Newsletter

ALA TechSource
www.techsource.ala.org

Smart Libraries™

Formerly Library Systems Newsletter™

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

Febryary 2004 Volume XXIV Number 2

ISSN 1541-8820

Digitizing historical newspapers

Historical newspapers are in great demand in both public and academic libraries. Millions of pages of newspapers have been converted to microfilm, largely through the U.S. Newspaper Program funded by the National Endowment for the Humanities, but only a tiny proportion of these newspapers is available digitally on the Web.

Digital newspapers have great advantages over microfilm:

- They can be accessed from anywhere.
- They don't require mechanical readers for viewing.
- They allow interactive full-text searching.

At the same time, newspapers are more difficult and costly to make available digitally than books and serials, and newspaper presentation systems require special features. In the typical production system, newspaper pages are digitized into TIFF images and then zoned, or separated into sections representing articles, letters to the editor, blocks of advertisements, and other types of content. The images are then processed by optical character recognition (OCR) software to create searchable text.

Because of the expense of editing OCR output, uncorrected or dirty text is usually used only for searching, and a display copy of the page or article is created in PDF format. Metadata is supplied for issues and articles, and the OCR output is marked up with XML tagging.

See **Newspapers** *on page 2*

Open source gains interest

In library automation, commercial companies dominate. Open source (freely available software), however, offers an alternative approach for creating, distributing, and supporting automation needs. Open source has yet to make significant inroads in the ILS arena, but interest continues to intensify. In library automation areas other than the ILS, open source already has more of a presence.

In the broad cycles of technology trends, open source is in its infancy. But many developments related to open-source library technologies have occurred in the last year.

Koha gains ground

Koha continues to be the open-source ILS that draws the most attention. (The September 2000 issue of *Smart Libraries Newsletter* introduced readers to Koha.) Originally developed by Kapito Communications for the Horowhenua Library Trust in New

See **Open Source** *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.

E-RATE PROGRAM ADDRESSES MISUSE

The schools and libraries program of the Federal Communications Commission (FCC), better known as the e-rate program, has distributed $9.8 billion in subsidies since its inception in 1997, but it also has been plagued with charges of waste, fraud, and abuse. In December the FCC adopted new rules to improve the administration of the program.

The measures adopted include:

- Prohibiting the transfer of equipment purchased with e-rate discounts to other locations for three years after purchase, with limited exceptions
- Creating a more formal and transparent process for annually updating the list of services eligible for support
- Limiting support for upgrading or replacing internal connections to no more than twice every five years (basic maintenance services exempted)

In related news, the FCC concluded that about $420 million in unused schools and libraries funds from prior years (1999–2001) will be carried forward for disbursement in Funding Year 2003, making more benefits available to more recipients.

The FCC also released two appeal decisions covering denials of nine funding requests totaling more than $700 million for IBM services. One decision (FCC 03-313) upheld eight of the denials. This decision is particularly important because it contains significant clarifications and prospective interpretations of e-rate rules related to competitive bidding and Form 470 requests. All funding year 2004 applicants should read this decision carefully and follow its guidance.

In addition, the Schools and Libraries Division (SLD) of the Universal Service Administrative Co. (USAC) Task Force on the Prevention of Waste, Fraud and Abuse issued its long-awaited final report. “Recommendations of the Task Force on the Prevention of Waste, Fraud and Abuse” concludes that the discount matrix, as currently structured, unintentionally encourages some waste, fraud, and abuse, and it should be revised to increase the percentage of some costs paid by applicants.

The report also makes extensive recommendations regarding the clarity of rules and the effective use of resources in reviewing applications. A companion report by the USAC, “Interim Response to the Recommendations of the Task Force on the Prevention of Waste, Fraud, and Abuse,” identifies the parties responsible for addressing each recommendation, and indicates the actions USAC has taken or plans to take.—*PLC*

Contact: General E-Rate information, www.e-ratecentral.com
SLD Task Force Recommendations,
www.sl.universalservice.org/taskforce

Newspapers *from page 1*

Each step presents special problems:

- Scanning from microfilm is sensitive to the quality of the microfilm. Scanning from paper can require special equipment to handle the oversized pages.
- Articles and photographs may require different scanner settings.
- Imperfections that don’t bother human readers have a significant effect on OCR accuracy, as do the use of different fonts and type sizes.
- Retrieval and presentation also is more complicated for newspapers.
- Multiple columns and wide pages often require using special techniques for the page to make sense.

Desirable features in display systems include the ability to zoom in and out of different sections of text, to highlight search terms in the display, to view articles in their original page context as well as individually, and to view an article in its entirety, even when it extends over several pages in the original newspaper.

These challenges exist for both open and commercial collections. Libraries undertaking historical newspaper digitization projects usually outsource the process of digitization, OCR, and markup to vendors. Most noncommercial projects in the United States use OCLC’s Digital and Preservation Services or iArchives (see side story).

Although state-of-the art newspaper systems are wonderful tools for researchers, basic problems remain. The technology is expensive. MetaSearch, or cross-system searching, is generally unavailable for newspapers. Collections created at great expense are not portable, owing to the lack of standards for marking up historical newspaper content. These factors tend to favor large, cooperative state and regional projects over smaller institutional ones.—*Priscilla L. Caplan*

Sources for digitization

OCLC's Digital Collection and Preservation Services offers an end-to-end solution for digitizing newspapers. OCLC will digitize newspapers from print or microfilm and prepare them for loading into the ActivePaper Archive software developed by Olive Software, Inc.

In 2002 OCLC became the sole distributor of this application to the library market. After collections have been digitized, libraries can host them on their own Olive servers or on OCLC's Olive server. Libraries maintain control of the content and its distribution.

A mix-and-match model is offered by iArchives, a Utah company that specializes in OCR, indexing, metadata creation, and markup. iArchives will digitize from paper or microfilm or accept digital images created by a third party. The output can be loaded into any system that supports newspaper display.

The Utah Digital Newspapers Project recently loaded 30,000 newspaper pages converted by iArchives into CONTENTdm, a digital collection management system from DeMiMa, Inc., distributed by OCLC.

Contact: iArchives,
www.iarchives.com
OCLC Digital Collections and
Preservation Services,
www.oclc.org/services/
preservation

New name for VTLS

VTLS, Inc., has given its initials new meaning. Though the company was originally named for the Virginia Tech Library System, it now stands for Visionary Technology in Library Solutions.

The new name reflects the company's emphasis on its multifaceted approach to library automation, including Visual Multimedia and Imaging Solutions, the Virtua integrated library system, and its Vtrax division, which focuses on its Fastrac RFID products.

Other news from VTLS includes a partnership with SF-Systems, a major bookbinder, to fully integrate the LincPlus binding automation software into the Virtua library automation system.

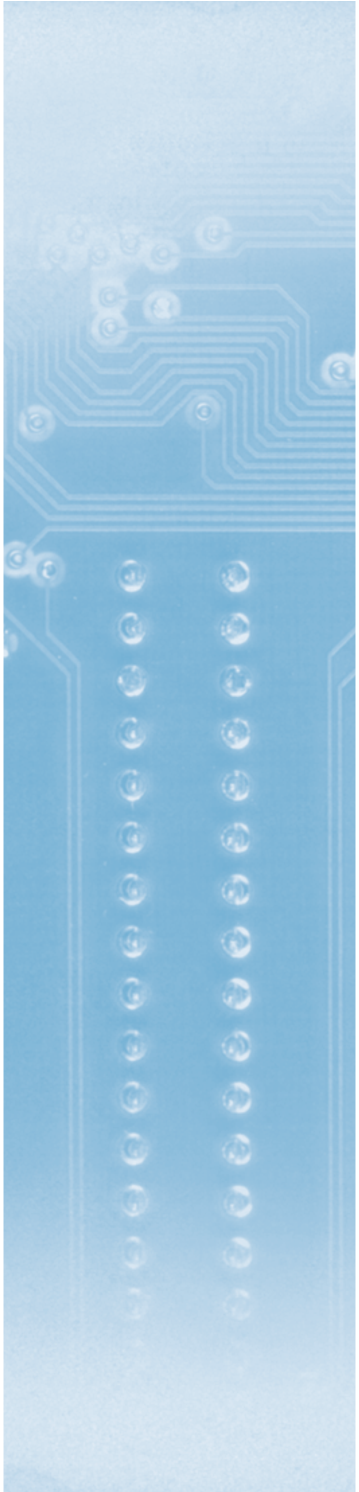
VTLS also entered a relationship with Sun to be an iForce Application Provider, which allows VLTS to designate model configurations and resell Sun systems to its customers, ensuring its software will be well supported on this platform.—*MB*

Contact: www.vtls.com
Sun's iForce partner programs,
http://au.sun.com/partners/become/whomayapply.html

Ex Libris excels in ARL market

With recent sales of SFX to Northwestern University and Vanderbilt University, Ex Libris now has its products in 51 of the ARL member libraries. Though its Aleph 500 library automation system has been sold to 20 ARL members, another 31 have purchased SFX, MetaLib, or DigiTool.

These numbers indicate Ex Libris' success in selling SFX to libraries that run a competing library automation system. Ex Libris' success is especially noteworthy considering that it was a late entrant to the North American library automation market. Its first major library, the University of Notre Dame, sign to purchase Aleph 500 in 1998.—*MB*



Open Source *from page 1*

Zealand, it has enjoyed continued development by a group of contributors around the world, following the vision of collaboration associated with the open-source model.

Development and distribution of Koha takes place on SourceForge, one of the primary repositories of open-source software. To date, only a handful of libraries have implemented this system. But at least one U.S. public library—the Nelsonville Public library in Athens County, Ohio—has migrated from a commercial system (Civica's Spydus) to Koha.

General use of open source

Open-source software continues to be used in some libraries' general technical infrastructure. Many types of organizations, including libraries, rely on the Linux operating system and the Apache Web server, two of the most successful open-source projects.

According to figures from Thomas Dowling, who administers the LibWeb directory of libraries on the Web, of the 6,892 library Web servers monitored, 39% use Apache, 42% use some version of Microsoft's IIS Web server software, with the remaining 19% using other Web server software. Dowling also notes, though, that the overall percentages of library Web servers based on the commercial servers from Microsoft are increasing and the library use of the open-source Apache software servers is decreasing.

See **Open Source** *on page 6*

Open source sources

Other open-source applications relevant to libraries include:

- D-Space. Institutional repository software jointly developed by Hewlett-Packard and MIT. https://dspace.mit.edu and www.dspace.org
- Fedora. Digital library software developed by the University of Virginia and Cornell University with financial support of the Andrew W. Mellon Foundation. www.fedora.info
- The Greenstone digital library software created through the New Zealand Digital Library Project at the University of Waikato with support from Unesco. www.greenstone.org
- SiteSearch. A Z39.50 gateway environment and digital library toolkit, originally created by OCLC. Originally licensed as a proprietary product of OCLC, SiteSearch has been transformed into the open-source framework. www.sitesearch.oclc.org
- MyLibrary is a customizable library portal application designed to organize information through a library Web site and present a personalized interface for library users. http://dewey.library.nd.edu/mylibrary.
- Many open-source utilities related to the Metadata Harvesting Protocol of the Open Archives Initiative are listed on the Open Archives website. www.openarchives.org/tools/tools.html

FREE PORTAL APPLICATION AIDS METADATA HARVEST

Libraries looking for a portal or metadata creation tool to support a particular project, collection, or community should check out the Collection Workflow Integration System (CWIS), a new application available without charge from the Internet Scout Project at the University of Wisconsin–Madison.

CWIS builds on the popular Scout Portal Toolkit to provide a complete portal application along with tools for creating, harvesting, importing, editing, and exporting metadata. (It should not be confused with the acronym for "Campus Wide Information System," an early predecessor of the academic website.)

CWIS was created to help sites contribute to the National Science Digital Library (NSDL). It allows content providers to create NSDL-compliant metadata and expose it for harvesting into the larger NSDL portal. Since the NSDL uses a profile of the Dublin Core Metadata Element Set and the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), CWIS supports these widely useful standards. This synergy is possible when communities coalesce around a compatible set of open standards.

The public portal supports both simple and advanced searching as well as browsing by subject categories. Result sets can be displayed in short or full format. Users can establish their own logins and customize their interface. They also can set up user agents, saved searches that can run on a daily, weekly, or monthly basis and e-mail back result sets.

The real strength of the public portal, however, lies in its support for communities of users through devices such as news, electronic bulletin boards (forums), end-user rating of resources, and user recommendations.

Authorized users also can access forms for central or distributed input and edit of Dublin Core–compliant metadata records. The data entry template can be customized with local fields and authority lists. Metadata also can be batch-loaded and exported, as well as exposed for harvesting in standard OAI format or the NSDL variant.

CWIS is turnkey software that can be installed and administered with a minimum of technical expertise. It runs under Linux and requires a Web server with PHP and MySQL. It is an unusually well-designed application that installs neatly. The user interface is clean and easy to use, and the help is actually helpful.—*PLC*

Contact: http://scout.wisc.edu/Projects/CWIS

Open Source *from page 4*

Librarians are asking if the growing general interest in open source in libraries will be reflected in an eventual increase in the adoption of Apache.

Web additions

Another opportunity for the use of open-source software lies in the area of managing the content of a library's Web environment. Many libraries have developed database-driven applications that manage access to e-journals, databases, and other resources.

The University Libraries Digital Library Development Lab at the University of Minnesota–Twin Cities developed a system it calls LibData using Apache, My SQL, and PHP. It has made the system available for use by other libraries through the GNU Public License—a software license that embodies the principles of the open-source community. LibData has been in production use for the University of Minnesota Libraries since fall 2003.

Related association activities

The Library and Information Technology Association (LITA) of the American Library Association (ALA) oversees many activities promoting open-source software in libraries. The LITA Open Source Systems Interest Group was established in 2000 as a forum for librarians interested in this topic.

As part of its Regional Institute program, LITA offers a full-day workshop on "Open Source in Libraries" presented by Eric Lease Morgan, the author of the MyLibrary open-source application. LITA offers each Regional Institute in venues across the country, typically scheduled three or four times a year.

Boost for U.S. market

Index Data, based in Denmark, has long developed library-specific open-source software. David Dorman, previously employed as a library consultant for the Lincoln Trails Library System and author of the Technically Speaking column in *American Libraries,* was named U.S. marketing manager for Index Data in November 2003. Although Index Data may not be a well-known name, the company has established itself as a developer of professional quality software that operates behind the scenes in library applications.

Software developed by Index Data includes the toolkits for the Z59.50 search and retrieval protocol that can be integrated into library applications to add Z39.50 capabilities. Specific products include the YAZ Z39.50 C toolkit, the YAZ++ toolkit for C++; ZAP!, a Z39.50 client for the Apache Web server; Zebra, an XML-based indexing and retrieval engine; and TKL, an XML-based content management system.—*Marshall Breeding*

Contact: SourceForge, http://sourceforge.net/projects/koha
Nelsonville's implementation of Koha, www.athenscounty.lib.oh.us
LibWeb directory of libraries, http://sunsite.berkeley.edu/Libweb
LibData, http://sourceforge.net/projects/libdata
LITA, www.ala.org/lita
Index Data, www.indexdata.dk

MILLENNIUM, ENDEAVOR tie for top ARL sales

The scorecard of Association of Research Libraries (ARL) member libraries changes once again with the announcement that the University of North Carolina at Chapel Hill will be migrating to Millennium from its current DRA Classic system. With this contract, Innovative ties with Endeavor for the top rankings in providing library automation systems to the ARLs—both companies now claim 35 of the 124 total members. Innovative held the largest number of these libraries from the late 1990s through the summer of 2003.

Although Millennium consistently enjoys strong sales, this sale marks the first new-name ARL that has chosen Millennium since 1999. The close competition among these libraries shows that no single company dominates in high-end academic library automation.

This selection also reflects the diversity of choices the Research Triangle Library Network (TRLN)—Duke University, UNC Chapel Hill, North Carolina State University, and North Carolina Central University (NCCU)—has made in automation strategies. This group of libraries previously participated in a library automation network based on DRA Classic. Each has gone a different direction. North Carolina State University selected Sirsi's Unicorn and Duke opted for Ex Libris' Aleph 500. NCCU has not made plans to migrate from DRA Classic.

This set of strategies seems somewhat contrary to the trend of ever-larger groups of libraries partnering to share a single library automation system. It does, however, highlight the expectation that libraries expect a high degree of interoperability and resource-sharing capabilities from their automation systems even when different companies provide them. Even though the TRLN libraries have selected different automation systems, the consortium remains strategically committed to sharing resources among its members.—*MB*

BiblioMondo offers enhanced serials control

The latest release of BiblioMondo's Portfolio offers a completely redesigned serials control module intended to set the industry standard. The highlight of Portfolio Version 6, released in November, is the new serials module, which can be run as part of the Portfolio library suite or as a standalone product linked to any other library system.

The new serials control module includes these abilities:

- Searching can be done by the Serial Item and Contribution Identifier (SICI).
- Routing management is more sophisticated.
- Check-in is improved.
- The improved check-in allows librarians to do the following:
- Record publication patterns to the full six levels defined in the MARC format for holdings data
- Change months and seasons to match the language of the publication
- Easily separate regular and supplemental content
- Add analytical entries to the issue's bibliographic record at the point of check-in

Portfolio Version 6 also contains many new features in other modules including support for EDIFACT, Z39.50 searching from all core modules, and URL validation in cataloging.

BiblioMondo was formed by the acquisition of ALSi, an U.K.-based library automation company, by the French Canadian firm Best Seller. Headquartered in Montreal, BiblioMondo provides integrated library systems, including Concerto for public libraries and Portfolio for public, academic, and special libraries. It has 1,400 customer libraries in 11 countries served by offices in Britain, France, the Netherlands, Germany, Canada, and the United States.—*PLC*

Contact: www.bibliomondo.com

ALA TechSource
www.techsource.ala.org

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

February 2004
Digitization solutions and open source sources

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.