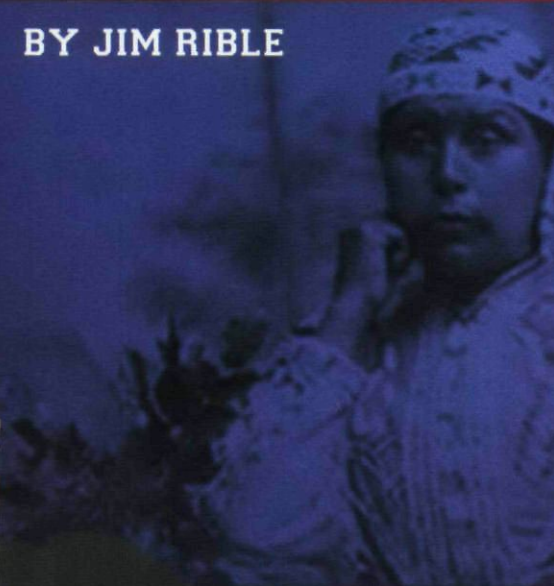


THE DIGITIZING PROJECT
THAT MADE

SODA

BY JIM RIBLE



WHEN WE UNEXPECTEDLY GOT A GRANT TO CREATE A DIGITAL LIBRARY, WE WERE BUBBLING WITH ENTHUSIASM FOR THE NEW CHALLENGE.

What would you do if you were suddenly given half a million dollars?

In late 2000, while looking for funds to help build an addition to the Southern Oregon University Library, we stumbled upon the opportunity of a lifetime. Our congressman Greg Walden, R-Ore., had the opportunity to procure funds to create a digital library. While he couldn't help us with our building project, he said, would we be interested in digitizing? We were now! So with a congressionally directed National Leadership Grant of \$470,000 from the Institute of Museum and Library Services (IMLS), the Southern Oregon Digital Archives (SODA) was born.

'Oregon-izing' Our Efforts

We started with many limitations. We knew very little, we weren't going to know a whole lot more soon, and we couldn't give up our current responsibilities. So, what were our strengths? Tenacity, commitment, and money! As professional librarians, we are involved in a business that thrives on the love of making the printed word available to others. But this was a whole new world of printed words.

We quickly formed a team and assigned duties. As the systems librarian, my role was to figure out the technical challenges and to act as the project manager. Our cataloger, Lisa McNeil (and another hired later, Kate Cleland-Sipfle), took on the task of creating a Document Type Definition (DTD), after

first learning what one was and did. Our associate director, Teresa Montgomery, took the role of project director and chief facilitator to manage the rest of us. Mary Jane Cedar Face, our collection development librarian, and Deborah Hollens, our government publications librarian, would become the overseers for what would become the First Nations/Tribal Collection and the Bioregion Collection, respectively.

We decided that the materials for our project would come from our library's rich collections of federal, state, and county publications—most of it gray literature. Southern Oregon is rich in Native American culture and history and is so unique in its biodiversity that the World Wildlife Fund, an international environmental organization, has one of its regional offices located here. Together, these collections would become the SODA database.

Doing Research on SODA: Formulas That Others Used

Our vision for SODA was to create something akin to one of our full-text periodical databases, where users could enter keywords in a variety of fields, retrieve a list of titles, and then view the full text. We began our research by reading all we could and by contacting librarians involved with digitization projects. What we found was surprising. Almost everything we read on digitizing came from the world of archiving and preservation. Most of the projects we viewed were small—less than 200

objects—and contained rare pictures and/or manuscripts. There was nothing like we wanted to create for SODA.

The only project we found that came close was the Making of America (MOA),¹ a collaborative effort between the University of Michigan and Cornell University. It was a magnificent project, but not one that a medium-sized library like ours could easily undertake. The Making of America contains more than 1.5 million images from approximately 5,000 early American journals and monographs. Every image is a separate digital object and is converted on-the-fly by a user's click from a TIFF to a GIF for viewing. Staff at Cornell and Michigan obviously had a tremendous amount of expertise and technology at their disposal, beyond anything we could imagine for ourselves with our current grant. They eloquently preserved old and rare documents and made them available on the Web for full-text searching.

Despite the enormity of it all, MOA was the closest thing we found to the formula we had in mind. So I called the creators. I had the good luck to find that Cornell Library's Department of Preservation and Collection Maintenance was offering a weeklong workshop on digitizing—just the thing we needed. (The research for that workshop has been made into a book, and I recommend it to anyone interested in digitizing.)² My time at Cornell was extremely rewarding. I had time to hang out with people who knew the ins and outs of scanning and to ask them lots of annoying questions. While that week was great, I think

»

I learned more about what we *didn't* want to do for SODA, or the *way* we didn't want to do it, rather than about what we actually *wanted* to do. However, one very important concept I picked up there was the "digital master."

The digital master is a digital object (an image file) created from the original document in its rawest and most uncompressed form, usually something like a 600 dpi TIFF. A TIFF is a lossless, high-resolution image that ensures you can use it again and again to create other copies at the same, or lower, resolution. For example, a photo placed on the Web only for viewing doesn't need to be any higher than, say, 72 to 100 dpi. But if you want to create an 8 x 10 photograph you will need a higher resolution, say a 300 dpi image. This too can be placed on the Web, but for viewing purposes it would be so large it probably wouldn't fit on your monitor. For our initial SODA collections, which consisted of mostly black-and-white printed documents, we felt that scanning 300 dpi TIFFs was adequate for our digital masters. But how would users search and retrieve them?

Having a TIFF over PDFs

While Cornell's digitizers had used a TIFF-to-GIF converter for MOA, they had also created their own navigation

system for viewing the documents. To make a long story short, this was way beyond our ability. So we decided that Adobe's Acrobat PDF would be the digital object of choice for SODA. The PDF file has the ability to incorporate a series of scanned images into one file. The Acrobat Reader was also ubiquitous in the Web world, at least when compared to readers for other image types, and this way we would only need one cataloging record per digital object.

There were arguments against using Acrobat, however. Adobe's Portable Document Format is proprietary, not an open industry standard like TIFF. And while the Acrobat viewer is currently free, what if Adobe managers changed their minds and started charging for it? Indeed, when you download the Acrobat viewer, you are already subjected to Adobe's marketing via little ads in the upper right-hand corner and prompts to download other software. Also, relying on Acrobat puts the navigation of the digital object in the hands of Adobe, unlike the navigation of the MOA project, which is completely in the hands of people at Cornell and the University of Michigan.

Despite these problems, it looked like Adobe PDF would be the best way to serve our users. It adequately displays the image of the original document, you can zoom in and out to re-size

the text to your desired viewing scale, you can search within the document, and you can easily bookmark and link to chapters. (However, for preservation purposes, we keep a copy of every page scanned as a TIFF tucked away on CD and tape.)

Back when I first learned we were going to be involved in a digitization project, I had bought a cheap scanner and a copy of Adobe Acrobat 5.0 and started experimenting. I was surprised to find that I couldn't search my scanned documents. OK, maybe it's obvious to you that an image has no "text" in it for searching, but I was truly mystified at the time. To actually be able to search the "text" of a PDF file, it first had to be subjected to Optical Character Recognition (OCR). The PDF file then becomes an image with "embedded text." This opened a whole new world in discovering how to scan our documents for full-text searching capability.

After a little poking around, I found that Acrobat 5.0 had a plug-in I could download, called "Paper Capture," for performing OCR on documents of less than 50 pages. When running this plug-in it would create the "embedded text" necessary to allow searching in the PDF file. However, it was only adequate for small documents. Some of the materials we intended to digitize had more than 400 pages. Later, we were introduced to a product that would solve that problem.

Finding a Vendor to Help Bottle the Data for SODA

In addition to our scanning research, we were also studying how to make the scanned objects available on the Web. By now it was obvious that we were not going to create our own database from scratch, so who would we turn to? We looked at various vendors' products, but none offered the full-text searching capability that we wanted. Endeavor's EN-Compass, Innovative Interfaces' imaging product ContentDM, and other open

Software We Used for SODA

ArchivalWare, for entering and maintaining metadata and searching
<http://www.ptfs.com>

InputAccel, scanning and image quality control
<http://www.captivasoftware.com>

Prime Recognition, OCR software that generates PDFs
<http://www.primerecognition.com>

Adobe Acrobat, for creating bookmarks and editing PDF files

Adobe Photoshop, for tweaking pictures
<http://www.adobe.com>



architecture products—some that don't even exist anymore—all offered "solutions," but none that seemed to have the capability of full-text searching without a great deal of effort on our part.

For example, ContentDM is a great product for small collections, but it only indexes the metadata you add to it. It will not search the contents of a collection of PDF files. To have a searchable, full-text database using ContentDM, we would have had to scan every page and assign the full text of that page to a metadata field. This is similar to what Cornell and Michigan did in their MOA project, but beyond what we were able to do with our resources.

Around this time, we heard about a company called Progressive Technology Federal Systems, Inc. (PTFS), which offered a product called ArchivalWare. It had a demo of its product up on the Web and it was exciting! Not only could it search the contents of PDF files, but a user could quickly scan the retrieved documents by her search words via Adobe's navigation buttons. ArchivalWare bridges the gap between database searching and PDF navigation. It offers the ability to search by a variety of fields and has full-text searching functions.

"I HAD TIME TO HANG
OUT WITH PEOPLE WHO
KNEW THE INS AND OUTS
OF SCANNING AND TO
ASK THEM LOTS OF
ANNOYING QUESTIONS."

For example, in addition to using Boolean logic, you can also select Pattern Searching or Concept Searching. You can set the Pattern Searching op-

Our vendor, PTFS, was able to customize ArchivalWare to include certain fields that we wanted.

tion to look for a maximum of 100 ways of spelling your search terms. You can set the Concept Searching option to look for strong synonyms, strong antonyms, and a variety of other words. Where did these kinds of searching options come from? ArchivalWare is actually PTFS's friendly repackaging of Convera's RetrievalWare, a "search and categorization solution." Convera has developed an extremely complex product designed to allow businesses to intelligently perform data mining on their documents.

Besides ArchivalWare meeting our searching needs, PTFS could also help us with our entire digitization project. Its employees recommended and configured our scanners, interface cards, server, workstations, imaging software, and even a backup system. They provided a week of training in scanning, entering data, system administration, setting up our scanning process with a product called InputAccel, and ongoing support and maintenance. In the end, they were extremely flexible in designing ArchivalWare to search and display a variety of the fields of our choosing.

They also introduced us to several scanning products we hadn't heard about.

PDFs Were 'Flat' Files, Then OCR Made Text 'Pop'

PTFS showed us a product called PrimeOCR, created by Prime Recognition, that would both create a PDF file and perform the OCR with a higher degree of accuracy than any existing product. With PrimeOCR, an image is subjected to more than one OCR engine, thus ensuring that the text is correctly identified.

PrimeOCR contains "voting" technology. For example, if two engines say a letter is an "a" and one says it's an "e," the "a" wins! You can buy the product with between three and five OCR engines; the more engines, the more accuracy you get (and the more you pay). In fact, Prime Recognition claims that OCR errors are reduced by 65 percent with three engines and by 80 percent with five. PrimeOCR is used by LexisNexis, H.W. Wilson, Boeing, Xerox, the U.S. Department of Defense, MIT, and other "pros" involved in dig-

itizing paper documents. PrimeOCR is not cheap. We paid an initial license fee of \$11,782 and the ongoing maintenance is approximately \$1,800 per year. However, when you compare that cost to hiring a library technician to perform quality control, the price doesn't look that bad. One other piece of advice here: Expensive scanners are worth it. Good scans give good OCR results.

Here's an experiment you can try: To view the text that we put through OCR, go to our Web site (<http://soda.sou.edu>), perform a search, and bring up a PDF. Use Acrobat's Text Select tool to highlight some text. Copy and paste it into a text editor and view the results. Almost every time we have done this we have found very few errors.

WE MADE MANY OF OUR PROJECT CHOICES BASED ON MY OWN TECHNICAL SKILLS, OR THE LACK THEREOF.

Employees' Knowledge Made All the Difference

Our project research began in July 2001 and SODA went live to the public in September 2002. In a project like this there are a number of ways to do things, and the reasons we chose to do things in certain ways were based largely on my own technical skills, or the lack thereof. I had to learn Oracle (which is used by ArchivalWare), the Windows 2000 Server operating system, and a little bit of Java, plus, of course, I used all the knowledge I already had for creating Web pages: HTML, JavaScript, and CSS. You can find more technical info on the SODA site at <http://soda.sou.edu/technical.html>.

I had to create an Access database to track all our documents. I entered the title and assigned a tracking number to each document (along with other notes). We could then print out the information on a cover sheet and attach it to the document, updating the database as it went through the process.

This is the document tracking sheet we designed to keep tabs on each document we processed.

One of the best personnel decisions we made for SODA was hiring our digitization assistant, Tatiana Fox. She worked as a student in the library for years and was extremely proficient at using technology. While she had little experience scanning, I knew that she paid attention to detail and would be perfect for overseeing the students who would actually scan the documents. Tatiana oversaw the work flow and assured all the pages of our documents wound up in the right place.

Finally, I cannot overstate the importance of Deb and Mary Jane in collecting these materials. Without them, the documents would merely be bits of data hoping to connect with nameless researchers. Their work ensured that our scanned objects truly served the patrons they were designed to serve. (They explained their efforts in an article if you'd like to know more.)³

Meeting Our Digitization Goals Was Sweet Success

We had begun our project with a great deal of anxiety over success. We had seen other databases be produced but then die for lack of expertise and un-

derstanding in how to maintain them. While we plan to continue using the software systems we have in place, if need be we can transfer everything over to any other kind of system currently out there. All our scanned images exist as TIFFs and all metadata is in XML. This gives us a degree of security knowing that our project, at its core, is in compliance with existing standards.

We are currently looking to develop cooperative arrangements with agencies producing documents that are appropriate for our Bioregion Collection. The SOU campus is fortunate to be home to the Klamath Network Inventory and Monitoring Program, one of 32 networks set up by the U.S. National Park Service to monitor the parks' natural resources. We are currently exploring ways to work with them and to possibly load the raw data from their numerous inventories (as Excel spreadsheets), as well as other reports. We are also exploring grant funding from a variety of sources to digitize some of the local history materials we have in our library, and to possibly coordinate with local history groups to digitize older materials in their collections.

PTFS will soon be updating ArchivalWare to take advantage of the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). Compliance with OAI-PMH is important to us because we want to be able to share our metadata with other digital projects, such as the National Science Digital Library (<http://www.nsd.org>). Why keep our SODA bottled up when others could be drinking it in?

We want to further explore the full-text searching capabilities of ArchivalWare by possibly incorporating our own thesaurus. There is a whole world of full-text searching techniques that we have yet to explore and a world of content management vendors that have a lot to offer libraries.

So that's our digitization project in a nutshell. The primary goal for SODA was to have a full-text, completely searchable



To Contact the Companies

Convera

<http://www.convera.com>

Endeavor ENCompass

<http://encompass.endinfosys.com>

ContentDM


<http://contentdm.com>

Innovative Interfaces, Inc.

<http://www.iii.com>

PTFS (Progressive Technology Federal Systems)

<http://www.ptfs.com>

database of hundreds of documents taken from our own collection, and we've tasted success. Other goals were to capture related transient documents we didn't own, to create a digitization infrastructure at SOU, and to serve as a model for regional libraries. Looking back, I would say that we met all our goals, except for being a model. We won't know if we have succeeded in that until we get more feedback. I encourage you to try out SODA's full-text searching capabilities and let us know what you think. We welcome all comments. 

Jim Rible is the systems librarian for the Lenn and Dixie Hannon Library at Southern Oregon University in Ashland. He has been involved in library technology for 20 years, including a

stint as the SOU campus Webmaster. His e-mail address is rible@sou.edu.

References

1. Making of America. <http://www.hti.umich.edu/m/oaagrp> and <http://cdl.library.cornell.edu/mao>
2. Kenney, Anne R. and Rieger, Oya Y. *Moving Theory into Practice: Digital Imaging for Libraries and Archives*. Mountain View, Calif.: Research Libraries Group, 2000.
3. Hollens, Deborah and Mary Jane Cedar Face. "A Digital Library to Serve a Region: the Bioregion and First Nations Collections of the Southern Oregon Digital Archives," *Reference & User Services Quarterly*, vol. 44, no. 2 (Winter, 2005), 116-121.

Further Reading

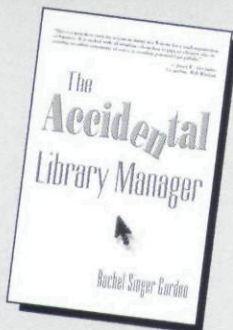
Colorado Digitization Project: <http://www.cdphheritage.org/resource>
Dublin Core Metadata Initiative: <http://dublincore.org>
Moving Theory into Practice: Digital Imaging Tutorial: <http://www.library.cornell.edu/preservation/tutorial/contents.html>
RLG DigiNews (Web-based newsletter): <http://www.rlg.org/preserv/diginews>

THE "ACCIDENTAL" SERIES

THE ACCIDENTAL LIBRARY MANAGER

A must-have guide for any librarian who wishes to succeed in management.

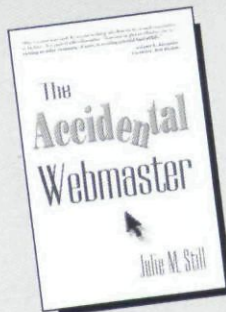
By Rachel Singer Gordon
368 pages/ISBN: 1-57387-210-5
\$29.50



THE ACCIDENTAL WEBMASTER

A lifeline for the individual not trained as a Webmaster.

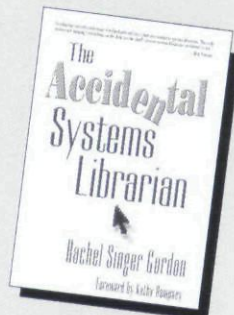
By Julie M. Still
208 pages/ISBN: 1-57387-164-8
\$29.50



THE ACCIDENTAL SYSTEMS LIBRARIAN

Essential for any librarian who wants to deal more effectively with technology.

By Rachel Singer Gordon
220 pages/ISBN: 1-57387-161-3
\$29.50



**Visit your local
bookstore or order
direct from the
publisher.**

www.infotoday.com

For more information, call 800-300-9868; outside the U.S. call 609-654-6266
Write to: Information Today, Inc., 143 Old Marlton Pike, Medford, NJ 08055
Order online at www.infotoday.com or e-mail: custserv@infotoday.com

Copyright of Computers in Libraries is the property of Information Today Inc. and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.