Newsletter

ALA TechSource
www.techsource.ala.org

Smart Libraries™

Formerly Library Systems Newsletter™

50 East Huron Street, Chicago, Illinois 60611-2795, USA

June 2004 Volume XXIV Number 6

ISSN 1541-8820

Amazon raises the bar, search engines lower it

As the search engine market has been slowly consolidating over the last year, Amazon.com's research arm, a9, has entered the fray with a new search engine that combines Google searches with Amazon's "Search Inside the Book" index.

The a9 service, still in Beta release as of early May 2004, also includes a downloadable toolbar and a search history tool. As Amazon and free Web search engines keep raising the bar on functionality, they keep lowering it on privacy and relevant search results. As libraries evaluate the service offerings of Internet entities, they should constantly be aware of the business those entities are in.

Amazon's toolbar will include Amazon searches, as well as searches of the Internet Movie Database (owned by Amazon), Mirriam Webster's online dictionary and thesaurus, and a cached version of Google. Since the history function stores search histories for registered Amazon users online instead of in a cookie, concerns over privacy have already been raised.

In fact, Amazon has released an anonymized version of the search

See **Amazon** *on page 2*

Two new uses for OAI

The Open Archives Initiative (OAI) has quickly evolved to become part of the basic infrastructure for digital libraries. Since its inception at the Sante Fe Convention in October 1999, the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) has steadily gained acceptance as the preferred approach for metadata exchange in the digital library community.

Two new uses for OAI-PMH have been proposed, both with involvement of one of the initiative's key architects, Herbert Van de Sompel. One new extension of the protocol involves its use in harvesting digital objects as well as the metadata that describes the objects. The other proposes the protocol as a tool for harvesting Web sites more efficiently than is possible with current Web crawlers.

New work done at the Los Alamos National Laboratory (LANL) by a group led by Van de Sompel proposes using the protocol to harvest digital objects themselves, not just the metadata that describes the objects. Though the original workflow of OAI centers on transferring metadata from diverse types of information repositories, a need also exists to systematically collect digital objects.

At the Spring Taskforce meeting of the Coalition for Networked Information in April 2004, Van de Sompel described the work the Prototyping Team at the LANL research library

See **OAI** *on page 4*

IN THIS ISSUE

Receive *Smart Libraries* via e-mail

Subscribers who would like an e-mailed version of the newsletter each month should forward their e-mail address and ALA identifier (the 7-digit number printed on the top line of the address label that appears on page 8 of your newsletter) to jfoley@ala.org. Type "e-mail my Smart Libraries" into the subject line. Issues will be e-mailed in addition to your print subscription and at no additional charge.

Amazon *from page 1*

engine on the Web. Others, however, have pointed out the collaborative opportunities offered by the feature in that professional searchers could share their search strategies and annotated queries simply by sharing a password with clients.

But don't forget that Amazon is in the business of building a customer database to which it can hawk its wares. Just as Google and its brethren—AskJeeves, Yahoo!, and MSN—are, for all intents, advertising portals.

Search results = bait

To continue to think that free search engines are altruistically generating relevant search results for the collective benefit of the Web community is ludicrous. Their business is the conversion of search traffic into advertising revenue. This conversion is improved if a search engine is more popular, but good search results are the bait, not the big catch (see sidebar).

A9 opened its doors in October 2003; it is a separately branded yet wholly owned subsidiary of Amazon.com. Nine letters are in the word *algorithm*, which is how the group got its name.

A9's charge is to improve the search experience for e-commerce applications. On the surface, a9 does just that. Adjustable column displays, search history, and customized displays are just a few features.

Signing in—a unique feature—allows users to save search histories and then manage the display of those histories. The Google-like toolbar includes a Diary feature that allows users to create diary entries about specific Web pages and then access those entries from any computer.

Libraries can model from business

Libraries could spend a lot of time (and they surely will) debating the privacy implications of Amazon's centralized service model. And they will be interested to see how Amazon customizes its marketing based on the searches entered by its users. Amazon will indeed use that data.

How search engines do biz*

All search engine hit lists redirect users to advertisers' websites. What most librarians and users fail to realize is that this redirection is the main goal of all search engines because that's how they make money. Some do it more subtly than others, but even Google and Amazon's a9 include sponsored links.

As search engine portals compete for the market share, these ads will become more and more prominent, edging out relevant search results in return for increased revenue for stockholders. How does the redirection work?:

- Keyword leasing. Advertisers pay for the keywords that users enter into the search box. When those words are searched, the advertised link appears—the sponsored link. If more than one advertiser wants the same keywords, then keyword bidding can occur to determine the order that sponsored links appear. Search Google for "laser printer" and you will see that several manufacturers and resellers have bid for prominent ownership of those words. Follow the link, and the company pays Google again.
- Paid inclusion. Paid inclusion is when advertisers pay a fee to ensure their links appear in search results. Inclusion does not ensure *where* the link will occur, so it is not popular and generates less solid revenue than keyword leasing or its cousin, paid placement.
- Paid placement. Placement is like inclusion but adds assurance that the advertised link will show up in a given spot on the hit list, usually near the top. Google combines placement and keyword bidding to place certain advertisers above search results, rather than just on the side of the page.
- Syndicated ads. Search engines also can syndicate their ad links to other websites and search engines. You might have seen "Ads by Google" in various places. With little overhead, the advertising arms of search engine companies can easily extend their reach. For example, *Library Journal* includes sponsored links on its website, the content for which changes based on searches of its site.—*AKP*

*From Andrew K. Pace's more detailed treatment of the business of search engines, *American Libraries*, "Technically Speaking," May 2004.

But rather than predicting the Cassandra-like ramifications of user data exploitation, libraries should think about how to build similar helpful features into their own interfaces. Detailed, persistent, and "available anywhere" search histories have more relevance to scholarly research than they do to point-of-need Google searches.

Alert services based on those searches have been around for a while (like those services offered by Ingenta), but almost all of them require registration with the hosted service. Given the broad range of resources to which libraries subscribe, brokering the storage of those searches, as Amazon is doing with a9, would provide an excellent service to users.

Moreover, the collaborative aspects of annotating Web pages and queries themselves would be an excellent area of research. Extending this type of service to deeply linked resources (those nested deep within a database, such as catalog records or full-text articles) would offer unique opportunities for the academic community.

Exciting opportunities exist for library technologists. Virtual collaboration will be an increasingly important area, just as use of the library for physical collaborative space is taking off.

Libraries should put effort into creating new services equal to the effort they put into complaining when an Internet service provider runs afoul of library ethics. Applying Internet technology within the library service model will distinguish libraries from their corporate counterparts.—*Andrew K. Pace*

Contact: See an anonymized version of the a9 search engine at http://generic.a9.com.

EJOS further binds ENDEAVOR and ELSEVIER

Endeavor Information Systems Inc. has created a new version of its Encompass family of products designed for the local management of electronic journals. The product, dubbed Encompass for Journals Onsite (EJOS), provides the infrastructure for a library or consortium to load and store electronic journals on local servers, providing their users with an interface to search, browse, and view the content. The new product supports content from multiple publishers.

Although the majority of libraries subscribe to electronic journal content hosted by publishers or aggregators, some large institutions prefer to manage this content locally. This approach, though requiring significant local resources and effort, provides improved security and performance and the ability to customize a search and presentation interface.

This approach also allows the institution to create a permanent archive of journal content. For large organizations, local management of information resources may be more cost-effective than the conventional remotely hosted alternative.

As a mark of the increasing convergence of the business activities between Endeavor and its parent company Elsevier Science, EJOS has been positioned as the technology to replace Elsevier's aging ScienceServer platform, part of a wholly owned subsidiary of Elsevier Science.

EJOS will rely on technologies from both the existing Encompass and ScienceServer products. The data ingestion and storage capabilities of ScienceServer will be blended with Encompass' Oracle framework for managing metadata records, its Fast Data search engine, and a newly designed user interface.

The first organization to implement EJOS will be Institut de l'Information Scientifique et Technique based in Vandoeuvre-lés-Nancy, France, migrating from ScienceServer.

The development of EJOS reflects Endeavor's continued integration into the business activities of its parent company, Elsevier Science, and its interest in electronic publishing and digital library technologies—areas of keen interest to academic libraries, the company's key customer base.—*MB*

OAI *from page 1*

that demonstrates the ability for OAI-PMH to harvest digital objects in addition to the metadata. Its project substitutes an XML structure capable of handling complex digital objects rather than the default metadata format of Dublin Core.

Possible XML structures that might work with OAI in this way include METS and MPEG-21. Both provide an XML structure for encapsulating a digital object and its descriptive and administrative metadata into a package for transfer among interoperable digital repositories.

How OAI-PMH works

OAI-PMH provides a data model and transfer protocol for the efficient transfer of metadata among an increasing number of digital library systems or repositories.

The original OAI architecture involves a set of data providers, each of which maintains a repository that makes available the metadata describing the objects it contains. Service providers operate metadata harvesters that visit one or more data providers, creating an aggregated database of metadata that can then serve as a comprehensive access point to the multiple repositories. The federated search performed on the comprehensive database of the service provider links back to the digital objects located on the data provider's repositories.

OAI-PMH has flexible requirements for the exchanged metadata. By default, the metadata exchanged by the protocol is formatted in unqualified Dublin Core, but any other schema that is mutually agreeable within a community of repositories that participates in an OAI harvesting relationship also can be employed.

The protocol harvests selectively, based on timestamp information for the records requested. An initial harvest might request all the records in a set. Subsequent harvests would need only request records created or changed since the last harvest.

OAI-PMH was conceived to transfer metadata for the purpose of providing enhanced access to digital objects among a set of distributed digital repositories, but the protocol has proven to be extensible to applications beyond the originally intended applications.—*MB*

The prototyping done at LANL uses MPEG-21 to harvest digital objects from a DSpace repository. DSpace is software jointly developed by MIT and Hewlett Packard that has become popular among universities and research centers that want to create a repository of scholarly works produced in their institution.

LANL's challenge is to create a mapping of the metadata of D-Space into the MPEG-21 Digital Item Declaration Language. Van de Sompel says the project proves the protocol can serve as a mechanism for automatically transferring digital objects and their metadata among digital repositories.

Computer scientists at LANL and Old Dominion University have been working together to advance OAI-PMH as a method for harvesting Web sites. The original intent of the protocol involved harvesting data from repositories that store their metadata in databases that are not generally available to the Web crawlers. This project, however, aims to show that OAI-PMH can be implemented as a plug-in to conventional Web servers, providing a more efficient mechanism for indexing the Web.

Most Web crawlers follow a brute force approach for gathering Web pages. They request a server to deliver all its Web pages on each pass. Well-behaved crawlers make these requests gently, by spreading the individual page requests over a period of time, so the server is not overwhelmed.

OAI-PMH includes some characteristics that may prove to be helpful to the process of crawling Web servers, such as the ability to retrieve selected pages—those that have been created or modified since the last crawl and to note deleted pages.

The project will create an add-in module for the popular Apache Web server software that will enable a website to deliver its content through OAI-PMH. The software created by the project will be made available as open source under the GNU Public License. The Andrew W. Mellon foundation is funding the project.—*Marshall Breeding*

Contact: www.modoai.org. For a full treatment of mapping of the metadata of D-Space into the MPEG-21 Digital Item Declaration Language, see Bekaert, Jeroen. "Using MPEG-21 DIDL To Represent Complex Digital Objects in the Los Alamos National Laboratory Digital Library," *D-Lib Magazine,* Nov. 2003. www.dlib.org/dlib/november03/bekaert/11bekaert.html.

Open-access journals register impact

In the David and Goliath struggle of subscription journals versus those with open-access (OA), scholars are still seeing only a small number of titles challenge high-cost subscription titles.

Nevertheless, in April 2004, Thomson ISI reported that the relatively small number of OA titles is having a big impact on scholarship. Of the 8,700 journals selected for inclusion in ISI's Web of Science, 191 are OA journals. Of the 2,000 titles Thomson reviews each year, only about 10% are accepted for inclusion. Comparing OA titles with other scholarly titles, ISI's citation metrics—such as impact factor and cited half life—have shown no discernible difference in terms of citation impact or frequency with which the journal is cited.—*AKP*

Contact: www.isinet.com/oaj

ADOPTION OF HIGH-SPEED INTERNET ACCESS EXPANDING

According to a study released in April 2004 by the Pew Internet and American Life Project, the number of Americans accessing the Internet through high-speed broadband connections has increased significantly in 2003.

Some of the key statistics of the report, which reflect the results of a February 2004 survey, include:

- 48 million adult Americans have high-speed broadband connections in the home. This number represents 39% of all Internet users or 24% of all adult Americans.
- 68 million adult Americans use high-speed broadband connections either at work or at home, or 55% of all Internet users, or 34% of all adult Americans. This group has increased 60% in the last year.

Some of the demographic details of the study show specific groups especially favor high-speed Internet options, such as the college-educated, younger-than-35 set and those with annual incomes greater than $75,000.

Regarding the technologies used, the survey indicated that DSL adoption grew at a faster rate than cable modem. Also note that high-speed Internet access is not universal. In rural areas, only 10% subscribe to a high-speed Internet service.

Although the report shows a steady increase in the number of Americans with high-speed Internet access, it also indicates that significant percentages of the people have with low-bandwidth or no regular access the Internet.

Dial-up connections continue to be available in large numbers and offer Internet access and much lower cost than the high-speed options, such as cable-based or DSL services. Although dial-up access works adequately for Web services based on text and minimal graphics, it is less suitable for sites with high-resolution images and photographs or streaming audio or video.

As more libraries offer services that include rich media such as images, sound, and video, they should consider the number of individuals and households and that have the types of Internet connections that can best use these services.—*MB*

Contact: www.pewinternet.org/reports/pdfs/PIP_Broadband04.DataMemo.pdf

ARTstor launches digital image database

ARTstor, a nonprofit initiative funded by the Mellon Foundation, will make available its huge digital library to nonprofit educational and cultural institutions in the United States starting this summer. It was established to use digital technologies to enhance scholarship, teaching, and learning in the arts and associated fields.

ARTstor's charter collection will contain 300,000 digital images from various cultures and academic disciplines, but with a focus on the arts. The collection will include images from The MoMA Architecture and Design Collection, the Huntington Archive of Asian Art, and the Image Gallery, a collection of 200,000 images of world art and culture focused on teaching, among several others.

Two institutional subscription fees will support the charter collection database. The Archive Capital Fee (ACF) will help to ensure that ARTstor will always be in a position to adapt ARTstor's content and tools in the face of emerging technologies. The Annual Access Fee (AAF) helps cover annual maintenance costs. ACF fees range from $600 to $40,000; and AAF fees range from $500 to $20,000, depending on the size and type of the institution.—*AKP*

Contact: www.artstor.org

UNPRECEDENTED HIGHS in SECURITY CONCERNS

Potential vulnerabilities to network security as well as constant attacks are forcing all organizations that rely on technology to channel more resources into this high-stakes battle.

During the week of April 20, 2004, a huge vulnerability in one of the core protocols that underlie the Internet was discovered that forced Internet service providers and router companies to scramble to create and implement fixes to the problem before the flaw was made public.

The protocol involved is the Transmission Control Protocol (the TCP in TCP/IP). Although the problem was discovered and corrective measures were taken before the vulnerability was exploited, potential occurred for a major interruption in the flow of data on the Internet.

Network administrators continue to fight the onslaught of viruses and worms that come in waves where each successive attack proves more obnoxious than its predecessor. Two families of worms, Netsky and Bagle, have waged an all-out online battle of egos, where each competes to create new variants of increasing sophistication.

Keeping operating systems and applications up-to-date with all the patches available is more imperative than any other period in the history of computing. In April 2004 alone, Microsoft issued four security patches, three labeled as critical and one as important.

Applying security patches isn't an optional activity—the likelihood of exploitation is high for unpatched systems. The problems aren't limited to Microsoft systems. Potential flaws in Linux and Unix continue to be discovered.

Combating worms, viruses, and spam continues to absorb an unprecedented amount of network and system administrator's time. Fewer and fewer libraries with limited technical expertise can maintain their own systems without the help of specialized staff from the IT department of their larger organization.

Fewer libraries can operate such systems independently. More libraries are participating in automation systems shared through a consortium or through a vender-hosted ASP offering.—*MB*

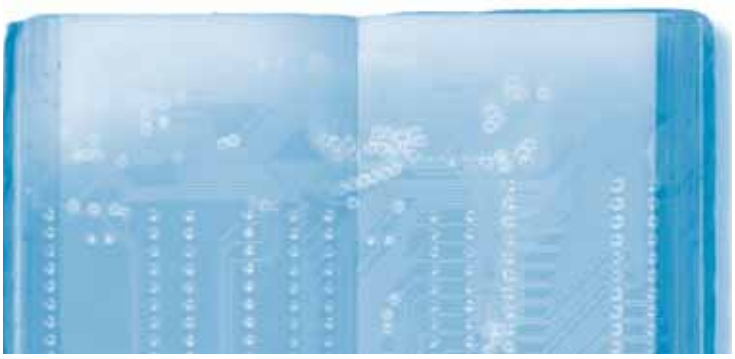
Streaming video at more libraries

As the number of potential library users increases that have high-bandwidth Internet connections, an increasing number of library services stand ready to take advantage. One bandwidth-hungry library application involves linking to streaming video of movie previews and trailers. Several Web-based OPACs now offer this feature:

- In April 2004, Innovative Interfaces, Inc., offered this capability as a new feature in the Millennium WebBridge linking product. Through a joint development effort between Innovative and Video Detective, LLC, libraries using WebBridge can provide access to movie trailers through streaming video for the VHS videotapes and DVDs in their collections.
- In January 2004, VTLS partnered with Video Pipeline, the parent company of Video Detective, offering similar capabilities through the Virtua Vectors portal.
- The Library Corp. formed an agreement with Video Pipeline in May 2003 to provide previews of videotapes and DVDs for both Library.Solution and Carl.Solution library automation systems.

Each of these products takes advantage of the more than 15,000 movie trailers maintained by Video Detective, which stores and streams the video. Libraries can offer this service without creating the large-scale technical infrastructure associated with streaming video applications.

Sirsi Corp. has offered a streaming video component as part of the enriched content option (optional subscription) in its iBistro and iLink interfaces since January 2002. Through a partnership with Bookstream, Inc., library users can view streaming video clips of best-selling authors interviews. Bookstream provides the video to Syndetic Solutions, who in turn serves as a data supplier for Sirsi's content enrichment service.—*MB*



E-BOOK BEST-SELLER LIST LAUNCHED

In April 2004, *Publishers Weekly* reported that e-book sales were up a dramatic 27% in 2003, resulting in sales of $7.3 million. Although the increase is substantial, compared with print sales, the e-book market is still trying to launch. The book publishing industry net sales reached $23.4 billion for 2003. Promising news, however, is that more than 1 million e-book titles have already been copyrighted in 2004.

The growth of e-book titles also has led to a best-seller list endorsed by the Open eBook Forum (OeBF). Publishers and resellers can contribute data to the list through the OeBF website.

The best-seller list announcement also came on the heels of a successful eBooks in the Public Library conference, in New York in March 2004. OeBF president Steve Potash noted at the conference that the e-book market is "set to explode."

Whether such an explosion will be felt by the print publishing industry is yet to be seen. What's more clear is that libraries—as early adopters of devices, collections, and services—will continue to be an important seed market for the e-book industry.

Admittedly, unit sales and revenues for e-books have a long way to go to catch up with a book industry that still brings in more money than Hollywood. Nevertheless, e-books have a market, and the ranks of loyal customers, including libraries, continues to grow. But e-books will take decades to catch up to an already mature print book market.—*AKP*

Contact: www.openebook.org/bestseller.htm

ALA TechSource
www.techsource.ala.org

Smart Libraries Newsletter
American Library Association
50 East Huron Street
Chicago, IL 60611-2795 USA

NON PROFIT
US POSTAGE
PAID
PERMIT 1479
ROCHESTER, NY

June 2004
Model from search engines

Smart Libraries Newsletter

Smart Libraries Newsletter delivers hard data and innovative insights about the world of library technology, every month.

Contributing Editors
Marshall Breeding
615-343-6094
marshall@breeding.com

Priscilla L. Caplan
352-392-9020, ext. 324
pcaplan@ufl.edu

Judy Luther
610-645-7546
judy.luther@informedstrategies.com

Andrew K. Pace
919-515-3087
apace@unity.ncsu.edu

Editor
Chris Santilli
630-495-9863
chris@wordcrafting.com

Administrative Assistant
Judy Foley
800-545-2433, ext. 4272
312-280-4272
jfoley@ala.org

Production and design by Angela Hanshaw, American Library Association Production Services.
Smart Libraries Newsletter is published monthly by ALA TechSource, a unit of the publishing division of the American Library Association.