically increase within the next year, with the increasing rate of adoption of CWIS by NSDL-related projects and the rapid growth of interest in the OAI-PMH protocol for disseminating collection metadata.

Some of the active CWIS-based sites include the Electronic Environmental Resources Library (http://www.eerl.org), a collection focused on environmental and sustainability resources for community college educators and students; the Journal of Chemical Education’s JCE-DLib repository (http://resgenchem15.chem.wisc.edu/spt/), which catalogs chemical education resources; and the Consortium for the Advancement of Undergraduate Statistical Education’s CAUSEweb project (http://www.causeweb.org/resources/). Active SPT-based sites include Duke University Libraries’ Classical Music Resources collection (http://www.lib.duke.edu/dw3/), the Tibetan & Himalayan Digital Library bibliography database (http://datastore.lib.virginia.edu/tibet/spt/), and the British Columbia History Portal (http://bchistoryportal.tca.ca/). SPT has also been used for several projects where Scout has had a more direct role, including LearningLanguages.Net (http://learninglanguages.net), a site collecting Spanish, French, and Japanese language education resources for K–12 students and teachers, and Access NSDL (http://accessnsdl.org), a portal intended to help NSDL collection builders and service projects cope with online accessibility issues. And, of course, one of the largest and most active SPT-based installations is Scout’s own Scout Archives (http://scout.wisc.edu/Archives/), which catalogs more than 17,000 online resources culled from the past ten years of Scout publications. All of these sites and more can be found on Scout’s SPT/CWIS site list (http://scout.wisc.edu/Projects/SPTCWISSites/), which is periodically updated to list new public installation of the software.

Acknowledgements

With funding from the Andrew W. Mellon Foundation and the National Science Foundation, respectively, the Scout Portal Toolkit and CWIS were developed by Edward Almasy, Barry Wiegan, Andy Yaco-Mink, Rachael Bowser, and David Sleasman. CWIS and SPT are open-source software, licensed under the GNU General Public License (http://www.gnu.org/licenses/gpl.txt) and available at no charge.

References


Gateway Standardization: A Quality Assurance Framework for Metadata

BRIAN KELLY, AMANDA CLOSIER, AND DEBRA HIOM

ABSTRACT
As digital library services develop from project demonstrators to mature, mission-critical services, it becomes necessary to develop and implement systematic procedures that will ensure the quality of the content, the functionality of the service, accessibility to a wide range of users and devices, and interoperability with other services. This article describes a quality assurance methodology that has been developed to support digital library programs in the United Kingdom higher and further education sectors. The article describes the approaches taken by the SOSIG subject gateway service in developing and maintaining a national service that is dependent on quality metadata. The article then outlines a quality assurance framework, which has been developed to support the Joint Information Systems Committee’s (JISC) digital library programs in the UK and its application to metadata. The article concludes by describing a self-assessment toolkit that can be used by service providers to ensure that they have addressed the key areas.

The Web has now established its importance for providing access to scholarly resources in teaching and research. As digital library services develop from project demonstrators to mature, mission-critical services, it becomes necessary to develop and implement systematic procedures that will ensure the quality of the content, the functionality of the service, accessibility to a wide range of users and devices, and interoperability with other services. In the UK we have been working toward this end by developing a “quality
assurance (QA) methodology” to support digital library programs in the UK higher and further education sectors. This article describes the approaches taken toward developing and maintaining a national service that is dependent on quality metadata. The self-assessment toolkit we have developed can be adopted by subject gateway service providers to ensure that they have addressed the important issues facing digital library services—standardization and quality control.

BACKGROUND
In the UK the Joint Information Systems Committee (JISC), which funds a range of networked services for the higher and further education communities, has played a key role in the development of digital library services. The JISC established the eLib program (eLib, 2001) in the mid-1990s, providing an opportunity for experimentation in multiple areas, including a strand for the establishment of pilot subject gateways. Following the success of the eLib program and the recognition of the Web as the key delivery platform for scholarly resources, the JISC subsequently established a strategy for accessing these resources seamlessly. Initially known as the DNER (Distributed National Electronic Resource) but later renamed the JISC Information Environment (IE), the implementation of this strategy is based on a number of JISC programs that fund the development of a wide range of projects. These projects will, together with related JISC service developments, help to provide the IE’s content and technical infrastructure.

An example of one of JISC’s national services is the Resource Discovery Network (RDN), which provides access to scholarly resources in various subjects. The RDN is an ambitious subject gateway system made up of eight area (or hub) subject gateways. These services (as indicated below) are hosted at particular universities throughout the UK and draw upon the expertise of over seventy educational and research organizations, including the Natural History Museum and the British Library. A summary of the RDN hubs is given in Table 1.

The RDN is now recognized as one of the Web’s most reputable scholarly resources, with clear missions and interfaces set in place. Although, inevitably, there will continue to be a need for experimentation as new formats and protocols are developed and different types of services are evaluated, there is now a need to ensure that project deliverables can be deployed into a service environment with ease. In other words, once a gateway is built, it is necessary to establish systematic maintenance procedures, as well as continue to add resources to it. What follows is a description of a quality assurance (QA) framework for maximizing digital library services.
Quality Assurance Framework

The SOSIG case study (explained in detail below) outlines a practical approach for ensuring the quality of the service’s metadata and hence maintaining the quality of the service. With the success of SOSIG’s quality assurance procedures, it became clear that these methods could be implemented on a wider scale—to other JISC-funded services. In this section we describe how the JISC has funded the development of a quality assurance methodology for its digital library programs and how this methodology can be applied to the creation and management of metadata.

QA Focus

In 2001 the JISC issued a call for a “Digitisation and QA Focus” service (Joint Information Systems Committee, 2001). The call recognized that “Past digitisation programmes tended to operate in an environment where technologies were relatively immature and unstable, therefore suggesting a research-orientated approach to the management of digitisation activity.” Following a successful bid the project (which was renamed “QA Focus”) was provided initially by UKOLN (a national center of expertise in digital information management based at the University of Bath) in conjunction with the Institute of Learning and Research Technology (ILRT) based at the University of Bristol (ILRT is the host organization for the SOSIG service.) UKOLN and ILRT are located close to each other and have been involved in a number of joint activities, including the EU-funded DESIRE project (DESIRE, 2000a). One deliverable from the DESIRE work was an Information Officer’s Handbook (DESIRE, 2000b), which describes best practices to support libraries and other organizations interested in setting up large-scale information gateways on the Internet. This handbook, which was jointly authored by staff at ILRT and UKOLN (and others), helped develop both

<table>
<thead>
<tr>
<th>Service</th>
<th>Area</th>
<th>Host</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALTIS</td>
<td>Hospitality, Leisure, Sport, and Tourism</td>
<td>University of Birmingham</td>
</tr>
<tr>
<td>Artifact</td>
<td>Arts and Creative Industries</td>
<td>Manchester Metropolitan University</td>
</tr>
<tr>
<td>BIOME</td>
<td>Health and Life Sciences</td>
<td>University of Nottingham</td>
</tr>
<tr>
<td>EEVL</td>
<td>Engineering, Mathematics, and Computing</td>
<td>Heriot Watt University, Edinburgh</td>
</tr>
<tr>
<td>GSource</td>
<td>Geography and Environment</td>
<td>Consortium of Academic Libraries in Manchester</td>
</tr>
<tr>
<td>HUMBUL</td>
<td>Humanities</td>
<td>University of Oxford</td>
</tr>
<tr>
<td>PSIgate</td>
<td>Physical Sciences</td>
<td>Consortium of Academic Libraries in Manchester</td>
</tr>
<tr>
<td>SOSIG</td>
<td>Social Sciences, Business, and Law</td>
<td>University of Bristol</td>
</tr>
</tbody>
</table>

Table 1. Examples of Resource Discovery Network Hubs
organizations’ expertise and knowledge of quality assurance processes for metadata and, indirectly, led to the work described in this article.

Today, QA Focus is a joint venture between UKOLN and the AHDS (the Arts and Humanities Data Service, based at King’s College, London). The change in partnership (which followed ILRT’s decision to refocus on their core activities after the first year of the QA Focus project) has strengthened the QA Focus team due to AHDS’s additional service responsibilities and experience in a wider range of digitization activities.

The role of QA Focus is to help ensure that project deliverables are interoperable and widely accessible. The remit of the work covers the areas of standards, digitization, the Web, metadata, software, and service deployment. QA Focus seeks to ensure that projects deploy appropriate open standards and best practices in these areas. The approach taken has been published elsewhere (Kelly, Guy, & James, 2003) and is summarized below.

**Initial Groundwork** Focus group meetings were arranged in the first year, providing an opportunity for QA Focus to inform projects of the service and to gain feedback on work areas that needed to be addressed. The meetings raised the following issues:

- a lack of awareness of recommended open standards in some cases
- difficulties in implementing standards in some cases due to lack of expertise, immaturity of the standards, or poor support for the standards
- concerns over changes in standards during the projects’ lifetime

Although it was pleasing to hear that many projects were committed in principle to the JISC’s open standards philosophy, it was also clear that implementing open standards would not be easy: projects faced other pressures such as lack of technical expertise, short time scales, investment in existing tools and products, and use of third-party applications and data that sometimes hindered deployment of open standards.

Another activity carried out in the first year was a series of benchmarking surveys of the Web sites provided by the JISC 5/99 projects. The surveys made use of a variety of automated tools, which analyzed the compliance with HTML and CSS standards for the projects’ home pages and other features, such as the number of broken links, use of embedded metadata, etc. Although such automated surveys have their limitations (automated accessibility tools need to be supported by manual tests in order to ensure pages are accessible, for example) the surveys were valuable in providing an understanding of common problems and in helping to identify and prioritize areas in which advice was needed.

**Briefing Paper** The findings of the focus groups and the surveys helped us prioritize the areas in which advice was needed. Since QA Focus was not funded to provide direct support to projects, our advice came in the form of short, focused briefing papers. Currently over seventy briefing papers...
have been produced, covering the areas of standards, digitization, the Web, metadata, software, and service deployment.

Advice on Testing Tools There is a clear need for tools to check that resources comply with standards and best practices, including tools such as HTML and CSS validators and link checkers. Although Web developers should be familiar with such tools, our experiences have revealed a number of factors that may result in misleading results:

- Definition of Links: Some links checkers will only check conventional hyperlinks and embedded images. However, links can also be provided using the <LINK> tag for links to external resources such as JavaScript files, CSS files, and metadata resources.
- HTTP Headers: Testing tools should take appropriate actions based on HTTP headers received. Some testing tools report on the output of an HTTP header rather than reporting on the header received.
- Misconfigured Servers: Servers, caches, firewalls, etc. can sometimes be misconfigured, giving misleading findings.
- Personalized Pages: There is an increasing need to be able to test personalized pages. The personalization may be due to a number of factors, including user preferences, browser type and environment, regional factors, etc.

Online Toolkit In order to help embed quality assurance procedures, we have developed an online toolkit that provides a simple checklist. The toolkit helps to focus the developer’s mind on key issues and provides advice on the main areas to be addressed. Online toolkits are available in several areas including standards selection, mothballing Web sites, and metadata.

Selection of Standards Although digital library services seek to make use of open standards, there can be dangers in making use of immature standards or not having the resources and expertise needed for the successful implementation of certain standards. We have published a methodology on the selection of standards (Kelly, Dunning, Guy, & Phipps, 2003).

QA Focus Methodology The key deliverable of the QA Focus project has been the development of a lightweight quality assurance methodology. The QA methodology has been informed by the ISO 9000 standard for quality management (International Organization for Standardization, 2004). The methodology requires projects to provide documented policies on their technical infrastructure and systematic procedures for ensuring they comply with their policies.

Case Studies In order to support the sharing of experiences across the JISC digital library community, QA Focus has also commissioned case studies that provide an opportunity for projects to share their approaches to technical developments. The SOSIG case study illustrates a typical example.
SOSIG Case Study

The development of a QA framework had its roots at the Social Science Information Gateway (SOSIG). SOSIG (Social Science Information Gateway, 2004) is a well-established Internet resource discovery service for the social sciences, business, and law. SOSIG is based at the Institute for Learning and Research Technology (ILRT), University of Bristol. It is funded by the UK’s Economic and Social Research Council (ESRC) and JISC and is part of the RDN. The SOSIG Web site is illustrated in Figure 1.

SOSIG began life as a pilot project in 1994 and is now considered by many as a pioneer amongst Internet subject gateways in the UK and worldwide. The core of the service, the Internet Catalogue, currently holds over 27,000 structured metadata records (across 17 top-level subject headings and over 1,000 subsections) describing Internet resources relevant to social science learning, teaching, and research. Since its inception, members of the SOSIG team have consistently worked on and developed tools, meth-
ods, and procedures that support the creation and ongoing maintenance of quality-controlled information gateways. One of the co-authors (Hiom et al., 2003) has been working for the SOSIG service since its launch and has worked closely in the development of the quality assurance procedures for the service. Though always an important consideration, the need for quality assurance procedures has emerged as a real issue with the increasing size and scope of the SOSIG service. To this end, SOSIG now has an established and comprehensive set of procedures that range from the selection of resources to the systematic weeding of the collection. The following case study documents the QA procedures at SOGIG that underpin the creation of high-quality metadata records. These procedures involve subject specialists who are carefully trained according to clear policies, guidelines, and criteria, as well as various automatic checking measures to further standardize the process.

Use of Subject Specialists The records are created and maintained by a geographically dispersed team of over forty subject specialists (known as Section Editors) who select and catalogue Internet resources within a particular subject area. SOSIG relies on a solid set of quality assurance methods that aim to ensure consistency and accuracy amongst the team of specialists. The Section Editors are responsible for seeking out, evaluating, and describing social science Internet resources within their specialized subject area. In addition, the service interoperates with other subject gateway services and therefore aims to ensure that all of its catalogue records are compatible with the wider Resource Discovery Network.

At SOSIG a great deal of time and effort has gone into developing procedures to ensure a consistent approach to the cataloguing process. A thorough training program is backed up with detailed and comprehensive printed and online reference material available to all Section Editors.

Training Each Section Editor receives training on all aspects of working with SOSIG. This begins with an overview of the service from the end-user perspective. An explanation of how the service is used by real people helps to set in context some of the editing procedures—identifying relevant keywords, for example. An end-user perspective is followed by sessions on best practice in locating and evaluating resources and practical training on the online cataloguing center. The workshop is supplemented by documentation in the form of a workbook, as well as a step-by-step guide to cataloguing that includes the following:

1. The SOSIG Scope Policy, which outlines the type of resources the Internet Resource Catalogue covers in terms of subject matter, geographical coverage, language, etc.
2. The Collection Management Policy, which offers a guide to the selection and deselection criteria for the collection.
3. The Evaluation Criteria, which explains how potential resources are evalu-
ated in terms of content, presentation, and any quality assurance procedures that may be in place.

4. *The Cataloguing Rules*, which aim to help SOSIG editors use standard practices when adding records to SOSIG to ensure that records within the database are consistent and of a high quality. The rules include an explanation of each of the metadata fields and how they should be entered (that is, particular formats for dates, names, etc.). The rules also include links to further information such as classification schemes, country codes, etc.

The cataloguing rules document is the most important one in terms of ensuring consistency across the service. With published and publicly available documentation on all areas of the selection and evaluation procedure, Section Editors have a constant resource to turn to while working on the catalogue, while end users can gain a better understanding of what to expect of the service.

**Online Tools and Checks** SOSIG has integrated a range of online tools and automatic checks (many at the request of Section Editors) into the cataloguing process in an attempt to eliminate errors and inconsistencies prior to the records being added to the catalogue. Controlled vocabularies or thesauri are used for assigning keywords to the records to help in the standardization of spellings, but more importantly, to help users find other related terms and records linked to their topic of interest. Wherever possible, SOSIG uses preformatted authority files to minimize the risk of typing errors creeping into records. Editors are also encouraged to cut and paste URLs into records to avoid errors. Conversely, Editors are encouraged to create freehand, textual descriptions for records. These are seen as an important and value-added aspect of SOSIG. To counter error the system operates a spell-check facility that checks the record as it is submitted to the database and highlights any words it does not recognize. Occasionally this can prove problematic, especially with proper names and technical terms, but SOSIG has included an override function as well as the ability to add particular words to the spell-check dictionary. Online help and access to the cataloguing rules are also provided for Editors through the cataloguing form.

**Post-Cataloguing Methods** The ideal situation for SOSIG and other digital libraries is to ensure that procedures for quality assurance are robust enough to minimize any editing work after the creation of the catalogue record. Given the volatile nature of information on the Internet, however, it is necessary to implement a number of quality checks on the existing metadata records.

**Automatic Confirmation of Record Creation** As metadata records are created, an email message is sent to the administrator of the catalogued resource or site to inform them that they have been added to SOSIG and
to give them the opportunity to read the description. Suggestions and amendments can be sent directly to the central administration team for approval. Email content, we have found, is essential for the maintenance of the SOSIG database. We conducted a major one-off “clean-up” exercise in 2003, contacting all administrators of sites that had been catalogued by SOSIG and requesting that they check their record on the SOSIG database for accuracy. This process provided multiple benefits: it not only allowed us to check the accuracy of the records, but it also served as a promotional tool for the service and often resulted in reciprocal linking, suggestions for additional useful material to add to the gateway, and a communication channel for administrators to notify us about major overhauls of their own sites.

Link Checking and Reviewing Given the dynamic nature of the Internet, and the Web in particular, collection development is a major task. Collection management (that is, removing broken links, checking and updating records) at this scale can also be something of a challenge. Many sites often change constantly or even disappear, only to reappear under a new guise. To counter this, an automatic link checker is run over the entire database of URLs on a weekly basis and errors are noted in a report that is made available to Section Editors.

Of course it is not only link errors that need to be considered. Records should also be reviewed on an ongoing basis to ensure that they are still accurate and suitable for inclusion within the catalogue. The Collection Management Policy outlines the principles and process for editing and deleting records. For example, if the information content of the resource has changed so that the resource description and keywords need to be updated, or if the currency or reliability of the resource has lessened over time, the policy has clear directives on how to handle such cases.

Section Editor Workshops Because Section Editors at SOSIG work as a geographically dispersed team, we feel it is important that they are able to get together on a regular basis in order to meet each other and exchange experiences. Consequently, the whole team meets annually to discuss the development of the overall service, to plan changes to their individual sections, or just to brush up on skills generally. Feedback from the workshops suggests that Editors find these events invaluable in that they help to reduce feelings of isolation that can so easily develop within virtual teams.

Summary SOSIG has grown into a large and significant resource. The size of the catalogue raises considerable issues in terms of collection maintenance and the management of a distributed team from many disciplines. The QA tools and procedures described above have developed over a considerable period of time. They are now considered a vital element of the service in that they support the needs of both the central administration of the gateway, the team of distributed Section Editors, and, most importantly, the needs and expectations of the end users.
Quality Assurance for Metadata

We have given a broad outline of the QA Focus work. We will now focus on the application of this work to the area of metadata.

Purpose of the Metadata

Decisions on the use of metadata in any digital library project should be based on the functionality to be provided by the metadata. The functionality required will influence the metadata standards to be used as well as the architecture for managing and deploying the metadata. Implementing appropriate quality assurance procedures into a project’s planning activities and workflow practices will help to ensure that the metadata is and remains fit for its purpose.

Cataloguing Rules

There are a number of problems that can arise for any project using metadata. Probably the most important is the issue of consistency. Ensuring that metadata consistency is maintained is important if interoperability is to be achieved. Where resources are catalogued by more than one person (or indeed organization), the potential for errors in the metadata multiplies. Thus it is vital to ensure that cataloguing rules and a consistent approach are implemented across the board. Services such as SOSIG have adopted a systematic approach to minimize the problems that a geographically distributed service faces when creating metadata. A well-defined interface for inputting metadata, which restricts variation as much as possible, can help this process. Selection lists populated from a controlled vocabulary or fields that only accept data in a particular format are useful ways of restricting variation in metadata creation. The use of authority files will help ensure that naming conventions are followed systematically.

Maintenance

In addition to ensuring that any metadata produced is consistent one must ensure its currency. The evolution of electronic resources is an almost constant activity, and it is important to update the resource’s associated metadata alongside the resource itself. Not only will project staff find inconsistencies unhelpful, but machine interfaces will not be able to spot out-of-date information in the way that humans can. The popularity of a resource may fall if users believe it to be out-of-date, even when in reality the resource has been revised recently.

Interoperability

It is important not to be too restrictive when thinking about creating metadata for a project’s resources. For metadata to be widely used it must be interoperable. While records may start out only being used in house, ensuring that the project’s metadata conforms to standards and maps easily to other metadata schemas will allow the metadata to be used more widely. SOSIG is an excellent example of a pilot project that has evolved
into a service and is now a major Internet gateway used on a national and international scale. Project staff will need to be aware that different cataloguing rules may be used in other environments. As an example, date formats often throw up inconsistencies between the United Kingdom and the United States.

Validation

It is important to ensure that any metadata a project or service produces is validated. If metadata is encoded in XML, it must be validated against a DTD or schema. Metadata creation and management tools should be configured to validate newly created metadata and output it in a controlled format.

Errors may occur in the workflow process: a Microsoft Windows character such as the © symbol could be entered into a database and then embedded in a metadata record in XML format. However, this character is an invalid character in an XML format. The impact of such errors in the record can be considerable: a record that is not spell checked or presented consistently will reduce the impact of your metadata, the service it provides, and its interoperability.

We present a fictitious scenario below in which some of the common problems that can arise when producing metadata have been drawn together.

A multimedia e-journal project is set up. The Dublin Core metadata element set is used to describe published articles. There are documented cataloguing rules in place but, unfortunately, due to a high staff turnover (many staff are on short-term contracts), there are many inconsistencies in the metadata (John Smith & Smith, J.; University of Bath and Bath University; etc.).

The metadata is managed by a home-grown tool. Unfortunately, the author details are output in HTML as DC.Author rather than DC.Creator. In addition, the tool outputs the metadata in XHTML 1.0 format, which is embedded in HTML 4.0 documents.

The metadata is created by hand (with no interface to simplify and control the process) and is not checked. This results in a large number of errors and use of invalid characters (for example, £,—, and &). Consequently, the quality of the records is low.

The metadata describing copyright and access information for the images associated with the articles becomes separated from the images during the workflow process. Since some resources can be freely used by all but others are restricted (used only by the host institution), the separation of the rights metadata from the resources means that the project deliverables cannot be used by third parties.
QA For Metadata Toolkit

We have described a number of areas in which there is a need to address metadata quality when supporting resource discovery. However, metadata can be used to support a wide range of areas, such as maintenance of Web sites, access to e-learning resources, or accessibility. Rather than providing detailed advice for every area in which metadata can be used, we have sought to develop a simple model that can be applied in many areas. Our online toolkit for QA for metadata seeks to ensure that projects have given due consideration to key areas. The QA for metadata toolkit is illustrated in Figure 2.

It should be noted that the toolkit is intended for self-assessment purposes only. A record of the responses is not kept.

The issues addressed in the toolkit are:

• clarification of the purpose for which metadata is being used
• use of an appropriate metadata schema and appropriate cataloguing rules
• appropriate technical architecture for creating and managing the metadata
• procedures for checking the metadata content and syntax
• appropriate training and staff development policies
• liaison mechanisms with potential remote users of the metadata

We have recommended to the JISC that those JISC-funded projects making significant use of metadata should address these issues as part of the project’s reporting procedures. We feel that this lightweight but important approach to the quality assurance of metadata can help minimize interoperability problems and can also be of benefit if a service is to be deployed in a service environment.

Conclusion

In this article we have described the approaches taken by mature subject gateway services such as SOSIG to ensure that they deliver and continue to provide the quality metadata that is essential for an effective subject gateway service. We have sought to generalize this work in the form of a quality assurance framework, which can be deployed by projects and services that wish to make use of metadata. Finally we have described how this quality assurance framework has been extended to support the broad interoperability of JISC’s digital library programs.

Metadata is critical to the effective deployment of many digital library environments such as open archives, e-learning environments, and semantic Web applications. Quality assurance procedures will be critical to the effective deployment and interoperability of such services. The authors hope that this article has outlined a quality assurance framework that can be of use to those involved in development work in this area.
Figure 2.

Self Assessment

Metadata

1. Have you a clear idea of the intended purpose of your metadata?  Yes ☑ No ☐
2. Have you identified the functionality to be provided by the metadata?  Yes ☑ No ☐
3. Have you chosen an appropriate schema for your metadata?  Yes ☑ No ☐
4. Do you have a mechanism for managing your metadata?  Yes ☑ No ☐
5. Have you identified potential problems areas for your metadata?  Yes ☑ No ☐
6. Have you made plans for ensuring the correctness of your metadata?  Yes ☑ No ☐

Your scores are given below.

Metadata

1. Have you a clear idea of the intended purpose of your metadata?  Points -10
   NO.
   You should have a clear idea of the intended purpose of your metadata before allocating resources on the creation and maintenance of your metadata. For further information see Metadata Deployment Briefing paper.

2. Have you identified the functionality to be provided by the metadata?  Points +10
   YES.
   Well done.

3. Have you chosen an appropriate schema for your metadata?  Points -10
   NO.
   Ideally you will choose a well-
REFERENCES


Strategies and Technologies of Sharing in Contributor-Run Archives

PAUL JONES

ABSTRACT
While we argue about and discuss the plusses and minuses of contributor-run archives, groups formed by people of shared interests and of varied technical competencies have been creating, maintaining, sustaining, and growing their archives for over a decade in several cases. These contributor-run archives make use of powerful open technologies to facilitate their projects. In this article I will focus on three different volunteer-run projects that involve worldwide cooperation using advanced technologies to further their ends. The Linux Documentation Project, the Degree Confluence Project, and Etree.org are all large projects that involve many contributors with technical teams of various sizes using a variety of technologies. Each project will be described in terms of its aims; its history; its rules, or lack thereof, for contribution; its technologies; and its current state of practice. From these examples we can draw some lessons as well as some enhanced awareness of technologies of cooperation. Among the technologies used by the projects are wiki, mailman, Shorten (SHN), FLAC, PHP, mySQL, PHPbb, Postnuke, BitTorrent, rsync, XML, and CVS. All of these technologies are “open” and available for installation, customization, and further sharing of their code.

Over my dozen years as director of ibiblio.org and its predecessors, sunsite.unc.edu and metalab.unc.edu, I have seen many projects flourish and many projects stagnate and more than a few projects die completely. At the time of this writing, ibiblio.org hosts and facilitates over 1,500 projects in addition to our extensive software collections. In May of 2001 I published

Paul Jones, Director of ibiblio.org, Clinical Associate Professor, University of North Carolina, Campus Box 3456, Chapel Hill, NC 27599–3456
© 2005 The Board of Trustees, University of Illinois
a brief article describing how open-source tools might be used in contributor-run libraries (Jones, 2001, pp. 45–46). For this article I aim to describe some successful examples of open-source contributor-run collections. I have selected three projects with worldwide contributor bases, innovative technology use, open management, and volunteer staff for consideration. All three projects solicit participation from their users, and amazingly they have consistently received reliable and enthusiastic contributions. As a result, each is—within its particular area—a must-visit resource.

Briefly, the Linux Documentation Project aims to provide reliable, accurate, and helpful documentation to Linux users from beginners to advanced systems administrators in every language in the world. The Degree Confluence Project aims to document the world by visiting every degree confluence on the earth. A degree confluence is defined as “the exact spot where an integer degree of latitude and an integer degree of longitude meet.” Confluence.org volunteers participate in creating a database of photographs and narrative descriptions of their visits to each degree confluence on, or near, dry land on the entire earth. Etree.org aims to provide a forum for the exchange of very high-quality concert recordings of “tape friendly” bands (Etree.org, 2004e).

Each of these projects has an education component as part of its mission—that is to say, guidelines and FAQs for new contributors and new users. These educational components also serve to advance the ideology of sharing the information, skills, and experience that is a part of each project.

**The Linux Documentation Project**

The Linux Documentation Project (2004), begun by Matt Welsh in 1992 not long after the first wide release of Linux itself, predates the World Wide Web (Garrels, 2004; M. Garrels, personal communication, May 14, 2004). The goal of the project, as described by volunteer David Lawyer in the Linux Documentation Manifesto, is “to create the canonical set of free Linux documentation. While online (and downloadable) documentation can be frequently updated in order to stay on top of the many changes in the Linux world, we also like to see the same docs included on CDs and printed in books” (Lawyer, 2000). Thus, while the Linux Documentation Project can be seen as a long-lived online community project, its goals are not limited to cyberspace; the project aims for world conquest—or at least to conquer the world of Linux Documentation. To a large extent, TLDP—as it is now known—has succeeded. Andy Oram of O’Reilly and Associates, a leading technology publishing company that might be considered to be the competitor of TLDP, has written that TLDP “is an impressive organization that has editors, guidelines for reviewers, procedures for updating documents, translators—in short, it’s an organization that has tried to reproduce everything about conventional publishers, but in an open and volunteer
manner” (Oram, 2004). Oram also praises TLDP as “a phenomenon we should all be following as a model for documentation in an open source community” (Oram, 2004). I would note that an organization of a dozen years is no longer a phenomenon but is, in the world of cyberspace, an institution.

Those wishing to contribute documents to TLDP are pointed to a detailed yet straightforward Author’s Guide that describes what and how to participate. The process goes as one might expect from any publishing company:

1. Become familiar with the Linux Documentation Project’s other works by looking over the site and joining the Discuss mailing list.
2. After having identified a gap in the documents or that a new document is needed, propose your document to the Discuss mailing list, including if possible an outline and description of the document.
3. Write your document.
4. Mark up your document or seek help in doing mark up. All documents published by TLDP are in SGML, Docbook XML, or LinuxDoc formats to allow for flexible republication. Obviously, this might constitute a high barrier of entry for contributing writers, but TLDP volunteers have agreed to work with new contributors by instructing or even providing proper markup for submitted documents (Sundaram, 2003).
5. Submit your document for review by sending a copy or a link to a copy to the Submit mailing list. A language editor, a technical editor, and a metadata editor review all documents. It is not unusual for all three editors to actually be the same person. This process could take up to two weeks. Of particular note to readers of *Library Trends* is the requirement that eleven metadata fields be complete and accurate before a document is accepted. Metadata editor Emma Jane Hogbin writes that the goal is for TLDP documents to be Open Archive Initiative–compliant within the year (E. J. Hogbin, personal communication, May 12, 2004).
6. After, or even during, the review process, the document is added to the Concurrent Version System (CVS) for TLDP. While the use of CVS is optional, it is a great innovation. CVS allows an author to keep an offsite copy of the document as well as allowing other authors or editors to make traceable changes in the document. Additionally, the change log may be included in the document automatically as an aid to readers. Ideally in the future, the change log will also interact with the appropriate metadata elements and will be used to announce the new or newly revised document to the TLDP Web site and appropriate lists (E. J. Hogbin, personal communication, May 12, 2004).

Other than mailing lists and CVS, TLDP’s use of technologies of cooperation is minimal but highly effective. Requiring metadata makes documents easier to find and to use in a trusted manner. Choosing an open markup
language, XML, and an open tag set, Docbook, as opposed to Word or even PDF, gives the documents flexibility and longevity. That flexibility includes the ability to use open-source publishing tools such as openjade, the dsssl stylesheets, libxml2, xsltproc, XSL, etc. (Hogbin, Komarinski, Godoy, & Merrill, 2004).

Because of the strong and mostly friendly involvement of the editors, there are few documents rejected. In fact, most rejections occur at the proposal stage and usually because of content overlap with an existing document. Even then the proposing author is encouraged to contribute to the overlapping document and so to be a part of the continuing process (G. Ferguson, personal communication, May 12, 2004).

The current TLDP archives hold works from over 500 different authors, although not all of the authors are active at one time. There are 345 subscribers to the discuss@en.tldp.org list, 52 active editors, and a core team of 19, as well as translation coordinators working in languages from Albanian to Walloon. The impact of TLDP cannot be overestimated. Not only are TLDP documents distributed with major Linux software distributions, but also the entire site is mirrored or copied completely on over 300 official sites around the world (see http://tille.soti.org/images/tldp-world.jpg).

The Degree Confluence Project

The goal of the Degree Confluence Project (DCP) is to document visits to “each of the latitude and longitude integer degree intersections in the world” by narrative descriptions and photographs. The photographic narrative usually includes at least one photograph of a GPS (Global Positional Sensor) device showing evidence of having been at the confluence. The first project visit documented a trip by project founder Alex Jarrett in February 1996 when he and friend Peter Cline visited 43 degrees North 72 degrees West near Hancock, New Hampshire (Jarrett, n. d.). Jarrett posted information about his confluence visit to a personal Web page and invited others to send him information about their own visits to confluences. Before long, new people and new technologies were put in place to support a grander plan: to document visits to all confluences on or within sight of land. As of this writing, 3,137 confluences in 148 countries have been visited and successfully documented (Degree Confluence Project, 2004a).

Degree Confluence is a masterpiece of PHP and mySQL coding that allows for interactions with other sites including Mapquest (for street maps) and Terraserver (for aerial maps of the United States) as well as links to the antipode, or exact opposite side of the globe, for each confluence and navigation to adjacent confluences. Additional customized mapping allows for a clickable view of the world composed of images taken by confluence visitors.

While the Degree Confluence Project might on one hand be thought of as an eccentric excuse for using techie toys (GPS, map-coding projects,
digital cameras, etc.), it clearly has other components, including social and educational. My own son’s fifth grade class has used the Degree Confluence Project in learning about geography, mapping, and global environments. The project is not without the occasional dramatic moments as volunteers trek to exotic confluences. Most recently in the search for 19 degrees South 49 degrees East, one half of a honeymooning confluence-seeking pair vanished for 23 hours—arrested as it turned out (Christen & Christen, 2004).

Besides being about technology, mapping, global awareness, and drama, the Degree Confluence Project is about volunteerism, sharing, and trust. As seen in the Linux Documentation Project, there is some gate keeping done by volunteers, but this gate keeping is minimal, usually in the form of rejecting incomplete submissions, attempting to clear up confusions about the proper geographic datum to use, or confirming the photographic guidelines (D. Patton, personal communication, May 12, 2004). In the FAQs for the DCP, the theme of trust is reiterated: “Basically, we trust people, unless we can show that it’s not at the right spot (usually by comparing with a map) or the narrative clashes with the pictures, then it will be more thoroughly verified” (Degree Confluence Project, 2004c).

The Degree Confluence Project handled an average of 120 submissions a month in the past year, June 2003 through May 2004 (Degree Confluence Project, 2004d) with a volunteer staff of 3 administrative coordinators, 2 technical coordinators, and 8 regional coordinators (Degree Confluence Project, 2004b).

The PHP/mySQL-based Web site and associated database, along with custom mapping software, support most of the contributor interactions for this project. The Degree Confluence Project also uses email and mailing lists to coordinate activities but, unlike TLDP, Degree Confluence exercises little in the way of version control or other oversight. Still, the fast-paced growth and internationalism of the site show that its contributors have earned the trust of the administrators and visitors to the Confluence Project, making it a valuable way to learn about the world.

**Etree.org**

Moving from software documentation and global adventure to trading music that is recorded live is not such a far leap. In many ways the culture of sharing concert tapes has been a model for ad hoc communities of interest organizing themselves in a variety of communications and exchange media going back to using the postal service or even trading tapes hand-to-hand at concerts.

As bands with fanatical followers go, few in history can compare in intensity and in longevity with the Grateful Dead. Phish, Dave Matthews, and Dead spin-off bands do not pick up the identical memberships but they do pick up nearly identical enthusiasms. In the online world—as well
as in “meatspace,” as Grateful Dead songwriter and net-spokesman John Perry Barlow has called the non-cyberworld—Deadheads are a community by any definition.

Since the 1960s, fans of these bands have been recording concerts and trading tapes. After a while and after many poor-quality tapes were circulated, the Grateful Dead created a policy for tapers that included a special tapers’ section at every concert and some simple ground rules for noncommercial use of the tapes. This practice became more institutionalized and is a part of every Grateful Dead concert to date (The Dead, 2004). Along the way, tapers themselves sought out standards for describing the recording steps, including concert conditions, microphone specifications, audio transformation, and varieties of ways in which digital audio could be distributed in near loseless formats. Most recently they have settled on SHN (or shorten), but they have an eye open for emerging formats such as FLAC (Etree.org, 2004c, 2004d).

It is worth noting that fan taping has not stopped the Grateful Dead from creating and selling their own CDs of their live shows. In fact, the Grateful Dead has developed not only one of the largest archives of live shows, but also—possibly in response to the tapers’ insistence on high quality and open description of conditions, etc—the band’s organization publishes, in detail, the steps with which their commercial live concert CDs are made (The Dead Summer Getaway, 2003). Grateful Dead fans also shared information online. From the earliest days of virtual communities, Deadheads filled the most active discussion groups on the WELL (Rheingold, 2000). Fans of later bands such as Phish took a page from the Dead’s book and built communities of their own both in concerts and online, including Phish.net.

Over the years, a number of Web sites have been developed to assist in the trading of high-quality concert recordings. The direct predecessors of Etree.org are PCP, People for a Clearer Phish, and Sugarmegs (a name in which the Grateful Dead song, Sugar Magnolia, meets megabytes). PCP is a CD trading tree organization that uses CDRs and postal mail to spread live recordings of Phish concerts from limbs to leaves (People for a Clearer Phish, 1999). Sugarmegs was originally dedicated to sharing recordings of Grateful Dead concerts in various formats by downloading and by streaming—the site was even hosted by Microsoft at one point in its life (Black, 2000). Etree.org prides itself on offering the highest-quality recordings with the least compression. As the site’s home page states,

etree.org is the award-winning leader in lossless digital audio distribution on the Internet! We are a community committed to providing the highest quality live concert recordings in a losslessly-compressed, downloadable format. All of the music on etree.org is free, and 100% legal to download, trade, and burn. We also assist new traders in learning to trade online through our extensive guides. (Etree.org, 2004b)
Of the three projects considered, Etree.org makes the greatest use of the largest variety of technologies. The Etree site is really several sites with several different, but interrelated, functions. Etree.org, the Web site, has been supplanted by wiki.etree.org (more below). Forums.etree.org offers threaded discussion groups to members via PHPbb. News.etree.org offers a syndicatable blog and announcement area via Postnuke. Etree.org/irc gives access to real-time chat among etree-ers via Internet relay chat. Db.etree.org is the gateway to the real business of etree; it is the database that describes in extreme detail the music that etree members have to share. And Bt.etree.org gives access to the actual music via the peer-to-peer BitTorrent protocol.

Etree also has the most individuals directly involved in managing the project’s resources. Volunteer Tom Anderson reports that there are over 280 people who have some kind of administrator privilege for db.etree.org. However, only three of those are involved with technical administration of the site and related database. The others are volunteer content administrators who are assigned to specific artists or who help out with artists who do not yet have an assigned administrator (T. Anderson, personal communication, May 14, 2004). Over 350 artists have had shows recorded and placed in the database.

Etree sites other than the database site require only a few administrators. The etree site with the next largest number of people involved is forums.etree.org, where up to eleven volunteers serve as moderators. According to volunteer Caleb Epstein, www.etree.org has often languished, been out of date, and was even occasionally inaccurate in large part because the administrators rarely found time to work on it. To alleviate this problem, etree.org chose a more open solution, a wiki, which has been wildly successful and popular (C. Epstein, personal communication, May 12, 2004).

A wiki is software that allows simple writing, linking, publishing, and editing in a collaborative and collective fashion. Thus, when readers find a mistake or have new news or additional information, they can quickly add their knowledge to the pages. While many fear that pages as open as wiki pages might be defaced or become full of misinformation, experience shows that this is not the case. Wiki pages do show a history of changes and do allow changes to be rolled back out if needed (WikkiTikkiTavi, 2004).

The wiki philosophy was already in place at etree.org in their use of the trading database at db.etree.org. Of the over 104,000 members who have created sign-ins that allow modifying or adding to the information in the database, Anderson reports that he has banned only 10 members since 1999 (T. Anderson, personal communication, May 14, 2004). Many of the static pages within db.etree.org are also wiki pages, which may be edited or added to by any registered member. The database itself is really a collection of metadata—rather than an audio collection—about shows performed, and hopefully recorded, by tape friendly bands. The set lists,
dates, venues, and the like are included, as well as, for cases in which there are SHN recordings, metadata describing the conditions under which the recordings were made.

An example of a successful metadata entry for a concert might look like this:

Band: Charlie Hunter Trio  
Date recorded: 05/21/2004  
Venue: Mr. Small’s—Millvale, PA  
Source: Soundfield ST-250 (stage lip, M-S) > Lunatec V3 (24/96) > M-Audio Firewire Audiophile > Sony PCG-V505AX > WaveLab 4 (24/96)  
Conversion: WaveLab 4 (dither and resample) > SoundForge 6 (M-S decoding) > CD Wave Editor > FLAC  
Recorded & Converted by Jef Fugh

This is followed by the set list and comments. The database also allows for additional comments by other volunteers and links to other recordings of the same show should they exist. Once submitted, the record must be reviewed by a volunteer before it is published (db.etree.org, 2003).

Bt.etree.org is a slightly different case from the database site in function, in size, and in the metadata review process. BitTorrent, the “bt” in the name, is a peer-to-peer file-sharing system that allows parts of large files to be downloaded from several sites at once, thus reducing the bandwidth and processing demands on any single machine. The Free Software Directory describes BitTorrent as

a tool for copying files from one machine to another. FTP [File Transfer Protocol, the most common protocol for downloading files] punishes sites for being popular. Since all uploading is done from one place, a popular site needs big iron and big bandwidth. With BitTorrent, clients automatically mirror files they download, making the publisher’s burden almost nothing. (Casey, 2001)

The person seeking to download a file uses a BitTorrent client. The client connects to a BitTorrent tracker, like bt.etree.org, to initiate a download.

The tracker’s job is to keep tabs on computers that have successfully downloaded files. The tracker does not have the music file itself but instead has a list of those who have that file. The actual download comes from the computers that most recently received the file themselves. Each of those computers contributes a portion of the file being sought. The client assembles the portions sent to create the complete file. At completion of the download, the new file receiver is added to the tracker so that that computer may now help with the next torrents. The more popular a file is, the more sites participate in each download.

While it is possible to only download with a BitTorrent client, you pay a performance penalty for being a “leech.” That is to say that refusing to allow uploads will result in a slower download (Etree.org, 2004a). This
download penalty is a practical solution to the “free rider problem” that occurs in so many volunteer projects.

Bt.etree.org requires similar metadata to that required by db.etree.org. Volunteers review the metadata at bt.etree.org, but those metadata are immediately viewable by all. Despite this lesser degree of gate keeping, Epstein reports that there have been only 88 banned torrents on bt.etree.org out of 3,941 total. Most of those banned torrents are a result of misunderstanding about the taping policies of the artists involved (C. Epstein, personal communication, May 12, 2004).

Etree.org contributors can and do create some sophisticated metadata within their special interest area by following some clear guidelines. The pressure of being responsible to one’s peers and to be corrected by them, if in error, might be one key factor for the high degree of reliable metadata. However, a pride in one’s collection and a willingness to share also contribute to the quality and reliability of the data and the metadata.

Of the three sites considered, Etree.org contributors face the least initial gate keeping and are offered the broadest array of technologies and forums for interacting and collaborating. Etree.org also involves the greatest number of volunteers, but interestingly enough the bulk of the volunteer work is not in the area of technology support. Instead, the volunteers mainly serve to facilitate communication and support quality assurance.

**Conclusion**

From a tightly focused technical content site to a site for recording adventures to a site for music sharing, we have seen that these three long-lived and heavily used sites follow different and somewhat unexpected models in their choices of technologies for cooperation. The Linux Documentation Project uses fairly base technologies. Confluence.org uses elaborate custom-developed software to record adventures. Etree.org seems to have left few technologies of cooperation untried.

The degree of gate keeping is also highly variable. The Linux Documentation Project operates with a strong, but writer-friendly, editorial structure that is managed using the same technologies that one might use to manage a software project. Contributors to Confluence.org and Etree.org face almost no initial gate keeping. Etree.org does rely on moderators for their discussions and assigns volunteers to be the representatives to each band in the database, but the main work of the project, the database and BitTorrents, are very open and rely heavily on trust.

All three projects report a very low rate of rejections or banning of materials or contributors, showing that the extension of trust to the community has been repaid by strong support and adherence to the behaviors defined by the projects. The overall message is that a variety of projects can be well served by extending trust to the communities of their users or clients. While these projects might have been implemented using pro-
proprietary software, the independent, open, and trusting spirit required for these projects is better represented by open-source code that allows the technologists to bend and extend the software to the particular needs of their communities.

**Table 1. Description of Technologies Discussed**

<table>
<thead>
<tr>
<th>Technology</th>
<th>What It Does</th>
<th>Who Uses It</th>
<th>Where to Find More Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>BitTorrent</td>
<td>Peer-to-peer file sharing</td>
<td>Etree.org</td>
<td><a href="http://bitconjurer.org/BiTorrent/">http://bitconjurer.org/BiTorrent/</a></td>
</tr>
<tr>
<td>CVS—Concurrent Versions System</td>
<td>Tracking and managing changes to software and documents</td>
<td>Linux Documentation Project</td>
<td><a href="http://www.gnu.org/software/cvs/">http://www.gnu.org/software/cvs/</a></td>
</tr>
<tr>
<td>Docbook</td>
<td>Open XML tag set for document publishing</td>
<td>Linux Documentation Project</td>
<td><a href="http://www.docbook.org/">http://www.docbook.org/</a></td>
</tr>
<tr>
<td>Mailman</td>
<td>Mailing list management</td>
<td>Linux Documentation Project</td>
<td><a href="http://www.gnu.org/software/mailman/">http://www.gnu.org/software/mailman/</a></td>
</tr>
<tr>
<td>mySQL</td>
<td>SQL database</td>
<td>Etree.org, Confluence.org</td>
<td><a href="http://www.mysql.com/">http://www.mysql.com/</a></td>
</tr>
<tr>
<td>PHP</td>
<td>Scripting language for creating dynamic Web pages</td>
<td>Etree.org, Confluence.org</td>
<td><a href="http://www.php.net/">http://www.php.net/</a></td>
</tr>
<tr>
<td>PHPbb</td>
<td>PHP- and mySQL-based bulletin board</td>
<td>Etree.org</td>
<td><a href="http://www.phpbb.com/">http://www.phpbb.com/</a></td>
</tr>
<tr>
<td>PostNuke</td>
<td>Content management and weblog system</td>
<td>Etree.org</td>
<td><a href="http://www.postnuke.com/">http://www.postnuke.com/</a></td>
</tr>
<tr>
<td>rsync</td>
<td>Fast incremental file transfer and archive synchronization</td>
<td>Linux Documentation Project</td>
<td><a href="http://samba.anu.edu.au/rsync/">http://samba.anu.edu.au/rsync/</a></td>
</tr>
<tr>
<td>WikkiTikkiTavi</td>
<td>Creating, editing, and maintaining Web pages collectively and collaboratively</td>
<td>Etree.org</td>
<td><a href="http://tavi.sourceforge.net/">http://tavi.sourceforge.net/</a></td>
</tr>
<tr>
<td>XML—Extensible Markup Language</td>
<td>Flexible and extensible text markup for electronic publishing</td>
<td>Linux Documentation Project</td>
<td><a href="http://www.w3.org/XML/">http://www.w3.org/XML/</a></td>
</tr>
</tbody>
</table>
There is no reason why these tried and true examples should not serve as models for future library collections and for community collaborations. Indeed, the world represented by Weblogs, Creative Commons licenses, and the Library of Science—as well as the examples in this article—point to an opportunity for a new a “Information Commons” movement (Kranich, 2004).

REFERENCES


About the Contributors

EDWARD ALMASY is Co-director of the Internet Scout Project, a research and development center in the Department of Computer Sciences at the University of Wisconsin–Madison. With almost twenty-five years of software engineering and design experience—including more than fifteen years developing technology and applications for the Internet—his primary focus is the development of new solutions for the digital library and research communities, with a recent emphasis on recommender systems and federated search mechanisms.

PETER BROPHY is Professor of Information Management and Director of the Centre for Research in Library and Information Management (CERLIM). He is currently the Director of the EDNER+ Project and has previously been director of the EDNER, LinkER, and EFX Projects. His research interests include digital and hybrid libraries; library performance assessment; and the impacts of information technologies on social exclusion. As director of CERLIM he has led over thirty projects, a significant number within JISC Programmes. His most recent books are *The Library in the Twenty-First Century* (2001) and *The Academic Library* (2000), and he has published a wide range of journal papers, reports, and book chapters. Peter was President of the Institute of Information Scientists in 1998–99 and was Chair of the North West (UK) Branch of the Chartered Institute of Library & Information Professionals in 2002–03.

AMANDA CLOSIER is a Research Officer at UKOLN, University of Bath. Her time is split between working for the several JISC-funded QA Focus IESR (Information Environment Service Registry) projects. Prior to this Amanda worked for a JISC 5/99 project as an Information Officer and has a background in library and information work. Amanda is based in UKOLN—a national center of excellence in digital information management, based at the University of Bath.
SAMUEL E. EBERSOLE, Instructional Design Specialist and Professor of Mass Communications at Colorado State University–Pueblo, has been studying new media and the convergence of digital technologies since the late 1980s. Professor Ebersole received his Ph.D. from Regent University and was awarded an Emmy award for his work with NBC Sports’ coverage of the 1988 Summer Olympic Games. Currently he is researching and teaching about electronic media, interactive multimedia and interface design, and media research. With numerous scholarly articles about media effects and online pedagogy, Professor Ebersole’s research has been cited in the New York Times on the Web and the Christian Science Monitor.

BETTINA FABOS is an Assistant Professor of Interactive Media Studies and Journalism at Miami University of Ohio. She is the author of Wrong Turn on the Information Superhighway: Education and the Commercialization of the Internet (Teachers College Press, 2004) and has written extensively about the role of the Internet in education. Fabos is also the co-author of a leading college textbook, Media and Culture (Bedfords/St. Martins) and is currently working to expand the way media literacy is taught, both at the university level and in public schools. While earning her doctorate at the University of Iowa, she received both an Iowa Presidential Fellowship and a Spencer Fellowship. Fabos is also an award-winning video producer and former print reporter. She received an M.A. at the University of Michigan and a Ph.D. at the University of Iowa.

JULIE FRECHETTE is an Associate Professor of Communication Studies at Worcester State College, where she teaches courses in media literacy, critical cultural studies, mass communication, and gender analysis. Formerly Assistant Director of the Five College Summer Institute in Media Literacy for the University of Massachusetts and the colleges at Hampshire, Amherst, Smith, and Mt. Holyoke, she has lectured and written on the importance of developing media literacy initiatives in K–12 schools and higher education. She is the author of Developing Media Literacy in Cyberspace: Pedagogy and Critical Learning for the Twenty-First-Century Classroom, as well as the co-author of the Media Literacy and Gender Equity Curriculum, which is distributed to teachers in the state of Massachusetts. Her scholarship was featured in the National Telemedia Council’s 50th anniversary edition of Telemedium: The Journal of Media Literacy in the summer of 2004. In addition, Frechette serves as the Director of Women’s Studies and the Co-Director of the Center for Teaching and Learning at Worcester State College.

JILLIAN R. GRIFFITHS gained her M.Phil. at Manchester Metropolitan University (MMU) (“Development of a Specification for a Full Text CD-ROM User Interface”) in 1996 and is currently working on a Ph.D. (“Success Factors in End User Searching of Full Text Information Sources”) also
at MMU. In 1999 she worked as Research Fellow on the DEvISE project, which has developed a methodology for evaluating Internet search engines from a user’s perspective, and since 2000 she has worked as a Research Associate on the EDNER Project and the subsequent EDNER+ Project. Jill has written a number of articles and presented at a variety of conferences, including articles on information research and Ariadne, and a presentation at the fourth DELOS workshop on Evaluation of Digital Libraries (Testbeds, Measurements, and Metrics, June 6–7, 2002, Budapest, Hungary) entitled “A Quality Management Approach to the Evaluation of Digital Library Services.”

THOMAS G. HABING is a Research Programmer at the Grainger Engineering Library Information Center at the University of Illinois, Urbana–Champaign (UIUC), where for the past six years he has worked on various digital library projects including UIUC’s Mellon-funded OAI Metadata Harvesting Project, UIUC’s D-Lib Test Suite project funded by CNRI, and UIUC’s Digital Library Initiatives I project funded by the National Science Foundation (NSF). He is currently providing programming support for the university library’s NSF-NSDL and IMLS grant projects and is the developer of the library’s OAI Registry service.

KAT HAGEDORN is OAIster/Metadata Harvesting Librarian at the University of Michigan Libraries. She is responsible for the OAIster project, a search gateway for OAI harvested records leading to digital objects, initially Mellon funded in 2001–02. She is also responsible for DLXS Bibliographic Class and co-coordinates the processing of Text Class materials. Her previous experience is in information architecture (with the Argus Associates firm) and ontology and taxonomy consulting (with the Food and Agriculture Organization in Rome). She graduated with an undergraduate degree in Biological Sciences from Cornell University and received her M.L.I.S. from the University of Michigan in 1996.

DEBRA HIOM is the Director of the Social Science Information Gateway, a nationally funded (UK) Internet service for social science research. Debra also teaches a Digital Libraries Unit on the MSc in Information and Library Management course at the University of Bristol. She has written extensively about the Internet and is co-author of the Information Professional’s Guide to the Internet and the Web-based tutorial Internet Detective.

PAUL JONES is founder and Director of ibiblio.org, the Public’s Library, at the University of North Carolina, where he also teaches on the faculties of the School of Journalism and Mass Communication and the School of Information and Library Science. Jones is also founding board member of the American Open Technology Consortium, a member of the Board of
Trustees of Chapel Hill Public Library, and a board member of the Linux Documentation Project. But he is most pleased to have been admitted into the Luxuriant Flowing Hair Club for Scientists and to have been selected in April 2003 as Best Geek in the Research Triangle by the Independent Weekly. In addition, Jones has published poetry in many literary magazines including Poetry (Chicago), Georgia Review, and Southern Poetry Review. His chapbook, “What the Welsh and Chinese Have in Common,” won the HarperPrints publication award from the North Carolina Writers Network.

BRIAN KELLY is Director of the UK Web Focus, a Web advisory post that provides support to the UK Higher and Further Education Communities and the museums, libraries, and archives sector. Brian is also the project manager of the QA Focus project and the NOF-digitise Technical Advisory Service, which provide support for digital library programs. Brian is based in UKOLN—a national center of excellence in digital information management, based at the University of Bath.

XIAOMING LIU received his B.S. (1994) in computer science from Shandong University, P.R. China; an M.S (1997) in computer science from Shanghai Jiaotong University, P.R. China; and a Ph.D. (2002) in computer science from Old Dominion University, Norfolk, VA. In March 2003 he joined the digital library research and prototyping team of the Los Alamos National Laboratory Research Library.

KURT MALY received the Dipl. Ing. degree from the Technical University of Vienna, Austria, and M.S. and Ph.D. degrees from the Courant Institute of Mathematical Sciences, New York University, New York. He is Kaufman Professor and Chair of Computer Science at Old Dominion University, Norfolk, VA. Before that he was at the University of Minnesota, both as faculty member and chair. He also is Visiting Professor at Chengdu University of Science and Technology, People’s Republic of China, and is Honorary Professor at Hefei University of Technology, P.R. China. He was a member of the Board of the Microelectronic and Information Sciences Center in Minneapolis. His research interests include digital libraries, very high-performance networks protocols, interactive multimedia remote instruction, and Internet resource access. His research has been supported by DARPA, NSF, NASA, CIT, ARPA and the U.S. Navy, among others.

STEVE MITCHELL is iVia and Data Fountains Projects Coordinator at the Library of the University of California at Riverside (UCR). A co-founder of INFOMINE, a virtual library of significant research and educational resources of value to the academic community and one of the first Web-based services offered by a library, Mitchell has been a Science Reference
Librarian at UCR since 1986. He is a graduate of UC Santa Barbara (B.A., Sociology, 1977) and UC Berkeley (M.L.I.S., 1986).

MICHAEL L. NELSON received his B.S. (1991) in Computer Science from Virginia Tech and his M.S. (1997) and Ph.D. (2000) in Computer Science from Old Dominion University. He worked at NASA Langley Research Center from 1991 to 2002. Through a NASA fellowship, he spent the 2000–2001 academic year at the School of Information and Library Science, University of North Carolina at Chapel Hill. In July 2002 he joined the Computer Science Department of Old Dominion University. He is a co-editor of the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) and has developed many digital libraries, including the NASA Technical Report Server. His research areas include digital objects and digital archives, as well as the relationships between them.

SARAH L. SHREEVES is Visiting Assistant Professor of Library Administration and Project Coordinator for the University of Illinois IMLS Digital Collections and Content Project. Previously she was a Project Coordinator for the University of Illinois Open Archives Initiative Metadata Harvesting Project funded by the Andrew W. Mellon Foundation. From 1992 until 2001 she was a member of the staff at the Massachusetts Institute of Technology Libraries. She has a B.A. in Medieval Studies from Bryn Mawr College, an M.A. in Children’s Literature from Simmons College, and an M.S. in Library and Information Science from the University of Illinois.

JILL WALKER is an Associate Professor at the Department of Humanistic Informatics at the University of Bergen. Her research is primarily on networked communication in weblogs and other online forums, with a particular focus on literary and artistic aspects of the net. Her weblog, jill/txt, was one of the first academic blogs and has earned sufficient links to make the site the top hit for “Jill” on Google.

JEFFREY A. YOUNG is a Software Architect for the OCLC Office of Research. He has worked at OCLC since 1987 and in the Office of Research since 1996. He holds a B.S. (Computer Science) from Ohio State and M.L.S. (Beta Phi Mu) from Kent State. Current research interests include web services, interoperability, and authority control. He first got involved with OAI in association with the Networked Digital Library of Theses and Dissertations, and was a member of the OAI Technical Committee that helped develop the OAI-PMH specification.

MOHAMMAD ZUBAIR has more than sixteen years of research experience in the area of experimental computer science and engineering both
in academia as well as in industry. His current research interests are digital libraries and eCommerce. In his tenure at the university, he has developed several software systems. Two of his research efforts have led to source code license agreements with major companies. His major industrial assignment was at IBM T. J. Watson Research center for three years, where some of his work was integrated into IBM products. His research has been supported by NASA, NSF, ARPA, Los Alamos, AFRL, NRL, JTASC, SANDIA, and IBM Corporation.
Index to Volume 53

Prepared by Susan Kelsch

Page references in boldface indicate major treatment of a topic. Italic page references indicate an illustration. Italic t, f, or n indicates information in tables, figures, or notes.

Academic libraries
  collaborating with other libraries, 440
  collection development software, 609, 611
  compensation issues, 104
  Harvard College Library, 164
  INFOMINE (virtual library service), 606
  learning organizations, 62
  organizational culture, 42, 43–45
  organizational development, 5–16, 244
  performance data, 130–131, 134–135, 136, 139–149
  process improvement, 94
  Ryerson University Library, 166
  Texas A&M University, 139–140
  Tri-College Libraries (Philadelphia), 227–228
  University of California, Riverside, Library, 604, 606
  University of Connecticut Libraries, 12–13
  University of Illinois, 578, 580–581
  University of Maryland Libraries, 62
  University of Nebraska-Lincoln, 62–66
  University of Pennsylvania, 145–146
  University of Virginia, 12, 146–147, 164–166
  University of Washington, 147–149
  user surveys, 160
  See also University of Arizona Library

Access issues
  children and Internet, 555–556
  classroom Web use, 530–531
  health information, 266, 278–280, 283–300, 316, 402, 434–452, 482, 493
  Iowa City Public Library, 497
  slow connections, 549
  video materials, 276

Accountability, job. See under Performance evaluation

Accreditation
  American Accreditation Healthcare Commission, 370
  library performance data, 150

ACRL. See Association of College and Research Libraries

Activity Book for African American Families: Helping Children Cope with Crisis, 295

ADA (American Diabetes Association), 294–295

Adhocracy-oriented culture, 39f, 40, 45
Administration and management
academic libraries, 44
assessing practices, 28, 163
compensation issues, 100–105
flat vs. hierarchical structure, 61
hospitals, 412
job redefinition, 168–169
vs. leadership, 190, 191
library performance data, 129–155
National Institutes of Health, 379
organization charts, 119–122
organizational culture, 39
performance management, 97
restructuring, 239
skill sets, 49
strategic planning, 11
systems design theory, 69
Teton County (WY) Library, 113,
117, 125
training, 250
University of Arizona Library, 72,
81, 86, 89, 99, 106
See also Leadership
Adult education. See Continuing education
Advertising
click-throughs, 527
content vs., 537
Iowa City Public Library workshop,
507
search engines, 520
targeting children, 555, 558–559,
564–565, 570–572
Aerospace information, 597–598
African Americans, health information,
283, 291–295, 296, 398,
400–406
African Americans and Diabetes (ADA),
295
African Americans Reach and Teach
Health (AARTH), 398, 401–406
Aftab, Parry, 561–562
Age
Iowa City Public Library workshop
on aging, 507
library communities, 271
proof-of-age software, 559
See also Elderly
Age Pages (NIA), 290
Aggregators (metadata), 597–598
ALA. See American Library Association
Algorithms, search engines, 525,
526–527, 536
Alternative medicine, 309, 310, 405,
492
AMA. See American Medical Association
Amazon.com, 548, 629
America Online, 559–560, 565
American Accreditation Healthcare
Commission (URAC), 370
American Association of Retired
People (AARP), 290
American Diabetes Association (ADA),
294–295
American Federation for Aging Research, 290
American Library Association (ALA)
performance measurement, 144–145
public library budgets, 113
Spectrum leadership program, 197
American Medical Association (AMA),
331
AmericanSouth.org, 578
Andrew W. Mellon Foundation, 578,
621–622
Animals, in library promotions,
115–116
Anticipatory learning, 56
AOL@School service, 559–560
Apple Computer, 23
Appreciative inquiry, 218–229, 220,
222
Appropriate subject matter
health information, 274–275
Web, 555, 564–565, 567
Arc (OAI search service), 591–599,
592t, 593f, 594f, 602
Archives
contributor-run archives, 651–662
Open Language Archives Community, 579
Scout Archives, 622
Archon project, 600
Area Health Education Centers
(AHEC), 313–314
Argyris, Chris, 56, 233–234
Arizona. See University of Arizona Library
ARL. See Association of Research Libraries
Articles. See Literature, professional
Arts and Humanities Data Service (AHDS), 640
Ashcroft v. ACLU (2004), 567
Assessment
appreciative inquiry, 224
compensation issues, 104
culture of, 136–138
digital library usage data, 144
gap analysis tool, 139–140
health information, 417, 430, 455
health information outreach projects, 442–444, 460–461
instructional design, 354
Iowa City Public Library, 501, 509–511
learning organizations, 60–61
metadata quality, 648, 649
organizational culture, 37–38, 49
organizational development, 91, 13, 27, 156–171, 247–248
process improvement, 91, 92, 95
University of Arizona Library, 71, 98–99
University of Nebraska-Lincoln, 64
University of Washington Libraries, 147–149
user needs, 269–272, 302, 400–401, 403
Web filtering software, 566
Web information, 532, 535, 551–552, 560–562, 570, 572–573
See also Information and data; Surveys and focus groups

Association of College and Research Libraries (ACRL)
compensation survey, 104
Leadership Institute, 197t, 210
Leadership Survival Kit, 198t

Association of Research Libraries (ARL)
leadership training, 192, 196, 197t, 203
library statistics, 134–135, 145, 150, 159
New Measures Initiative, 139
Asynchronous communications, 351
Audio and video materials
health information, 276, 290, 295, 307, 323–324
music recordings, 656–659
searching strategies, 548
Web-based instruction, 352

Aurora Leadership Institute (Australia), 197t

Authors
Argyis, Chris, 56
Beckhard, Richard, 6
Cubbins, E.M., 294
de la Peña McCook, Kathleen, 302
Grudin, Robert, 20–21
Howard, Ellen, 303
Lancaster, E.W., 131–132
Light, Paul, 19
Rees, Fran, 230–231
Sample, Steven, 249
Schön, Donald, 56
Schwartz, Charles, 8
Schwarz, Roger, 230, 231, 233, 234, 244
Senge, Peter, 56–57, 69, 70, 112–113, 114, 122–123, 124, 225

Automation
automated library systems, 132–133
collection development software, 606–607, 608, 615, 616, 617
ERRoLs (Extensible Repository Resource Locators), 583–585
Internet gateway surveys, 640, 644–645
OAI metadata harvesting, 581, 582
Award, Malcolm Baldrige National Quality, 41

Baby boom generation, 490–491
Backus, Joyce, 332
Balanced Scorecard (assessment tool), 11, 12–13, 37, 141, 160–161
Bandage holders, 507–508
Barriers to access. See Access issues
Beckhard, Richard, 6
Behavior and psychology
appreciative theory, 221, 222–223
leadership, 189
psychological contracts, 248
society after September 11 attacks, 488

Benchmarking
compensation issues, 104
library performance, 140
online gateway quality assurance, 640
organizational development, 27
Bertlsmann Foundation International
Network of Public Libraries, 197
Bibliographies
medical textbooks, 340–341
Open Archives Initiative citations, 582
See also Links, Web
Bilingual services
health care, 312
Hmong Health Web site, 308–309
Neuro-Patient Resource Centre, Montreal, 428–429
Bill and Melinda Gates Foundation, 284
Biomedical libraries. See Medical libraries
BitTorrent (file-sharing system), 658–659
Blocking software. See Filtering and blocking software
Blogs, 313, 537
BOBBY® (Web site tool), 289
Books, physical aspects, 275
Boolean operators, 400
Boston Medical Library, 436
Boston Public Library, 435–436
Brainstorming
instructional design, 355
process improvement, 93
Branch libraries, 57
Branding, 519, 613
Brochures. See Pamphlets and brochures
Browsing. See Search and browse strategies
Bryn Mawr College, 227–228
Budget. See Finance and budget
Building issues, Teton County Library (WY), 116
Bureaucracy, 61, 406, 450
Bush, George W., 557
Business and private sector
appreciative inquiry, 220
change management, 183
innovation, 19, 23, 37
organizational development, 240–241
search services, 416–417
self-managed work teams, 178–179
Web commercialization, 519–520, 533, 533t, 557–558, 564–565
Web filtering in workplace, 568–569
C. Everett Koop Community Health
Information Center (Philadelphia), 458–459
Cable television
Internet service providers, 569
Iowa City local access, 505–506, 507, 509
Caches (metadata harvesting), 598
California
California Continuing Education Plan, 191
California Library Association, 193
California State Library, 192
Oakland, 278
San Jose Public Library, 194
Stanford-California Institute, 200t, 208, 209, 210, 211–212
University of California, Los Angeles, 196, 201t, 292
University of California, Riverside, 604, 606
University of California, San Francisco, 332
University of Southern California, 249
Campbell, David P., 27–28
Campbell Organization Survey (COS), 27–28
Canada
health practices, 429
library usage, 432
Neuro-Patient Resource Centre, Montreal, 425–431
Ontario public libraries, 473–474
Ryerson University Library, 166
Toronto Public Library, 194
Toronto Reference Library, 474–475
University of Alberta, 200t
CaPHIS (Consumer and Patient Health Information Section), 275
Career development
assessment efforts, 166–170
change management, 182
leadership training, 210–211
staff learning plans, 227–228
Career Development and Assessment Center for Librarians (CDACL), 194
Caregivers, family, 491
Cary Area (IL) Public Library District, 163–164
Case studies
assessment projects, 163–164
Hmong health information, 316–323
library statistics usage, 138
Neuro-Patient Resource Centre, Montreal, 425–431
North Suburban Library System, 177–178
Social Science Information Gateway, 642–648
University of Nebraska-Lincoln, 64
Cataloging
Collection Workflow Integration System, 624, 628–629
Social Science Information Gateway, 643–645, 646
CBO. See Community-based organizations
CCNet (Cochrane Communication), 484
CDACL. See Career Development and Assessment Center for Librarians
Celebrities and health information, 295
Cellular phones, 30
Censorship, 556, 557, 565–566, 570
Censorware. See Filtering and blocking software
Census information, 270, 499t
Center for Children with Special Needs, Seattle, 466–467
Center for Creative Leadership, 28
Centre for Health Information Quality (CHIQ), 371
Centre for Research in Library & Information Management (CERLIM), 540
Centre for Studies in Advanced Learning Technologies (CSALT), 540
Change
academic libraries, 44
commitment to, 15
communications issues, 23, 162
continuous process, 5
human resources role, 174, 181–183, 185
incremental steps, 10
leadership training, 211–212
learning organizations, 60
organizational development, 238–257
organizational structure, 158–159
process improvement, 93
staff education, 50
Teton County Library (WY), 127
University of Arizona Library, 74, 145
University of Nebraska, 66
University of Virginia Library, 166
University of Washington Libraries, 148
URLs, 618
Web site updates, 595
Characters (text), 647
Charles R. Drew University of Medicine, 292
Chicago suburb libraries, 163–164, 177–178
Children
Internet usage, 555–556, 557, 558–559, 564, 570–572
society after September 11 attacks, 488
See also Students and learners
Children’s Hospital and Regional Medical Center, Seattle, 466–467
Children’s Internet Protection Act (CIPA, 2000), 556–557, 567
Children’s Online Privacy Protection Act (COPPA, 2000), 556, 558
CHIQ. See Centre for Health Information Quality
Churches. See Religious organizations
Clan-oriented culture, 39f, 40, 42, 45
Classification systems
Collection Workflow Integration System, 626
job classification, 65
Classroom environment, 351–352, 530–531, 557, 564
See also Students and learners
Cleveland Clinic, 220
Click-through advertising, 527
Clientele, library. See User services and needs
ClinicalTrials.gov (NIH), 381
Clinton, Bill, 556
Cochrane Communication, 484
Colgate Palmolive Co., 571–572
Collaboration
  Cochrane Communication, 484
  contributor-run archives, 651–662
  health literacy, 313, 454–455
  nature of libraries, 245
  online collection building, 604–619
  Philly Health Info project, 463
  public and medical libraries, 411–421, 467–468
  Web searching, 542
Collection development
  health information, 275–277, 278–280, 404, 417, 455
  MedlinePlus en español, 382
  MedlinePlus® Health Topic pages, 377, 384
  Neuro-Patient Resource Centre, Montreal, 428–429, 430
  online collaboration software, 604–619, 620–636
  Teton County Library (WY), 120–121
Collection Workflow Integration System (CWIS), 620–636, 625f, 628f, 630f, 632f
College of Physicians of Philadelphia, 457–460
Colleges. See Universities and colleges
Colorado Consumer Health Information Libraries Listserv (CCHILL), 411–412
Commercialization, Web, 519–521, 523, 530–538, 555–575
Committees, library
  assessment projects, 160
  teams vs. committees, 179, 180t
Communications issues
  assessment projects, 161–162
  facilitative leadership, 233–234
  health information, 332, 407, 426–427, 476, 484
  health information outreach projects, 450, 455
  language of librarianship, 35, 401
  leadership, 190
  limitation of librarians’ role, 475
  managers and employees, 168–169
  Open Archives Initiative, 586
  organizational dialogue, 28–29
  performance management, 98
  Teton County Library (WY), 125
  University of Arizona Library, 88–90
  user service changes, 22
  Wake County (NC) Public Library System, 236
  Web instruction, 351–352
Community access television, 505–506
Community-based organizations (CBO)
  Harm Reduction Coalition, 280
  health information, 279, 333, 436–438, 440, 444, 445–449f, 472, 505
Community-based participatory research, 437–438
Community involvement
  alternate health information sources, 277
  community opinion leaders, 272–274
  digital library communities, 578
  health information needs, 269–272, 334, 397–410, 436–437, 438
  Iowa City Public Library, 497, 508–509, 511
  Teton County Library (WY), 115–116, 127
Compassion (facilitative leadership), 233
Compensation, employee, 100–105
Competing Values Framework (CVF), 37–43, 39f, 48
Complimentary medicine. See Alternative medicine
Computer industry
  innovation, 23
  Web content control, 557–558, 564–567
Computer skills training, 441, 468, 504
Concert recordings, 655–659
Conferences and workshops
  health care professionals, 320
  health fairs, 316–319, 318
  Internet/Online Summit, 556
  Iowa City Public Library outreach, 506–507
leadership, 196–204
library performance data, 150
Missouri health librarians, 414
Santa Fe Workshop (1999), 591
Social Science Information Gateway editors, 645
“Stake Your Claim to Health Literacy” (symposium), 328–329, 331–334
Web health information workshops, 348–359
Conflict, managing, 78
Confluence.org. See Degree Confluence Project
Connecticut
University of Connecticut Health Center, 477
University of Connecticut Libraries, 12–13
Conservatism, 565
Consortia
health information databases, 415
Sheet Music Consortium, 579
Constructive Dialog (tool), 79
Consultants and specialists
Hamel, Gary, 26, 30
leadership trainers, 207
library assessment, 159–160, 167
McClure, Charles, 98
organizational development specialists, 30, 74
Teton County Library (WY), 125–126
Consumer and Patient Health Information Section (CaPHIS), 275
Consumer health information. See Health information
Consumer Health Information Network (CHIN), 436
Consumer Health Information Resource Service (CHIRS), 413
Consumer Health Information Service (CHIS), 474–475
Content management
document management, 653, 657
Etree.org, 657–659
Internet Scout Project, 621–622
iVia (collection development software), 611, 613–618
metadata harvesting, 577
Social Science Information Gateway, 643–647
Content producers. See Information providers
Contextual literacy, 337
Continuing education
health education, 316–319
leadership, 191
role of public libraries, 127
Contributor-run archives, 651–662
Cooperrider, David, 220
Copyright
Hmong Health Web site, 323
Open Archives Initiative projects, 587
Web links, 526
Core competencies. See Skill sets, library staff
Corporations. See Business and private sector
COS (Campbell Organization Survey), 27–28
Cost-of-living adjustments, 103
Council on Library Resources, 196
COUNTER (Counting Online Usage of NeTworked Electronic Resources), 143
Court case, Ashcroft v. ACLU (2004), 567
Crawlers. See Web crawlers
Credibility
health information Web sites, 364, 377
Neuro-Patient Resource Centre, Montreal, 430
Crime, 557
Critical thinking
Web commercialization, 572–573
Web health information, 360–361, 362–365, 373
Cross-functional teams
Teton County Library (WY), 120, 121–122
University of Arizona Library, 73
Cubbins, E.M., 294
Cultural issues
African Americans and health information, 293–294
cultural profiles, 272–273, 273t
health care, 312
health literacy, 397–410, 429
Hmong and health information, 306–309, 310, 320–321
refugees and health information, 303–304
Currency
health information, 430–431
online subject gateways, 646
Customer service. See User services and needs
Customization
collection development software, 605, 613, 631–633, 634, 641
medical information, 476
CVF. See Competing Values Framework
CWIS. See Collection Workflow Integration System (CWIS)
CyberPatrol (filtering service), 563
Cybersitter (Web blocking software), 566
CyberTipLine, 558
Data Fountains (collection development software), 606–607, 608f
Data-gathering. See Information and data
Databases
iVia (collection development software), 610–611, 614
medical information, 415
MedlinePlus® publishing system, 386
student searching research, 546, 548
dela Peña McCook, Kathleen, 302
Death and dying, 310
Decision making
core values, 232
library statistics, 134–135
Decision support systems (DSS), 132–134
Definitions
appreciate inquiry, 219
health, 454
health information, 274
innovation, 18
leadership, 187–188, 193, 212–213
library communities, 269, 271
literacy, 423
Medical Subject Headings, 285
organizational culture, 34
organizational development, 6–7, 7t, 241, 249
quality, 550t
Degree Confluence Project, 654–655, 659, 660t
Deming, W. Edwards, 69
Democracy
cyber-democracy, 565
role of public libraries, 127
Demographics
African Americans, 291
baby boom generation, 490
elderly, 285
immigrants, 304
Internet users, 461
Iowa City, 499t
librarians, 191–192
library communities, 269, 272, 399
Department of Health and Human Services, 289
See also Healthy People 2010 (health objectives)
Des Plaines (IL) Public Library, 163–164
Diabetes in African Americans (NIDDK), 295
Diet and nutrition, 483
Digital Gateway to Cultural Heritage Materials (U. of Illinois), 578
Digital libraries
Arc (OAI search service), 590–603, 592t
INFOMINE (virtual library service), 606
Joint Information Systems Committee projects, 637–650, 639t
OAI Protocol for Metadata Harvesting, 576–589
organizational culture, 47–48
performance assessment, 142
University of Arizona Library, 108
Digital Library Federation (DLF), 138
Digital Library Grid, 601–602
Digitization projects, 143
Directories
medical libraries, 382
Philly Health Info (PHI), 460
Disabled Web users, 289
Diseases and conditions
baby boom generation, 491
health information, 426, 430–431, 474–475
heart health, 324
immigrant population, 310, 311
medical textbooks, 341t, 345, 345t
Metabolic Syndrome X, 483
unexplained symptoms, 487–488
Disney Corp., 571, 572
Disruptions to library service, 22, 24
Distribution, language specialists, 275
Diversity issues
- health information, 268–282, 301–328, 398, 400–406
- Iowa City Public Library, 499
- library leadership, 192, 215
- manager skill sets, 49
- organizational assessment, 158
- Teton County Library (WY), 115
See also Languages, other than English
Doctors. See Physicians
Document management. See under Content management
Domains (Web), 533, 534
Downloading, music, 658–659
DP9 (metadata gateway service), 600–601
Drebbel, Cornelis, 20
Drugs. See Medications; Substance abuse
Dublin Core, 585, 595, 623
DVD. See Audio and video materials
E-commerce, 520
E-mail
- Collection Workflow Integration System, 628–629
- health information outreach projects, 450
- patient-physician communication, 484
- Social Science Information Gateway, 644–645
- University of Arizona Library, 75, 77–78, 84
Editors
- The Linux Documentation Project, 653
- Social Science Information Gateway, 643–647
Education
- education levels and health literacy, 338–339, 341, 345
- educational theory, 399
- Pygmalion Effect, 221
- Web filtering software, 568
- Web use, 530–538, 557, 564
See also Continuing education
EDUCAUSE Leadership Institute, 197
Ela Area Public Library (Lake Forest, IL), 163–164
Elderly
- aging, 507
- baby boom generation, 490–491
- health information, 283, 285–291, 296, 505
- library communities, 271
- retirement planning, 506
Electronic books, 295
Electronic vs. print information, 58, 142–144
Emory University, 578
Employment. See Staff, hiring, and employment issues
England. See United Kingdom
English as a Second Language (ESL) programs, 315
Ephemeral materials, 279
Equations (metadata searching), 600
ERRoLs (Extensible Repository Resource Locators), 583–585, 587
Errors
- organizational development, 56
- Social Science Information Gateway, 644, 647
ESL. See English as a Second Language (ESL) programs
Ethics, 475
Ethnicity. See Diversity issues
EthnoMed, 303, 399–400
Etree.org, 655–659, 660
Evaluation. See Assessment; Performance evaluation
Everquest (online game), 527
Exhibits
- health information, 317
- Johnson County (IA) fair, 508
- Experimental OAI Registry (U. of Illinois), 580–581
- Expert patient program (England), 485
Export-import formats, 633
Extensible markup language. See XML (Extensible Markup Language)
Facilitative leadership, 230–237, 234
- group process improvement, 239
- University of Arizona Library, 75, 77–78, 84
Faculty, college leadership trainers, 207
librarians, 103
University of Virginia, 147
University of Washington, 148
See also Teachers
Fair, Johnson County (IA), 508
Faith-based organizations. See Religious organizations
Families
health caregivers, 491
Hmong culture, 305
Feedback
employee performance management, 97
health information seeking, 468, 471
instructional design, 355
leadership training, 205, 205t
MedlinePlus®, 386
Philly Health Info project, 460, 462
QA Focus projects, 640
360-degree evaluation, 235, 252
Web health information training, 503–504
The Fifth Discipline: The Art and Practice of the Learning Organization
(Senge), 57–58, 70, 112, 122, 225
The Fifth Discipline Fieldbook (Senge), 70
File-sharing systems, 658
Filtering and blocking software, 558–564, 565–570
Finance and budget
health care costs, 424
health information, 275, 325–326, 406, 415, 418, 439
Hmong Health Web site, 323–324
hospital marketing, 412
leadership training, 207
library performance assessment, 158, 161
National Network of Libraries of Medicine projects, 434, 443–444, 451
Open Archives Initiative projects, 578
organizational development, 9t, 15t, 49
public libraries, 113–114
Teton County Library (WY), 115
University of Arizona Library, 88
University of Pennsylvania Library, 146
Flesch-Kincaid Grade Level (assessment tool), 341, 342–343, 343t
Flesch Reading Ease (assessment tool), 341, 343–344, 343t
Focus groups. See Surveys and focus groups
Formats
Collection Workflow Integration System, 633
iVia (collection development software), 611
metadata, 585, 595
4-D Cycle model, 223–225
Francis A. Countway Library of Medicine, 436
Free Library of Philadelphia, 460
French language materials, 428–429
Frye Leadership Institute, 198t
Full text processing, 615
Funding. See Finance and budget
Future of libraries, 252–253
Galbraith’s Star Model, 11
Gale Group, 485
Gallup Corporation, 65–66, 167
Games, Everquest, 527
Gateways, portals, and repositories
AOL@School service, 559–560
Degree Confluence Project, 654–655
Digital Gateway to Cultural Heritage Materials, 578
Etree.org, 656–659
INFOMINE (virtual library service), 606
iVia (collection development software), 604–606
Joint Information Systems Committee projects, 637–650
The Linux Documentation Project, 652–654
Resource Discovery Network, 638
Scout Portal Toolkit, 620–636
Gemba (system design concept), 69
Genetics Home Reference (NLM), 381
Geography, of library users, 271
Georgia
Emory University, 578
Gerould, J.T., 130
GetNetWise (Web site), 558
INDEX 679

Gillispie, Mary Alice, 324
Girls, Iowa City Public Library workshop, 507
Go Local project (NLM), 387
Goals
human resources, 173
instructional, 355
Iowa City Public Library, 497, 501
job redefinition, 168
Neuro-Patient Resource Centre, Montreal, 425
organizational development, 15
performance management, 97–98
teams vs. committees, 180t
University of Arizona Library, 72, 82, 86
Google Inc.
branding, 519, 520
industry mergers, 536
Open Archives Initiatives repositories, 581
robots, 595, 596t, 597t
student Web searching, 539, 545–546, 550–551
Web links strategy, 525, 526–527
Goto.com, 520
Government agencies
alternative medicine, 492
collaboration, 379, 445–449t
innovation, 19–20
organizational development, 242
The Grace of Great Things (Grudin), 20
Grateful Dead, 655–656
Group learning, 64–65
Grudin, Robert, 20–21
Hackers, computer, 566–567
Hamel, Gary, 26, 30
Hardware, 634
Harm Reduction Coalition, 280, 280, 281
Harvard University
Harvard College Library, 164
Leadership Institute, 197t, 210
Haverford College, 227–228
Health and Wellness Web site (AARP), 290
Health care
alternative medicine, 492
costs, 424
health information collaboration, 314, 437, 438
immigrants and refugees, 311–312
impact of September 11, 487–489
national health issues, 453–454, 456, 483
in the news, 490
patient responsibility, 464–465
role of information, 402, 484
Health care providers
health information training, 319–320
language issues, 308
role of librarians vs., 417, 475, 491–492
Web sites, 380
See also Hospitals and clinics
Health Compass (AFAR), 290
Health fairs, 316–319, 318, 509
Health Info Iowa, 500, 500f
Health information, 265–267, 268–282, 279f, 395–396, 510–511f
cultural issues, 397–410
health literacy, 329–335, 422–433
immigrant and refugee communities, 301–328
Iowa City Public Library, 496–511
medical textbooks, 336–347
MedlinePlus®, 375–388
outreach projects, 434–452, 453–456
Philadelphia project, 457–463
public and medical libraries, 411–421
public libraries, 480–495
reference services, 464–479
vulnerable communities, 283–300
Web-based workshops, 348–359
Web information seeking, 360–374
Health Information Literacy Task Force, 331–332
Health insurance, 311–312
Health librarians, 453–456, 466
Health literacy, 337–340, 422–433
cultural factors, 397–410
ESL community, 315
Hmong immigrants, 309–310, 317–319, 320
information accessibility, 266
language issues, 329–335
national social issues, 480
role of librarians, 419
role of public libraries, 481–493
See also Consumer health information
Health on the Net (HON), 369–370, 503
Health Round Table (HEART), 414
Health Summit Working Group, 370
Health Topic pages (MedlinePlus®), 376–379
HealthInfoQuest (Web site), 349, 349
Healthnet (U. of Connecticut), 477
Healthy People 2010 (health objectives)
  health indicators, 483
  health literacy defined, 330–331, 337, 425
  prioritizing health issues, 453–454, 456
Healthy Roads Media project, 324
Heart health information, 324
Herbal remedies, 492
Hierarchical management structure, 39f, 40, 61, 88
Hierarchical subject gateways, 551
Hiring. See Staff, hiring, and employment issues
Hispanic Americans, health information, 310–311
History
  appreciative inquiry, 220
  health information, 265
  health literacy, 330
  Hmong, 304–305
  Internet safety, 556
  leadership, 191
  library statistics, 130–135
  National Library of Medicine, 375–376
  Open Archives Initiative, 591
  organizational development, 8
  process improvement, 91
Hmong Health Care Professionals Coalition, 316–317
Hmong Health Education Network, 302
Hmong Health Information Promotion, 302
Hmong immigrants, 301–328
HON (Health on the Net), 369–370, 503
Hoshin Planning, 11, 85–88
Hospitals and clinics
  Children’s Hospital and Regional Medical Center, Seattle, 466–467
  immigrants, 311–312
  medical libraries, 412, 445–449t
  Michael Callen-Audre Lorde
  Community Health Center, 271
  organizational culture, 41, 220
  patient health literacy, 337, 338–339
  under-served communities, 314
See also Health care providers
Howard, Ellen, 303
HTML (Hypertext Markup Language), 583
Human resources, 172–186
IBM Corp., 23
Identification, Web, 559
Illinois
  Chicago suburb libraries, 163–164, 177–178
  Illinois State Library, 489
  University of Illinois at Chicago, 437
  University of Illinois Graduate School of Library and Information Science, 352
  University of Illinois Library Research Center, 489
IMLS (Institute for Museum and Library Services), 141
Immigrants, health information, 301–328
Implementation issues
  Open Archives Initiative Protocol for Metadata Harvesting, 586
  organizational development, 10–12
Import-export formats, 633
Improving Health Professionals’ Access to Information (NLM), 438
INFOMINE (virtual library service), 606, 607f, 611, 612, 615, 619
Information and data
  census information, 270, 499t
  job description analysis, 176–177
  leadership training, 213
  library performance, 129–155, 159, 162–163
  process improvement, 92, 95
  teams vs. committees, 180t
  understanding core values, 232
University of Arizona Library, 88–89, 90
Information Environment (UK), 539–540, 638
Information literacy, 149
Information providers
advertising vs. content, 537
commercial health Web sites, 377
health information, 373, 416, 435, 485
licensed content, 380
metadata available to repositories, 577–578, 594–595, 596–597
See also Publishing
Information schools. See Library and information science graduate schools
Information seeking behavior
African Americans, 291–292, 402–403
cultural issues, 397–398, 405
elderly, 285
health information, 360–374, 435, 468, 469, 482
Iowa City Public Library project, 509, 510–511
Philadelphia health information project, 461
student searching research, 545–552
See also Search and browse strategies
Information technology, 8, 9t
Innovation, 17–32
See also Technological innovation
Institute for Museum and Library Services (IMLS), 141
Institute of Learning and Research Technology (ILRT), 639, 642
Instructional design, 351–358
Interactivity, 541
Interlibrary loan, 469
International Adult Literacy Survey, 423–424
Internet. See Web and Internet
Internet Content Rating Association (ICRA), 561
Internet/Online Summit, 556
Internet Scout Project, 621–622
Internet service providers, 559–560, 565, 568, 569
Interoperability, 590–591, 646–647
Interpreters, 312
Interviews
appreciative inquiry, 224
health information outreach projects, 440
health information seeking, 291–292
health terminology, 338
organizational development, 6, 6t
See also Reference interviews
Inventors
Drebbel, Cornelis, 20
Shewhart, Walter, 91
Iowa
Iowa City Public Library, 496–511, 502
Iowa Consumer Health Information Project (I-CHIP), 500
State Library of Iowa, 500
IQ-Tool (Web assessment), 370
iVia (collection development software), 604–606, 605f, 607, 609–619
Jackson Hole (WY), 114
Jargon. See Terms and subject headings
Job descriptions, 117–118, 167, 169, 175–178
Johnson County (IA), 498, 499t, 505, 508
Joint Information Systems Committee (JISC), 552n, 637–650
Journal of Cardiovascular Nursing, 417
Journal of MedSurg Nursing, 417
Journals
online subject gateway, 647
print vs. electronic, 142
student searching research, 546–547
Kansas
health information librarians, 414
Kaplan, Robert, 11
KEYS to Creativity (assessment tool), 28
Keywords. See Terms and subject headings
Kid’s Wave (Web site), 571
King’s College, London, 640
Kiosks, 460, 461–462
Knowledge, in Web links, 528
Kraus, Jan, 314
Kurtz-Rossi, Sabrina, 332
Lancaster, F.W., 131–132
Lancaster University, 540
Languages, other than English
community needs, 269
health information, 275, 276,
279t, 306–309, 319–320, 429
health terminology, 266
mandatory forms availability, 278
Teton County Library (WY), 115
See also Communications issues
Laos, 304
Latinos, Teton County (WY) Library
patrons, 115
Laurinburg (NC), 235–236
Leadership
academic libraries, 44
appreciative inquiry, 223, 227
assessment, 137, 161–162
characteristics of, 189t
facilitative leadership, 230–237
human resources, 184
innovation, 20, 31
learning organizations, 59
vs. management, 191t
Organizational Culture Assessment
Instrument, 39, 49, 50
organizational development, 6, 9t,
12, 238–257
training programs, 187–217
University of Arizona Library, 73,
75–80, 84
See also Administration and man-
agement
The Leadership Challenge (Kouzes and
Posner), 189t
Learning organizations, 54–66
appreciative inquiry, 225–226
change management, 181
Teton County Library (WY), 112–
113, 123, 127
University of Arizona Library,
70–74, 100
Left-hand column exercise, 233–234
Legislation, children Internet protec-
tion, 556–557, 558, 567
LibQUAL+R (survey tool), 13, 139,
148, 160
Librarians
Kraus, Jan, 314
Phipps, Shelley, 116
Stoffle, Carla, 113, 136, 145
Wescott, Beth, 332
Librarianship
academic librarians, 609
expertise vs. play, 25
health advocates, 455
language and artifacts, 35
leadership, 191–192
limitations of role, 475, 478
marketing importance of, 415, 419
organizational development,
244–245
patron communications, 400
professionals and paraprofession-
als, 44
qualities of, 406, 476
reference services, 472–473
understanding the community,
398, 405
worldwide, 432
Library Administration and Associa-
tion Management (ALA), 198t
Library and information science graduate
schools
leadership training, 213–214
research subjects, 546–547
University of Illinois, 352
University of Pittsburgh, 291–292
University of Texas, 292
Library Channel (Iowa City), 505–506,
507
Library directors, 120, 124, 126
Library Leadership Institute, 198t
Library Leadership New Mexico, 200t
Library Leadership Ohio, 198t
Library of Congress, 198t
Licensed content, 380
Light, Paul, 19
Links, Web
Degree Confluence Project, 654
Iowa City Public Library Web site,
502, 502f
MedlinePlus® Web links, 376,
379, 380, 382, 387
Open Archives Initiative reposi-
tories, 581
Philly Health Info, 459–460
power of, 524–529
Social Science Information Gate-
way, 645
The Linux Documentation Project. See
TDLP (The Linux Documenta-
tion Project)
Listservs. See Mailing lists and listservs

Literacy
- adult literacy programs, 487
- Hmong and health information, 306–309
- national levels of, 423–424, 486
- Web design tools, 289
See also Health literacy

Literature, professional
- health information, 416–418
- health literacy, 330, 338–340, 366–368
- health research, 381
- leadership, 193
- organizational development, 241–249
- Web searching, 540–541

Llamas, 115–116

Local health departments, 314

LookSmart, Ltd., 568–569

Los Alamos National Laboratory
- library organizational development, 30
- Static Repository Gateway, 586–587

Machine processing. See Automation

Macroevaluation of data, 131–132

Mailing lists and listservs, 313, 411–412, 413, 414

Maintenance learning, 55–56

Making Your Web Site Senior Friendly (NLM, NIA), 289

Malcolm Baldrige National Quality Award, 41

Managed health care programs, 314–315

Management. See Administration and management

Management by Objectives (planning system), 86

Management information systems (MIS)
- library performance data, 132–134
- University of Virginia, 147

Management Review and Analysis Program (MRAP), 135

Manchester Metropolitan University, 540, 544

Maps
- Degree Confluence Project, 654–655

U.S. health outreach projects, 439f

Market-oriented culture
- buying Web links, 527–528
- compensation systems, 102
- organizational culture, 39f, 40, 47–48
- Web advertising, 564–565

Marketing. See Publicity and public relations

Markup languages
- HTML, 583
- The Linux Documentation Project, 653
- XML, 583–584, 633, 647

Mars Hill Graduate School, 398

Maryland
- University of Maryland Libraries, 62

Mass media, 519

Massachusetts
- Boston libraries, 435–436
- Harvard College Library, 164
- Harvard University, 197f, 210

Mathematics content. See Science and mathematics materials

Matthews, Suzanne, 303

Mature vs. young libraries, 23, 24

McClure, Charles, 98

McCutcheon, D., 132

McGill University Health Centre, 429

McGoogan Library of Medicine, 413

The Measurement and Evaluation of Library Services (Lancaster), 131

Measurement and quantification
- Balanced Scorecard model, 11, 12–13
- LibQUAL+® (survey tool), 13
- library performance data, 129–155
See also Information and data

Measuring the Difference (NN/LM), 442, 443–444

Medical advice, 470, 473–474, 475

Medical libraries
- collaboration with other libraries, 411–421, 435–436, 440, 445–449t, 467–468
- College of Physicians of Philadelphia, 458–459
- MedlinePlus® links, 382
Neuro-Patient Resource Centre, Montreal, 426–431
Pacific Southwest Regional Medical Library, 529
PlaneTree Health Library, 333
See also National Library of Medicine (NLM)
Medical Library Association (MLA)
Health Information Literacy Task Force, 331–332
health literacy definition, 337
information access issues, 436
MLA News, 291
Medical textbooks, 340–346, 341t, 343t
Medications
Canadian health practice, 429
herbal remedies, 492
medical research, 490
MEDLINE (NLM), 375–376, 380–381, 438
MedlinePlus en español (NLM), 382
MedlinePlus® (NLM), 375–388
easy-to-read resources, 332
history, 438
inclusion criteria, 404
Interactive Health Tutorials, 350, 350
publishing system, 385f
training, 355–356, 508
user survey, 435
Memory Jogger II, 77
Mental health
immigrants and refugees, 311
Iowa City Public Library workshop, 507
residential care, 504
Mentors, 59, 206
Merit pay, 103, 105
MeSH (Medical Subject Headings), 285, 378–379, 380, 483
Metabolic Syndrome X, 483
Metadata
Arc (search service), 591–598, 599f, 596f, 597f
Collection Workflow Integration System, 622–624, 625f, 627, 633
Data Fountains (collection development software), 606–607
DP9 (metadata gateway service), 600–601
Etree.org, 657–658
INFOMINE (virtual library service), 606
iVia (collection development software), 605–606, 610, 614–615, 617–618, 619
Joint Information Systems Committee projects, 637–650, 649f
The Linux Documentation Project, 653
Open Archives Initiative Protocol for Metadata Harvesting, 576–589
speed of innovation, 24–25
Metasearching, 614
MetaWest, Inc., 98
Methodology
Open Archives Initiative projects, 581
organization assessment, 159–161
organizational development interviews, 6, 6t
outreach program evaluation, 443
performance assessment, 141, 156–157
process improvement, 92
QA Focus project, 641
student searching research, 544–545, 549
Michael Callen-Audre Lorde Community Health Center, 271
Michigan
Michigan Leadership Academy, 198t
University of Michigan, 201t
Microevaluation of data, 131–132
Microsoft Corporation
filtering software, 563–564
Web site, 289
Word® software, 341
Migrant workers, 310–311
Migration, library users, 273–274
Minnesota
Hmong Health Care Professionals Coalition, 316–317
University of Minnesota, 201t
Minorities. See Diversity issues
Misinformation, 373, 430, 470, 473–474
Mission, library
Iowa City Public Library, 497
leaders’ support of, 59
Neuro-Patient Resource Centre, Montreal, 425
Missouri
health information librarians, 414
University of Missouri-Columbia, 196
INDEX 685

Mitretek Systems, Inc., 370
MLA News, 291
Models and tools
Arc (OAI search service), 591–593
Balanced Scorecard, 11, 12–13, 37, 141, 160–161
BOBBY® (Web site tool), 289
Campbell Organization Survey, 27–28
Competing Values Framework, 37–43, 39f, 48
Constructive Dialog, 79
facilitative leadership, 235
Flesch-Kincaid Grade Level, 341, 342–343, 343t
Flesch Reading Ease, 341, 343–344, 343t
4-D Cycle model, 223–225
Galbraith’s Star Model, 11
Google Web toolkit, 581
Hoshin Planning, 11, 85–88
IQ-Tool, 370
Joint Information Systems Committee projects, 641, 648, 649/KEYS to Creativity, 28
left-hand column exercise, 233–234
LibQUAL+®&trade:, 13, 139, 148, 160
Making Your Web Site Senior Friendly (NLM, NIA), 289
Management by Objectives (planning system), 86
Memory Jogger II, 77
Mod_oai Project, 587
organization charts, 122, 123
Organizational Culture Assessment Instrument, 39–40, 43, 49–50
organizational development, 243–244
Peabody Individual Achievement Test, 339
Rapid Estimate of Adult Literacy in Medicine, 339
Santa Fe Convention, 591
Scout Portal Toolkit, 620–636
SERVQUAL®, 139
Social Science Information Gateway, 644
team learning model, 124f
Venn diagrams, 400
See also Standards of practice
Mod_oai Project, 587
Monastery, Wat Tham Krabok, 23
Monroe Country Library System, 198t
Montreal, Neuro-Patient Resource Centre, 425–431
Mortensen Center for International Library Programs, 199t
Mountain Plains Library Association Leadership, 199t
Multimedia materials. See Audio and video materials
Music
Etree.org, 655–659
Sheet Music Consortium, 579
NASA Technical Report Server (NTRS), 597–598
National Adult Literacy Survey, 423–424, 486
National Cancer Institute, 289
National Council on Interpreting in Health Care (NCIHC), 306, 308, 312
National Heart, Lung, and Blood Institute, 295
National Institute on Aging (NIA)
Age Pages, 290
Making Your Web Site Senior Friendly, 289
National Institute on Diabetes and Digestive and Kidney Diseases (NIDDK), 294–295
National Institutes of Health (NIH) collaboration projects, 379
history, 375
SeniorHealth.gov, 290
National Library of Medicine (NLM)
AAHSL Leadership Fellows Program, 199t
Hmong outreach program, 302, 326
Making Your Web Site Senior Friendly, 289
outreach project funding, 434, 436–439, 451
Web site, 332
See also MedlinePlus® (NLM)
National Network of Libraries of Medicine (NN/LM)
health literacy symposium, 329
HealthInfoQuest (Web site), 349, 349, 477
outreach project funding, 434, 438–439, 439f, 442–443, 451
public libraries health information, 477–478
Web design tools, 289
National Science Digital Library (NSDL), 611, 622
National Science Foundation (NSF) Internet Scout Project, 621–622
National Sciences Digital Library, 579–580
National Standards for Culturally and Linguistically Appropriate Services in Health Care, 307–308
Native Americans
health information projects, 436–437
Techniques for Evaluating American Indian Web Sites (Cubbins), 294
Nebraska
health information librarians, 413
Nebraska Library Commission, 469–470
Nebraska Library Institute, 199
University of Nebraska-Lincoln, 62–66
Neighborhood centers, 505
Neighborhood Internet Protection Act (NIPA, 2000), 556–557
Nielsen/NetRatings, 293
Net Nanny® (filtering software), 562–563, 568–569, 571
Networking
health information symposium, 334
Hmong Health Web site, 324–325
leadership training, 206–207, 212
Philly Health Info project, 463
Web links, 525
Neuro-Patient Resource Centre, Montreal, 425–431
Nevada Leadership Institute, 199
New Jersey Library Association (NJLA), 193, 199
New Measures Initiative (ARL), 139
New Mexico
Library Leadership New Mexico, 200
Los Alamos National Laboratory, 30, 586–587
New York City
health warnings, 274
library communities, 271
non-English language requirements, 278
Queens Public Library, 194
News, health information, 490
NIA. See National Institute on Aging (NIA)
NIDDK. See National Institute on Diabetes and Digestive and Kidney Diseases
NIHSeniorHealth (NLM Web site), 332
Nokia (corporation), 30
Nonprofit organizations
American Accreditation Health Care Commission, 370
Andrew W. Mellon Foundation, 578, 621–622
Bill and Melinda Gates Foundation, 284
College of Physicians of Philadelphia, 458
Health on the Net, 369–370
innovation, 19–20
user services, 21
Normalization, 585
North Carolina
North Carolina LA Leadership, 200
University of North Carolina at Chapel Hill, 387
Wake County Public Library System, 235–236
North Suburban Library System (Chicago), 177–178
Northern Exposure (leadership training), 200
Northern Wisconsin Area Health Education Center (NAHEC), 302
Northwest Career Development and Assessment Center, 208, 209–210, 213
Norton, David, 11
N2H2 (Web filtering software), 563, 568
Nurses
Neuro-Patient Resource Centre, Montreal, 426
organizational culture, 41
using health information, 417
Nutrition and diet, 483
NUValues (University of Nebraska), 65
NVivo software, 440
OAIster project (U. of Michigan), 578
Oakland (CA), non-English materials, 278
Obscenity, 555
Observation studies, 363
OD. See Organizational development
Office of Disease Prevention and Health Promotion (ODPHP), 453–454
Office of Minority Health, U.S., 307
Ohio
Cleveland Clinic, 220
Library Leadership Ohio, 198t
Ohio State University, 189–190
Online chat reference services, 57
Online games, Everquest, 527
Ontario, Canada, public libraries, 473–474
OPACs (Online Public Access Catalogue), 540, 546, 547
Open Archives Initiative (OAI), 586, 590–591
Open Language Archives Community (OLAC), 579
Open source software, 610
Oracle Corp., 386
Oral history tradition, 306
Oram, Andy, 652
Organization charts
MedlinePlus® (NLM), 383f
Teton County Library (WY), 119–122, 121f, 123
Organization Development: Strategies and Models (Beckhard), 6
Organizational culture, 33–53, 35t
culture of assessment, 137, 148, 162, 164–165, 171
dialogue, 28–29
organizational development, 242, 245, 247
prospective employees, 106
team-centered vs. hierarchical, 84
Organizational Culture Assessment Instrument (OCAI), 39–40, 43, 49–50
Organizational development, 1–4, 238–257
academic libraries, 5–16, 6t, 7t, 9t, 12t, 14–15t
appreciative inquiry, 218–229
assessment benefits, 156–171
human resources, 172–186
strategy and innovation, 27–31
University of Arizona Library, 68–111
Organizational Learning (Argyis and Schön), 56
Outreach, health information
Cultural competence, 272–274
Health librarians, 453–456
Hmong health information project (WI), 302–303, 315–316
in hospitals, 429–430
Iowa City Public Library, 497–498, 504
National Network of Libraries of Medicine projects, 434–452
Philadelphia area, 458–463
public libraries, 466, 472, 473
See also Publicity and public relations
Owens, Christine Wilson, 332
Pacific Southwest Regional Medical Library, 331, 529
PageRank (Google), 525, 526–527
Pamphlets and brochures, 276, 279, 280, 281, 404, 428
Parades, 115–116
Paraprofessionals, 44
Parent organizations
academic libraries, 8, 10, 44, 103, 104–105
government agencies, 242
medical libraries, 412, 415
organizational development context, 246–247
school and special libraries, 46–47
University of Nebraska, 65
Parents and parenting skills, 251
seeking health information, 468–469, 472
Web filtering software, 558–559
Parks, 253
Partnerships. See Collaboration
Patients
health beliefs, 310
hospital visits, 337
Neuro-Patient Resource Centre, Montreal, 425–427
patient-physician relationship, 287, 484
personal health responsibility, 464–465, 485
Patterns of use, 543
Peabody Individual Achievement Test (PIAT), 339
Peacefire (anti-Web filtering group), 566
Peer review
Teton County Library (WY), 118
University of Arizona Library, 105
PEMSysytem (U. of Arizona Library), 98–100
Penn Library Data Farm, 145–146
Pennsylvania
Philadelphia, 457–463
Tri-College Libraries (Philadelphia), 227–228
University of Pennsylvania Library, 145–146
University of Pittsburgh School of Information Sciences, 291–292
Performance evaluation
accountability, 83, 179
data, 128–155
leadership training, 208, 212, 214
pay for performance, 101
peer review, 118
recognition programs, 228
360-degree evaluation, 235, 252
University of Arizona Library, 99
Web-based instruction, 352, 357
See also Staff, hiring, and employment issues
Personnel. See Staff, hiring, and employment issues
Pfizer (company), 331
Philadelphia
Philly Health Info, 457–463
Tri-College Libraries, 227–228
Philosophy, appreciative inquiry, 222t
Phipps, Shelley, 116
Photographs
Degree Confluence Project, 654–655
picture frames, 318
Physicians
College of Physicians of Philadelphia, 458
cultural differences, 405
patient-physician relationship, 287, 426–427, 464, 484
Schillinger, Dean, 332
Physics information, 600
Picture frames, 318
Picture Stories for Adult ESL Health Literacy, 309
Placebo effect, 221
PlaneTree Health Library, 333
Planning. See Strategic planning
Platform for Internet Content Selection (PICS), 560–561
Play, in libraries, 25
Policies, library
online subject gateway, 643–644
supporting learning organizations, 59
Political aspects of Web, 555–575
Politics, organizational
health information outreach projects, 450
innovation, 20–22
See also Social aspects
Pornography, 555, 557
Portals, Internet. See Gateways, portals, and repositories
Positive thinking, 221, 223, 228, 240
Post-traumatic stress disorder, 311, 488
Posters, health information, 279
The Power of Appreciate Inquiry (Whitney and Trosten-Bloom), 221
Pregnancy information, 507
Print vs. electronic information, 58, 142–144
Private sector. See Business and private sector
Problem solving
appreciative inquiry, 219, 220t
learning organizations, 58
organizational development, 244
Process improvement systems, 90–96
Professionalism. See Librarianship
Project management, 159–160
Proof-of-age software, 559
Proprietary environments, 559–560
Proxies (metadata harvesting), 598
Psychological factors. See Behavior and psychology
Public health. See Health care
Public health care assistance, 311–312
Public libraries
Chicago suburb libraries, 163–164
collaboration with other libraries, 411–421, 435–436, 440, 445–449t
computer access study, 284
Free Library of Philadelphia, 460
health information, 376, 381, 438–439, 480–495
Hmong and health information, 316, 321–322, 322
Iowa City Public Library, 496–511
organizational culture, 45–46
performance data, 136
Queens (NY) Public Library, 194
reference services, 464–479
self-checkout systems, 21–22
team management structure, 112–128, 121f, 124f
Toronto Public Library, 194
Wake County (NC) Public Library System, 235
Public schools, 530–531
See also Classroom environment
Publications
Activity Book for African American Families: Helping Children Cope with Crisis, 295
Age Pages (NIA), 290
American Libraries (ALA), 113
The Fifth Discipline: The Art and Practice of the Learning Organization (Senge), 57–58, 70, 112, 122, 225
The Fifth Discipline Fieldbook (Senge), 70
The Grace of Great Things (Grudin), 20
Healthy People 2010 (health objectives), 330–331, 337, 425, 453–454, 456, 483
Improving Health Professionals’ Access to Information (NLM), 438
Journal of Cardiovascular Nursing, 417
Journal of MedSurg Nursing, 417
The Leadership Challenge (Kouzes and Posner), 189t
The Linux Documentation Project, 652–654
Making Your Web Site Senior Friendly (NLM, NIA), 289
The Measurement and Evaluation of Library Services (Lancaster), 131
Measuring the Difference (NN/LM), 442, 443–444
National Standards for Culturally and Linguistically Appropriate Services in Health Care, 307–308
Organization Development: Strategies and Models (Beckhard), 6
Organizational Learning (Argyis and Schön), 56
Picture Stories for Adult ESL Health Literacy, 309
The Power of Appreciative Inquiry (Whitney and Trosten-Bloom), 221
QA Focus briefings, 640–641
Research-Based Web Design & Usability Guidelines (HHS, NCI), 289
Restructuring Academic Libraries: Organizational Development in the Wake of Technological Change (Schwartz), 8
Scout Report (Internet Scout Project), 622
Social Science Information Gateway, 643–644
Sustaining Innovation (Light), 19
Techniques for Evaluating American Indian Web Sites (Cubbins), 294
The Thin Book of Appreciative Inquiry (Hammond), 221–222
Weight Information Network (NI-DDK), 295
Publicity and public relations
change management, 21–22
health information, 412–413, 416–418, 442, 473
health warnings, 274
Iowa City Public Library, 507–509
Neuro-Patient Resource Centre, Montreal, 430
Teton County Library (WY), 115–116
See also Outreach, health information
Publishing
The Linux Documentation Project, 652–654
MedlinePlus® system, 385–386, 385f, 387
Neuro-Patient Resource Centre, Montreal, 431
Web as publishing tool, 296
wikis, 657
PubMed (NLM), 380–381
Puget Sound region, 398
Purpose stories, 124
Push technology, 627–628
Pygmalion Effect, 221
QA Focus project, 640, 649f
QSR International, 440
Quality
health information, 485–486
Internet gateways, 639–648
leadership models, 76
metadata standardization, 637–650
student searching research, 544, 549–550, 550f, 551–552
Web health information, 360–374, 376–377, 384, 483–483
Web sites, 543–544
Queens (NY) Public Library, 194
Questionnaires. See Surveys and focus groups
QUICK (Quality Information Checklist), 371

Race. See Diversity issues
Radio programs, 319
Raleigh (NC) public library, 235
Rambo, Neil, 331–332
Rapid Estimate of Adult Literacy in Medicine (REALM), 339
Ratings
collection development gateway, 629
Web site ratings, 560–562, 563, 570
Reading levels
medical textbooks, 340–346, 345–346
Web design tools, 289
See also Usability
Recommendations, user, 629
Recordings, music, 655–659
Rees, A.M., 443–444
Rees, Fran, 230–231
Reference (health) books, 485
Reference interviews, 278, 469–470, 477
Reference services
health questions, 435, 464–479
Iowa City Public Library, 498
online chat, 57
University of Virginia, 147
Referral services, 313
Refugees, health information, 301–328
Registration, Open Archives Initiatives projects, 580–583, 598
Regulation of Web content, 560
Religious organizations
health information outreach, 398, 401, 402, 404
library communities, 271–272
Reports and research
compensation issues, 104
elderly and Internet use, 286–288
health literacy, 291–293, 338–345, 361–369, 477
leadership, 188–190, 189t, 193, 207–213, 214
library role in parent organizations, 247
library usage statistics, 134, 138
literacy, 423–424
medical research, 381, 490
performance standards, 140
process improvement, 91
public library computer access, 284
September 11 attacks, 489
student Web use, 531–535, 534f, 541–550, 550t
University of Arizona Library, 89
Web filtering software, 566
Web searching, 540–541
See also Surveys and focus groups
Repositories. See Gateways, portals, and repositories
Research-Based Web Design & Usability Guidelines (HHS, NCI), 289
Researchers
Argyris, Chris, 233–234
Campbell, David P., 27–28
Cooperrider, David, 220
Deming, W. Edwards, 69
Gerould, J.T., 130
Kaplan, Robert, 11
McClure, Charles, 132
McCutcheon, D., 132
Norton, David, 11
Rees, A.M., 443–444
Schein, E.H., 34, 35–36
Scholtes, Peter, 69
Trostten-Bloom, A., 221
Whitney, D.K., 221
Residential care facilities, 504
Resnick, Paul, 560
Resource Discovery Network (RDN), 638
Restructuring Academic Libraries: Organizational Development in the Wake of Technological Change (Schwartz), 8
Retirement communities, 504–505
Retirement planning, 506
Reviews, non-English materials, 279
Riffenburgh, Audrey, 332
Rights and permissions
Collection Workflow Integration
System, 624–625
Open Archives Initiative projects, 587
See also Copyright
Risk, in innovation, 25–26
RSS (content format), 633
Ryerson University Library, 166

SafeSurf (Web safety group), 560–561, 567, 570, 571
Safety, Internet, 556, 557, 558–563, 564–570
Salaries. See Compensation, employee
Sample, Steven, 249
San Jose (CA) Public Library, 194
Santa Fe Workshop (1999), 591
Scalability, 597
Schein, E.H., 34, 35–36
Schillinger, Dean, 332
Scholtes, Peter, 69
Schön, Donald, 56
School libraries, 46–47
Schwartz, Charles, 8
Schwarz, Roger, 230, 231, 233, 234f, 244
Science and mathematics materials
aerospace, 597–598
National Sciences Digital Library, 579–580
physics, 600
Scout Portal Toolkit, 622
Scout Archives, 622
Scout Portal Toolkit, 620–636, 630f
Scout Report, 622
Search and browse strategies
Collection Workflow Integration
System, 624, 626–628, 628f
health information, 363, 400
iVia collection development software, 612–613, 614
Web searching, 538, 540–547
See also Information seeking behavior
Search and Retrieval Web/URL Service (SRW/U) protocol, 582, 587
Search engines
health information seeking, 363, 365
industry mergers, 536
metadata harvesting, 600–601
revenue strategies, 519–520, 521, 569
search strategy research, 540–547, 550–551, 552
Web links, 524–527
Search logs
health Web sites, 378
MedlinePlus® (NLM), 386
Search services
Arc (OAI search service), 591–599, 602
Digital Library Grid, 601–602
health information companies, 416–417
Open Archives Initiative projects, 582
Seattle, Center for Children with Special Needs, 466–467
Security. See Rights and permissions; Safety, Internet
Selection. See Collection development
Self-checkout systems, 21–22
Self-managed work teams, 178–179
Self-regulation, Web content, 560
Senge, Peter, 56–57, 69, 70, 112–113, 114, 122–123, 124, 225
Senior Fellows Program (UCLA), 196, 201
SeniorHealth.gov (NIH), 290
September 11th terrorist attacks, 487–489
SERVQUAL®, 139
Sexual orientation, library users, 271
Shared leadership model, 73, 75–80
Sheet Music Consortium, 579
Shewhart, Walter, 91
SHRM. See Society for Human Resource Management
Skill sets, library staff
compensation issues, 102, 104
job redefinition, 167
leadership, 188, 192–193, 194, 206, 214
managers, 49, 250
occupational culture, 37
personal mastery, 49
team-building, 185
underutilized staff, 239
University of Arizona Library, 79, 84–85
University of Nebraska-Lincoln Libraries, 63–64
University of Virginia Library, 165, 166
SLA. See Special Library Association
Slang, 285
SMOG (Simplified Measure of Gobbledygook), 332, 346
Snowbird Institute, 200t, 208, 210–211, 212
Social aspects
  baby boom generation, 490–491
  health literacy, 480, 481, 483, 487–489
  innovation, 20–21
  role of libraries, 34
Social Science Information Gateway (SOSIG), 642–648, 642f
Society for Human Resource Management (SHRM), 173
Software
  Arc (search service), 591–593, 592t, 599–601
  BitTorrent (file-sharing system), 658–659
  blocking and filtering software, 558–564, 565–570
  DP9 (metadata gateway service), 600–601
  LibQUAL+R (survey tool), 13, 139, 148, 160
  Microsoft Word®), 341
  NVivo (data analysis), 440
  online collection development platform, 604–619, 605f, 620–636, 625f
  PEMSystem (U. of Arizona Library), 98–100
Spanish materials
  health information, 275–276, 320
  MedlinePlus en español (NLM), 382
Special libraries, 46–47
Special Library Association (SLA), 193–194
Specialists. See Consultants and specialists
Speed of Internet access, 549
Sports
  Hmong community, 316–317
  positive thinking, 221
  Squigly’s Playhouse (Web site), 571
Staff, hiring, and employment issues
  compensation system, 100–105
  human resources, 172–186
  job redefinition, 167–170
  learning organizations, 54–66
  MedlinePlus® (NLM), 383–385, 383f
  organizational assessment, 157, 162
  organizational culture, 41
  organizational development, 9t, 14t, 242–243
  Philly Health Info volunteers, 462–463
  psychological contracts, 248
  Teton County Library (WY), 116–118, 125
  University of Arizona Library, 79, 84, 96–100, 106–107
  University of Virginia Library, 165–166
  Wake County (NC) Public Library System, 236
  “Stake Your Claim to Health Literacy” (symposium), 328–329
Standards of practice
  concert recordings, 656
  cultural competence, 407–408
  Dublin Core, 585
  health information, 307, 476–477
  innovation, 24
  internal library studies, 140
  leadership competency, 193–194
  library data, 130
  metadata, 610, 637–650
  online usage data, 143
  Open Archives Initiative Protocol for Metadata Harvesting, 522,
  576–589, 590–603
  performance reporting, 141
  RSS (content format), 633
  Search and Retrieval Web/URL Service protocol, 582, 587
  Web ratings systems, 560–562
  Web site evaluation criteria, 368–369, 371–372, 503
  See also Models and tools
Stanford-California Institute, 200t, 208, 209, 210, 211–212
Start-up libraries. See Young vs. mature libraries
State health departments, 314
State Library of Iowa, 500
Static Repository Gateway (Los Alamos National Laboratory), 586–587
Statistics
African Americans and Internet use, 293
community languages, 260
elderly and Internet use, 287
health information seeking, 435
immigrants, 304
Internet access, 481, 556
Iowa City Public Library users, 498
librarians, 191–192
library community research, 272
library performance, 130–131, 133–134, 143, 146
literacy, 424
process improvement data, 95
Web domains, 533–535
Stoffle, Carla, 113, 136, 145
Strategic planning
data usage, 135, 136, 149
health information outreach projects, 444
health information symposium, 333–334
Hmong health information project, 303
human resources, 174
innovation, 26–31
Iowa City Public Library, 496–497
learning organizations, 60–62
organizational culture, 33–53
organizational development, 10–11
strategic plans, written, 27
University of Arizona Library, 85–88
Students and learners
academic Web use, 530–538, 534f, 539–554
anti-drug abuse workshop, 506
Hmong health information projects, 315
Iowa City elementary, 509
medical textbooks study, 342
Pygmalion Effect, 221
University of Virginia, 147
Web in classroom, 564, 572–573
Studies. See Reports and research
Style sheets, 583
Subject headings. See Terms and subject headings
Substance abuse, 506
Supreme Court, 567
Surveys and focus groups
community assessment, 400–401
elderly and Internet use, 286–287
health information outreach projects, 443, 461
leadership training, 192
learning organizations, 65–66
online gateway quality assurance, 640
organizational culture, 42
organizational development, 13, 27–28, 167
student Web use, 531–535, 542
user needs, 139, 147, 148, 160–161
Web filtering software, 568
Web health information training, 503
See also Assessment; Reports and research
Sustaining Innovation (Light), 19
Swarthmore College, 227–228
Sweden
library usage, 432
Symposia. See Conferences and workshops
Systems design, 68–111
appreciate inquiry, 219
Arc (OAI search service), 592–597, 594f, 598–599
collection development software, 608–609, 609–611, 615
Collection Workflow Integration System, 634
Digital Library Grid, 601–602
Etree.org, 658
evaluation of data, 132
Joint Information Systems Committee projects, 641
MedlinePlus® publishing system, 385–386
TALL Texans, Texas Library Association, 201t, 210
TDLP (The Linux Documentation Project), 652–654, 659, 660t
Teachers
anti-drug abuse workshop, 506
Pygmalion Effect, 221
Web-based instruction, 351–352, 353
See also Faculty, college
Teams, library
  group process improvement, 239, 242
  in public libraries, 112–128, 121f, 124f
  team-based organizations, 72, 73, 80–85
  team-building, 178–181, 180t, 185
Technical support activities, 165
Techniques for Evaluating American Indian Web Sites (Cubbins), 294
Technological innovation
  artifacts of librarianship, 35
  Etree.org, 657–658
  online repositories, 660t
  skills of library staff, 165
Teton County Library (WY), 35
University of Arizona Library, 108–109
users’ expectations, 158
Teenagers, Iowa City Public Library workshop, 507
Telecommunications industry, 30
Television programs
  Hmong and health information, 319
  Iowa City outreach projects, 505–506, 507, 509
Terms and subject headings
  Collection Workflow Integration System, 624, 626, 627
  health Web sites, 288, 350, 386
  hierarchical subject gateways, 551
  iVia (collection development software), 612
  leadership, 191
  MeSH (Medical Subject Headings), 285, 378–379, 380, 483
  Open Archives Initiative projects, 581, 587
  Social Science Information Gateway, 642f
  Web filtering software, 562–563, 566, 567
Terrorism, 487–489
Testing, Web-based instruction, 352, 357
Teton County Library (WY), 114–127, 121f, 124f
Texas
  TALL Texans, Texas Library Association, 201t, 210
  Texas A&M University, 139–140
  University of Texas School of Library and Information Sciences, 292
Textbooks, 340–346, 341t, 343t
Thailand, 305
The Thin Book of Appreciative Inquiry
  (Hammond), 221–222
360-degree evaluation, 235, 252
Toronto Public Library, 194
Toronto Reference Library, 474–475
Total Quality Management, 85
Traditional medicine, 309, 310
Training and education
  elderly Internet users, 287
  health care providers, 319–320
  online subject gateway editors, 643–644
  Philly Health Info volunteers, 462
  public library health information, 468, 485, 498, 503
  Web health information, 349, 372–373, 508–511
Training and education, library staff
  change management, 50, 182
  compensation systems, 103
  health information outreach, 441
  human resources management, 184
  innovation, 25
  Iowa City Public Library, 503
  leadership, 187–217, 197–202t, 205t, 250, 251–252
  learning organizations, 54–67
  learning plans, 227–228
  organizational development, 14t
  performance management, 98
  specialized skills, 37
  Teton County Library (WY), 117
  University of Arizona Library, 77, 78–79, 82
Training Institute for Librarians of Color, University of Minnesota, 201t
Transactional leadership, 190
Transcopyright, 526
Transformational leadership, 190
Translations
  health care, 312
  Hmong and health information, 306–307, 323
Treatment, medical
  currency of health information, 430–431
  informed patients, 427
  See also Medications
Tri-College Libraries (Philadelphia), 227–228
Tribal Connections project, 436–437
Trosten-Bloom, A., 221
Turning Point initiative, 437

UKOLN (UK Office for Library Networking), 639
ULC Executive Leadership Institute, 201
Undocumented immigrants, 304
Unexplained symptoms, 487–488
Uniform Resource Locators. See URLs (Uniform Resource Locators)
United Kingdom
  Centre for Health Information Quality, 371
  expert patient program, 485
  Information Environment, 539–540
  Internet initiatives, 552n
  Joint Information Systems Committee projects, 637–638, 639–640, 642
  student Web searching research, 543
  UK Office for Library Networking, 639
Universal Preprint Service (UPS), 591
Universities and colleges
  Hmong health information projects, 315
  organizational culture, 42
  student searching strategies, 543, 544
  See also Academic libraries
University of Alberta, 200
University of Arizona Library
  mentor role, 113, 116, 119
  organizational development, 10, 13, 19, 30, 68–111
  performance assessment, 136, 145
University of Bath, 639
University of Bristol, 639, 642
University of California, Los Angeles
  health information survey, 292
  Senior Fellows Program, 196, 201
University of California, Riverside, Library, 604, 606
University of California, San Francisco, 332
University of Connecticut
  Health Center, 477
  Libraries, 12–13
University of Illinois
  at Chicago, 437
  Digital Gateway to Cultural Heritage Materials, 578
  Experimental OAI Registry, 580–581
  Graduate School of Library and Information Science, 352
  Library Research Center, 489
  University of Maryland Libraries, 62
University of Michigan
  OAIster project, 578
Public Library Leadership, 201
University of Minnesota Training Institute for Librarians of Color, 201
University of Missouri-Columbia, 196
University of Nebraska-Lincoln Libraries, 62–66
University of North Carolina at Chapel Hill, 387
University of Pennsylvania Library, 145–146
University of Pittsburgh School of Information Sciences, 291–292
University of Southern California, 249
University of Texas School of Library and Information Sciences, 292
University of Utah Health Sciences Center, 289
University of Virginia Library
  assessment projects, 164–166
  organizational development, 12
  performance data, 146–147
University of Washington
  Libraries, 147–149
  Tribal Connections project, 436–437
University of Wisconsin, Madison, 622
URAC (American Accreditation Healthcare Commission), 370
URLs (Uniform Resource Locators), 532, 583–585, 618

U.S. Health and Human Services. See Department of Health and Human Services

Usability
African Americans and Internet use, 294
elderly and Internet use, 288, 289, 290
health information, 266, 289
internal library studies, 140
Web-based instruction, 358
Web health information, 369
See also Reading levels

User interfaces, 631–633, 632f

User services and needs
organizational assessment, 171
organizational development, 9t, 14t, 46, 248–249
process improvement, 92, 93, 95
Social Science Information Gateway, 643
surveys, 139, 147, 148
technology, 158
user-centered libraries, 17–32, 71–74, 136–137, 242

Users
evaluating health information, 361
Iowa City Public Library, 498
MedlinePlus® feedback, 386
migration, 273–274
perspective of health reference, 466–472, 474, 478
resource gateway ratings, 629
search strategies, 541
user agents, 627–628
varying constituencies, 43

Utah
health information librarians, 413–414
University of Utah, 289
Utah Consumer Health Information Network (UCHIN), 413–414
Utah Library Association, 414

Validation, metadata, 647

Values
organizational, 35–36, 35t, 37–38, 42, 227, 247
personal, 206, 231–232

Vendors, journals, 142–143, 146
Venn diagrams, 400
Version control, 653
VHS. See Audio and video materials
Vietnam War, 304

Virginia
University of Virginia Library, 12, 146–147, 164–166
Virginia Library Association, 202f
Virtual communities, 656
Virtual libraries, 604–605, 606, 612
See also Digital libraries

Vision, organizational
appreciative inquiry, 224–225, 227
Iowa City Public Library, 497
leadership, 190
organizational development, 15, 57–58, 61, 242
Teton County Library (WY), 124–125, 126–127
University of Arizona Library, 72, 86, 87
University of Nebraska-Lincoln, 63

Vision Literacy (organization), 333

Vocabularies. See Terms and subject headings

Volunteers
Degree Confluence Project, 654–655
Etree.org, 657–659
health information, 313
The Linux Documentation Project, 652–654
Philly Health Info project, 462–463
WiredSafety (Web safety group), 562

Wake County (NC) Public Library System, 235

War
after September 11 attacks, 488
Vietnam War, 304

Washington
Seattle hospitals, 466–467
University of Washington, 147–149, 436–437
Washington Heights, NYC, 274
Washington State Fathers Network, 472
Web and Internet

Web sites

Weight Information Network (NIDDK), 295

Wescott, Beth, 332
Western Electric Company, 91
Whitney, D.K., 221
Wikis, 657
WiredSafety (Web safety group), 561–562

Wisconsin

Women
health information seeking, 291–292 health issues, 265 Iowa City Public Library workshop for girls, 507
library health information programs, 277
Word® (Microsoft) software, 341
Work styles and habits
process improvement, 95
University of Arizona Library, 83
Workflow management, 624–625
Workshops. See Conferences and workshops
World library issues
Degree Confluence Project, 654–655
library usage customs, 273, 432
The Linux Documentation Project, 654
non-English materials distributors, 275
usage of digital materials, 143–144
Writing styles, 341–342, 344
Wyoming
health information librarians, 413
Teton County Library, 114–127
Wyoming Library Association Leadership, 202t
XML (Extensible Markup Language)
Collection Workflow Integration System, 633
ERRoLs (Extensible Repository Resource Locators), 583–584
Social Science Information Gateway, 647
Yahoo! Inc., 545, 560
Young vs. mature libraries, 23, 24
Younger generation librarians, 109
YS Lead MA (leadership program), 202t
Zip codes, 271