

DIGITAL COLLECTIONS MANAGEMENT

Prepared by Richard W. Boss

Libraries increasingly seek to maintain digital collections as well as print and audio-visual collections. The digital collections may include text, images, audio, and video. As of mid-2006, the most common digital collections in public libraries were images of local newspapers and historical photographs.

The motivation for maintaining digital collections may be to preserve fragile originals by limiting handling, improve access by making the information available on the Web, or both. The question of preservation and access usually is addressed before “capture” of the documents in digital form, rather than at the time a digital collections management product is purchased. For that reason, this TechNote does not address a library’s motivation.

Digital collections management, as used in this TechNote, has sometimes been confused with electronic resources management. The former deals with the management of the digital files of a library or consortium; the latter deals with the management of subscriptions to electronic resources supplied by others. Electronic resources management is the topic of a separate TechNote.

While community archives may be part of a digital collections program, they are treated separately in the TechNote entitled “Institutional Repositories.”

History

Digital collections management began in the late 1980s as a tool for use by large businesses and organizations. The first systems were developed in house by companies with large Web sites such as 3M and Amazon.com. The early efforts by libraries were by the Library of Congress and academic research libraries.

The Library of Congress began its digital collection in 1989 with a five-year pilot project. It

launched its American Memory Project (<http://memory.loc.gov/ammem>) in 1995 and reached its goal of five million documents by 2000. LC has raised substantial funds from private sources to support its National Digital Library Program, one that seeks to digitize not only materials from its own collections, but to stimulate efforts around the country.

The University of California began development of its collaborative California Digital Library (www.cdlib.org) in 1996. A number of other major collaborative efforts were launched shortly thereafter. A directory of scores of collaborative projects that focus on cultural heritage materials may be found at www.mtsu.edu/~kmiddlet/stateportals.html

In the past five years, a number of academic and public libraries have undertaken independent digital collections projects, many of them linked to collaborative digital collections sites in their states.

Digital Collection Management Products

When packages began to be offered commercially as CMS (Content Management Systems) in the late nineties, libraries were among the organizations that purchased them. By 2001 there were some 98 software products available and more than 5,000 systems had been implemented, however, purchases by libraries were few because none of the products were specifically designed for the library market and few were comprehensive solutions that included hardware, software, and services. They were also very expensive. The need to adapt the products to the needs of libraries added to the expense. Among the unique needs of libraries are the incorporation of the Dublin Core plus metadata scheme and support for Z39.50.

There were a few open-source alternatives to the expensive commercial products available as early as 2002, among them DSpace (www.dspace.org), a software package developed by MIT and Hewlett-Packard; Greenstone (www.greenstone.org), a collaborative effort from New Zealand; and MyCoRe (www.mycore.de), an initiative of several German academic institutions.

By 2005, libraries had a choice among several digital collections management systems designed specifically for libraries by vendors of integrated library systems. They included Endeavor's ENCompass for Digital Collections (www.endinfosys.com), ExLibris' DigiTool (www.exlibris-usa.com), Innovative's MetaSource (www.iii.com), OCLC's ContentDM (www.oclc.org/contentdm) SirsiDynix's Hyperion Digital Media Archive and Horizon Digital Library (www.sirsidynix.com), and VTLS' VITAL (www.vtls.com). As of the end of 2005, these vendors had at least 135 digital collection management contracts. The purchasers had the option of purchasing software only or "turnkey," including hardware, software, and services.

Libraries that launched digital collections on PCs, Web servers, and as special files on integrated library systems before purchasing a digital collections management system, have sometimes moved them to the new system, but just as often have created links from the digital collections management system to wherever the older digital collections were mounted.

All of the products developed for the library market have the ability to import, catalog, edit, store, search, retrieve content and metadata (bibliographic information about digital documents), and generate statistics. They generally do not include hardware and software for the digitization of source documents. It is assumed that libraries will do that in-house on their own equipment or will contract it out to a firm that specializes in digitization.

Importing Documents and Metadata

The ability to import digital documents is a standard component of all of the products. The documents can be downloaded from a digital storage medium, by FTP, or directly from any TWAIN scanning device. The documents can be images (GIF, JPEG, PDF or TIFF), plain-text, HTML, audio, or video. Generally only a few parameters are required to load documents. It is also possible to load metadata that already exists for the documents being imported.

Cataloging

The bibliographic information for a document in digital form is generally referred to as metadata. [See the TechNote entitled “Metadata: Always More Than You Think.” There are several options, including MARC21, EAD, TEI, and Dublin Core, but the most widely used for text is MARC21, for archival materials it is EAD, and for images it is Dublin Core. TEI (Text Encoding Initiative) was developed so that scholars could encode their own text, but it has had only limited use.

Several products offer the option of cutting and pasting information from the actual document to facilitate cataloging. Most of the systems do not limit the number of fields of metadata that can be created. At a minimum, the metadata will include the name of the author, title, language, the source, date, condition, access restrictions, and content description. The metadata can be linked to a single document, also known as an object, or to many.

After the documents have been cataloged, virtual collections can be created. A single document can appear in several virtual collections.

Editing

Digital content can be edited, including reducing high-resolution JPEG or TIFF documents to lower resolutions or creating thumbnails. Metadata can not only be created, but edited on a digital collections management system.

Storage

Storing digital files demands substantial disk storage capacity because digital files other than text files can be extremely large, especially color images and videos. Multiple disk drives are recommended in order to limit I/O (input-output) contention, thus improving response times.

Searching

Searching can be done directly against the digital collections management system or through an interface from a library's patron access catalog, portal, or Web server. When a portal is used, simultaneous searching can be undertaken not only against the documents on the digital collections management system, but also against other electronic resources within the library or on the Web.

Because they have been developed for libraries, the systems support Z39.50 in addition to XML and HTTP gateways.

Searching may be limited to specific documents, individual collections, multiple collections, or may encompass all collections.

Searching can be done either against both the full-text of the documents and the associated metadata, or against just the metadata. Non-text documents searching is, of course, limited to searching the metadata. In most cases, searches can be natural language, Boolean, proximity, wildcard, truncation, date and date range, and pattern. Pattern searching makes it possible to find words with incorrect or missing characters either in the search term or the body of the text.

Most of the system return merged result sets. Some also return relevancy-ranked results.

All of the systems include an image viewer for JPEG and TIFF files, enabling image manipulation and printing. There is also the option of including thumbnails for faster retrieval and display, with the option of clicking to the higher-resolution image. A user may be limited to viewing only the thumbnail when there are restrictions on access, or when access is subject to the payment of a fee.

Security generally includes patron authentication to control access to the documents based on

the profile of the documents and the users. For example, some documents may be subject to copyright, access restrictions required by donors, or the desire of a library to charge for printing or downloading. Some categories of users may not qualify for access because they are not registered borrowers, are not adults, or are delinquent borrowers.

Statistics

All of the systems include a number of reports that summarize loading, cataloging, editing, and searching activity.

Costs

The cost of digital content management systems is related both to the number of documents and number of concurrent users supported. An entry-level system supporting fewer than 10,000 documents and ten concurrent users will cost a minimum of \$25,000, including \$7,000 or so for software licenses, \$14,000 for hardware, and \$4,000 for training and other services. Annual hardware/software maintenance costs are a minimum of \$3,000 per year after the first year.

Libraries that require a small system may wish to consider a hosted service. Most of the vendors mentioned in this TechNote will mount the documents on a hosted server at the vendor's site for an annual subscription fee.

As the size of the system increases, it becomes more cost effective to purchase the system.

Sources of Information

Currency is extremely important when seeking information about digital collections management. The Web sites of the vendors cited in this TechNote are highly current sources of information. Also highly current is *D-Lib Magazine* (www.dlib.org). It characterizes itself as the “magazine of digital library research.” The Digital Library Federation (www.diglib.org), a

consortium of “libraries and related agencies that are pioneering in the use of electronic information technologies to extend collection and services” also has highly current information on its Web site.

Three good examples of digital collections maintained by public libraries are those of the Chicago Public Library (www.chipublic.org/digital/digital.html), the Everett (WA) Public Library (www.epls.org/nw/digital_collections.htm) and the New York Public Library (www.nypl.org/digital).

Other useful sources of information are <http://www.ifla.org/II/metadata.htm> for metadata, <http://libraries.mit.edu/guides/subjects/metadata/standards/tei.html> for TEI, <http://www.loc.gov/ead/ag/agappc.html> for EAD, and <http://dublincore.org> for Dublin Core.

Prepared August 18, 2006