

IT Feature

UMI's Digital Vault Initiative Project

With experience and content, UMI is poised for a conversion megaproject

by Péter Jacsó

I have always enjoyed ALA conferences for many reasons. Beyond the obvious and noble ones (talking to librarians, vendors of library automation projects, and individual developers showing their remarkable software at exhibit desks), there are also the more mundane ones: the product announcement breakfasts (or lunches, five o'clock teas, or time-independent buffets). This summer at ALA I was impressed by a number of new projects and products, but the showstopper for me was a UMI breakfast program—not so much for culinary reasons as for intellectual content.

I knew that Pulitzer Prize-winner Doris Kearns Goodwin, the featured speaker at the breakfast program, was an outstanding historian and biographer of LBJ, the Kennedys, the Roosevelts, and even the Dodgers baseball team, but I did not know what a mesmerizing speaker and person she was until I heard her talk. She was the perfect speaker for this event, at which UMI announced its Digital Vault Initiative, even though she never uttered the word UMI, let alone Digital Vault, and mentioned microfilm only twice. Nevertheless, her genuine, warm, and engaging storytelling made it clear how important it is to have access to original contemporary documents to learn and corroborate the details of events and to uncover the truth.

And what organization is better suited to undertake such a project in the era of the nascent digital library than UMI? The company's information assets, software, marketing, and technology know-how make UMI the prominent architect, civil engineer, carpenter, information broker, and real estate agent in building the digital library. UMI has been poised to do this, leveraging its experience and its assets in the process as no one else can.

60 Years and Counting

None of UMI's promotional materials for the Digital Vault Initiative sported a logo featuring the number 60, but it was exactly 60 years ago that UMI recorded on microfilm the Early English Books collection of the British Museum to preserve and distribute it to libraries around the world in a compact form: microfilm. That was then followed by some 6 billion pages that have included books, journals, newspapers, dissertations, BBC broadcast summaries, legal documents, and historical papers.

Although UMI stopped using its long name a few years ago, many remember that indeed "UMI" was an acronym for University Microfilms International, and UMI did not stop doing microforms. On the contrary, it made them more useful by first making the indexes to much of its microform collection available in computer-readable format, then the abstracts, then the full-text version. In the early 1990s the company introduced the idea of digital facsimile images of newspaper and journal articles through its PowerPage databases. It is this latter type of digital infor-

mation that best justifies the use of the term digital library. Many of the books and, especially, the magazines, journals, and newspapers have characteristic typography, page design, and layout that only the facsimile page-image version can reproduce or "make it similar," as the Latin original (facet simile) implies. In April 1994, in an *Information Today* column entitled "Document Cloning Software: The Kiss of Death for Microfilm

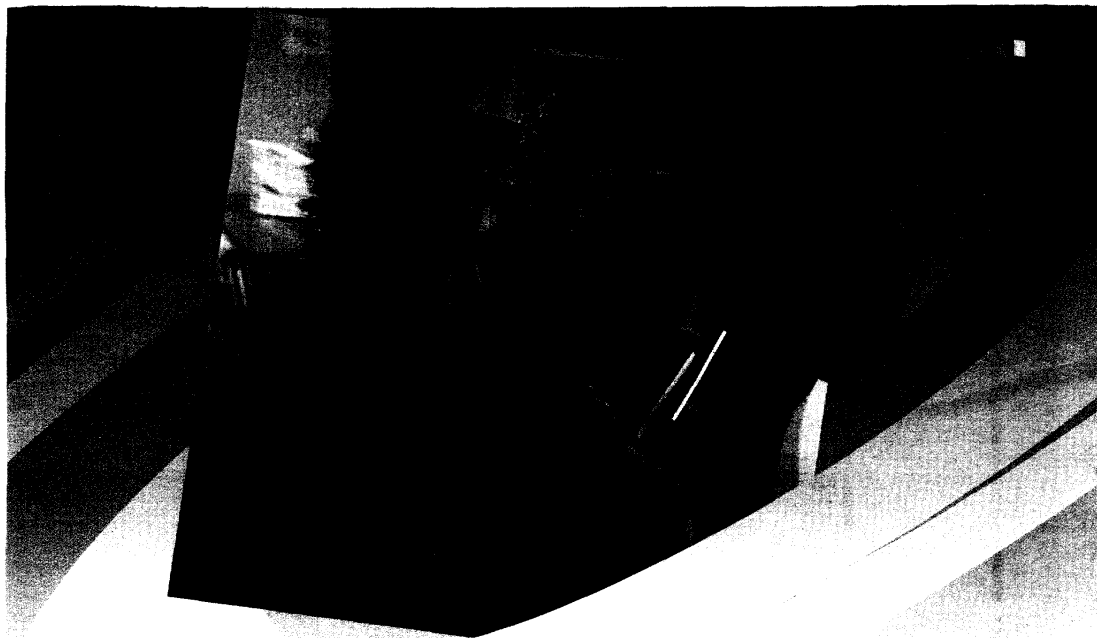
and Microfiche," I extolled the virtues and power of page image documents.

I have praised the various PowerPage databases in my columns in *Information Today* and *Computers in Libraries*, and I have been pleased to see these databases improve in quality, become colorful (selectively), and change to the most widely used standard document image format (Adobe PDF). This latter aspect is especially important because

it paves the way for one feature that many consider to be another important trait of digital libraries—the interlinking of documents through references and citations.

Ovid, using SGML coding of texts, has perfected this technology, allowing users to click on a reference or on citations in one item to display the abstracting-and-indexing record from MEDLINE or the full text of the article from one of Ovid's Biomedical Core Collections along with separate images. UMI does not have this feature yet, but Adobe keeps developing its Acrobat Exchange and Capture programs, and third-party developers offer PDF plug-ins that would allow similar interlinking.

(continued on page 61)



You've done difficult.

Are you ready for easy?

While they may be familiar, the old ways are downright difficult. But checkouts, searches, reports, and inventory are all a breeze when automated. And nothing makes them easier than Winnebago Spectrum®.

"Winnebago Spectrum is very user-friendly from both the librarian and student standpoint."

Marie Carter, Media Specialist

GUI means it's easy to use

Winnebago Spectrum's intuitive, button-based GUI (graphical user interface) makes searches and circulation tasks simple. You'll spend your time using your automation system, not learning it. And Spectrum's multi-tasking/multi-windowing ability lets you keep circulation, catalog, and reports windows open simultaneously—you just click in the window you want to work in.

The power to grow with you

Award-winning Winnebago Spectrum will handle the most demanding library's needs—including yours. You can choose Windows or Mac OS for

your server operating system, run Spectrum alone or on a network, and even connect Windows and Macintosh client computers together for circulation and catalog stations. Plus, with TCP/IP Spectrum is Internet ready, so patrons can access your collection via the Internet.

'Year-2000 safe'—like all Winnebago programs

Your library will move smoothly into the new millennium because Spectrum is "year-2000 safe!" Call 1-800-533-5430 ext. 1565 and tell our Library Automation Specialists you want more information and your FREE Guided Tour of Winnebago Spectrum—the program with the ease of Windows and the power of Winnebago!

Call for your FREE Guided Tour demo
1-800-533-5430 ext. 1565



Winnebago Software Company • 457 East South Street • Caledonia, Minnesota 55921 USA • Worldwide 507-724-5411 • Fax 507-724-2301 • sales@winnebago.com • www.winnebago.com

September 1998

UMI's Digital Vault Initiative (continued from page 15)

UMI has had experience with different solutions. In ProQuest Direct, the online version of PowerPage databases, UMI offers three different types of full-text records: plain ASCII format, ASCII with separate images, and the PDF format. This is very useful as some documents do not really call for the facsimile reproduction of the original page, and the plain ASCII text format is the best (the least expensive and the most legible), especially if the original font size is rather small and the type is dense.

The advantage of linking to external documents is obvious when you look at the very legible reproduction of Klemperer's excellent resource guide about digital libraries in ProQuest Direct and at the HTML version (http://www.lita.org/ital/1603_klemperer.htm) with hotlinks to the cited documents. However, publishers of scholarly journals have begun to make their articles available in HTML form only selectively and only exceptionally, whereas UMI offers an immense digital collection of a great variety of prestigious sources right now.

An Immense Digital Collection

UMI has a head start because it began the document digitization project even prior to the 1993 start of the Digital Library Project funded by the National Science Foundation. The first phase of that project is just ending as you read this, and grant recipients will soon be announced for Phase 2. The six recipients of Phase 1 have reported impressive achievements for the relatively small amount of about \$25 million dollars received between 1994 and 1998, and NASA, the Library of Congress, and the National Library of Medicine have also already filled "digital shelves" in the Digital Library. Not surprisingly, funding will double in Phase 2—but will remain minuscule compared to the amount of money wasted on less noble projects.

For its part, as of mid-1998, UMI has converted well over 50 million pages into one or more of the above-mentioned digital formats. The result of a simple search I conducted in the full ProQuest Direct database illustrates the variety of journals and types of documents that are available in full text in one or more of the digital formats. There were articles from trade journals (*Computers in Libraries, Database, Online, Library Software Review*), newspapers (*Information Today, Chronicle of Higher Education*), and scholarly journals (*Science, Communications of the ACM, Information Technology and Libraries, ACM Computing Surveys*).

Getting the rights for digital publishing from the publisher is the bane of every digitization project. The copyright issues involved are vexing, but UMI has extensive experience in the field, having worked out agreements with 8,000 publishers and having learned how to handle royalty accounting and payments to publishers deftly—no small feat. Other digital libraries may show superb technical solutions, but copyright negotiations can wear even the strongest advocates down. Some publishers in our own field are extremely conservative and do not make articles available in full-text format for third parties. With the advance of digital libraries, they will soon realize that their most important authors (the ones

A SPECIAL WEB SITE TO DEMONSTRATE THE DIGITAL VAULT

This article is also available not only on the Information Today, Inc. Web site at <http://www.infotoday.com> but additionally as a vastly enhanced HTML document on a special Web site hosted by UMI, accessible to anyone, at <http://www.umi.com/hp/News/Reviews/SiteBuilder.html>. The electronic version on the UMI Web site offers hyperlinks to many documents in full text and/or Adobe PDF format from UMI's existing digital vault, and to Web sites as well. I wish to thank UMI and, in particular, Tom Saltsman, senior project manager for UMI's SiteBuilder software for the help I received in implementing this special Web site. The site illustrates some of the current capabilities of the UMI's Digital Vault Initiative, a project that promises to offer Web access to 5.5 billion pages of historical documents in digital facsimile format in the near future. Details of the project are laid out in the July 13, 1998, NewsBreak story by Paula Hane that appears both on the Information Today, Inc. Web site and in this issue of *Information Today* on page 4. —P. J.

who are no longer under the pressure of tenure-driven publication needs, preferring now to write for a wider audience) will desert them if their journals are only available through the publisher's digital boutique library rather than the digital superlibraries.

UMI also has a very wide scope in terms of subject fields and its asset of 1.5 million dissertations. The dissertation abstracts are widely available through third parties, but digital facsimile equivalents of dissertations are available only from UMI itself. Currently, the number of dissertations in digital facsimile format is about 80,000 (amounting to about 20 million pages). The technology and the production process are clearly well-established, and the alternatives for the financial remuneration for authors are exemplary. While not as well known as its digital dissertation and periodical collection, UMI also has a Digital Book Vault of out-of-print books for on-demand printing. It is only a short leap to make this available on the Web.

The recently announced Digital Vault Initiative project will dwarf these numbers, but it is credible precisely because of UMI's many years of experience. As we went to press, few details were available about the priority to be given to sources in the digitization. UMI's press release mentions prominently the Early English Book Collection. I understand the homage to be paid to the pioneering work, but I wonder if Culpeper's *The English Physician* will generate as much interest as *The English Patient* did. Aside from the pun, I believe it would be much better to start with more recent material and work backward in time, for two reasons.

First, even in the humanities, the citation half-life is measured in decades rather than centuries, which clearly suggests priority for documents of the 1900s. The topics of the dissertations, which UMI must have a good grasp of, certainly would reconfirm this priority of scheduling.

The other reason has to do with the quality of information and—possibly later—

optical character recognition (OCR) in the conversion process. In my modest experience in the field, I have seen remarkable improvements in the past 18 months in scanning and OCR technology, not to mention reductions in pricing. By the time UMI would need to deal with digitizing its microform collection of the 15th century—4 or 5 years and a few billion pages from now—improvements in the technology and the company's own immense first-hand experience could make that phase of the project far more efficient.

Software for Finding the Gems

It is one thing to digitize print and microform documents in the billions, and it is another to provide efficient access tools to find the most relevant articles, books, or dissertations. UMI has a significant advantage in this regard, too. Full-text searching in itself is not a panacea. Searching through authority files, controlled subject vocabularies, and thesauri often represent a better solution. UMI has extensive experience in indexing source documents, especially business journals, newspapers, and general-interest periodicals.

The same applies, although to a lesser extent, to abstracting, where H.W. Wilson's experience and quality are unbeatable. It's no wonder that UMI recognized this and teamed up with that company to include several of the H.W. Wilson full-text files enhanced with abstracts in its offerings. Abstracts will remain an important asset both in optimizing retrieval by limiting the search terms to title, descriptors, or abstracts, and in judging the relevance of retrieved documents. H.W. Wilson already has top-notch abstracts going back to the mid-1980s for the most widely subscribed-to news magazines. This also supports the idea of starting with more recent material and working backward in time in the Digital Vault Initiative project.

The other aspect of search efficiency is of course the software. UMI has been doing well in this arena as well, even though it has not been doing pioneering work—until now. The current version of ProQuest—version 1.6—offers three search modes, field-specific searching with Boolean and proximity operations, and convenient date searching. Version 2.0, to be released about the time this article is published, will also offer thesaurus browsing and marking of items in the results list—both badly needed. The ability to browse the content of other indexes, such as the author name field, and to sort the results would also be helpful.

Even with these limitations, the ProQuest software runs circles around many of the search software applications offered directly by the publishers of journals that go on the Web on their own, and it even beats those of several commercial online publishers. In addition, ProQuest Direct users will need to learn only one search language to search megadatabases. This is an immense advantage even while we are waiting for further results of studies of users' search behaviors and preferences.

Building the Digital Carrel

There is an area in software development where UMI definitely is doing the pioneering work. Anyone who has recently been on either side of the cathedra in the classroom will understand the nightmares of making available current and appropriate reading materials for students. To create a course pack, the profes-

sor must go through an arduous process of getting copyright clearance from publishers. Often, professors get no response at all, let alone a positive one. If the response is positive, the bottom line may be quite discouraging. A further issue is reserve rooms and collections, which are nice when they work. But neither librarians nor students are crazy about them, and they are often closed when students need them most—the night before exams, for example. And giving students URLs in the reading list is a solution for only a tiny fraction of the typical reading materials. Half of the URLs will yield the "404 File Not Found" message 2 weeks after the reading list is created.

Enter UMI's ProQuest Direct SiteBuilder. I cannot do justice to this software in this space, but suffice it to say that with help from SiteBuilder project manager Tom Saltsman, I used SiteBuilder to create the Web site on which the enhanced HTML version of this article resides. Further, with the help of the upcoming SiteBuilder Wizard, I plan to use it on my own for a course pack or virtual carrel for a course in the fall semester. Searching the ProQuest Direct database, I can select the most appropriate articles in the most appropriate format, then drag and drop them into my virtual carrel. There, I can organize them in the desired sequence, annotate them if needed, and simply give my students a URL.

Students will have access to the facsimile or full-text ASCII versions of the articles day or night (apart from breaks for maintenance) without any hassle and without having to track down journals on the shelves and carry 2 pounds of nickels and dimes for a photocopy machine. If UMI spoils today's students rotten, it will have a captive audience when they become managers, decision makers, information specialists, professors, journalists, or biographers who will have to find, select, and organize the documents retrieved from the Digital Library as a customized, ready-to-use document package.

While Doris Kearns Goodwin brilliantly manifested the beauty and power of the spoken word in her talk at UMI's breakfast program on the Digital Vault Initiative, she just whetted our appetite for the power and durability of the written word. Every guest received a complimentary copy of her latest book, which she kindly autographed. Many of us could also have had the option of heading for the nearest bookstore or public library to pick up one of her other books, to prolong and savor the precious hour of her talk. But there were probably others in the audience from countries where her books are not available. UMI's Digital Vault Initiative may be the answer for them, and for millions whose best chance for attaining literacy in the natural sciences, humanities, arts, and social sciences is through such initiatives. I applaud UMI for its Digital Vault Initiative as much as I applauded Doris Kearns Goodwin for her speech. I'd like to see her books make it to the Digital Vault first as a symbolic gesture.

Péter Jacsó is associate professor of library and information science at the department of information and computer sciences at the University of Hawaii. He won the 1998 Louis Shores/Oryx Press Award from ALA's Reference and User Services Association for his discerning database reviews. His e-mail address is jacsop@hawaii.edu.