digitalnewspapers.org:

The Digital Newspapers Program at the University of Utah

By
John Herbert[i]
and
Kenning Arlitsch[ii]

University of Utah
Marriott Library

Abstract

This article describes the Utah Digital Newspapers Program at the University of Utah's Marriott Library. Background information regarding the historical importance of newspapers, the current state of commercial newspaper digitization and the problems with small newspaper digitization are reviewed, and the solution provided by the University of Utah is offered. Details are provided for the program's history, funding, goals, and future plans. Other topics covered include the trade-offs between microfilm and hard copy source materials, how titles were selected, the processes and technologies utilized, website organization, displaying PDF files with Macintosh computers, and using bi-tonal or grayscale images.

[i] John Herbert – Academic Degrees: BA Computer Science and Mathematics, University of Kansas, 1978; MBA, University of Utah, 1990. Current Title: Project Director – Utah Digital Newspapers Program. Street Address: University of Utah – Marriott Library, 295 S. 1500 East, Room 418, Salt Lake City, UT 84112. Email Address: john.herbert@library.utah.edu

[ii] Kenning Arlitsch – Academic Degrees: AAS, Business Administration, Suffolk County Community College, 1985; BA English, Alfred University, 1987; MLIS, University of Wisconsin-Milwaukee, 1993. Current Title: Head of Digital Technologies. Street Address: University of Utah – Marriott Library, 295 S. 1500 East, Room 463, Salt Lake City, UT 84112. Email Address: kenning.arlitsch@library.utah.edu

digitalnewspapers.org:

The Digital Newspapers Program at the University of Utah

Background

Newspapers are a primary source of historical information and are useful to scholarly researchers and laypeople alike. They are a major publishing source, representing a wide variety of national, regional, and local voices. Historical newspapers are immensely popular, particularly with genealogists and historians, but until now they have only been accessible in central locations to those diligent enough to browse reels of microfilm or read aged originals. These collections generally are organized by date, and indexes, which are rare, usually aren't standardized or searchable in an expedient way. With time and effort, these research methods can be effective, but they are never efficient. Traditional search methods simply are too slow and difficult to create a broad audience for newspapers research. However, the advent of digitized newspapers available on the Internet has fundamentally changed this paradigm. Now the possibility exists for newspaper research to be performed by practically anyone with an Internet-connected PC.

There are digital collections of large daily newspapers available online today, but they usually hold only the most recent issues whose original form was digital. The recently completed digital collections of the *New York Times* and *Wall Street Journal* in the *ProQuest Historical Newspapers* product has brought easy, though fee-based, archival access to two of the nation's most important historical resources. Discussions with ProQuest have confirmed that small regional newspapers are unlikely to be digitized by commercial firms because of their low financial return. This creates an opportunity for cultural heritage institutions because regional newspapers contain enormous amounts of local news and information, as well as unique perspectives on larger, national issues. They reflect local conditions and are the only place where local issues are covered.

Without any financial incentive for the private sector to pursue newspaper digitization on a broad scale, it falls to the public sector to move things forward. Numerous non-commercial attempts at digitizing newspapers have been made over the past ten years, mostly at academic institutions around the world[1]. In most efforts processing and storage costs and the lack of good optical character recognition (OCR) technology have outweighed the achievements.

Today, within the public sector, there are two major newspaper digitization processes, one of which has been developed by Olive Software Inc. The Olive software suite presents an attractive interface, has some successful implementations[2], and most likely will remain an alternative for digital newspaper projects. Olive is available from the Online Computer Library Center (OCLC)[3].

The second digitization process, recently developed and implemented by the University of Utah (U of U), also represents a viable digitization alternative for cultural heritage institutions nationwide. While we did examine Olive, our choice for the technology platform, CONTENTdm, delivers high quality images from microfilm and print, a robust database, and a powerful search engine, all at a relatively low cost[4]. The cost makes it affordable so that small institutions can launch a newspaper project of their own, and their stepped pricing enables collections to be funded incrementally. CONTENTdm supports not just newspapers but many different digital formats, and presents a unified interface for all digital collections at a given institution. In addition, distributed collections can have their metadata harvested and centrally aggregated to present to the user what appears to be a single collection for searching. Finally, it allows users to view newspapers with multiple Web browsers as described below. CONTENTdm is also available from OCLC.

The process and technology we have developed are working well with the volumes we have processed to date, and we expect them to continue to scale well as we further expand the collection. We hope that with the broad implementation of this process, a new generation of "citizen

historians" will utilize our website and others like it as their own in-home research tool. After a decade of research and development, the opportunity for broad-based online newspaper research is finally here.

History of the Utah Program

In the 1980's the Marriott Library at the U of U received a grant from the National Endowment of the Humanities (NEH) *United States Newspapers Program* to collect, catalog, and microfilm Utah newspapers. The project was successful and identified over 920 newspapers[5] in the state. More than a decade later the idea of digitizing the newspapers arose, prompted in part by inquiries from community libraries in Utah that were responding to increased use of their newspaper collections. In addition, increased expertise with other digital formats at the U of U and Brigham Young University (BYU) triggered innovative thinking that suggested newspaper digitization was not only possible but also had definite and realistic potential.

2001-2002

In 2001, the Marriott Library received a $93,000 Library Services and Technology Act (LSTA) grant to digitize 30 years of three weekly newspapers. The two main goals of the grant were: 1) develop a newspaper digitization process employing existing digital collection management technologies; and 2) post a significant amount of content on a website. Issues of cost, server space, file format, and newspaper presentation mechanisms also were to be addressed during this year-long project.

The 2001-2002 project resulted in a new Library website comprising 30,000 pages from three rural Utah weekly newspapers: the *Wasatch Wave* from Heber City, the *Times Independent* from Moab, and the *Vernal Express* from Vernal.[6]

This initial collection was launched in December, 2002, and word of the new website quickly spread through the state's library community. Librarians from the three newspapers' communities wrote to say how delighted they were with the improved access and asked when they could see more of their newspapers digitized. Other communities wrote to say they had found original hard copies of their local newspapers. (Details of the hard copy collections that were digitized in 2003 are outlined below.) Based on this information, we estimated that there are 3-5 million pages of historical newspapers in the state that await digitization. More importantly, we quickly realized that there is a tremendous demand for digital newspapers in Utah and a high need for our collection to be expanded.

2003

In November, 2002, the Utah Academic Library Consortium (UALC[7]), led by the Marriott Library, was awarded a $278,000 LSTA grant to continue digitizing newspapers. Cooperative support was unprecedented with nearly $100,000 in matching funds contributed by the UALC, Weber County Library, Murray City Library, the Uintah County Library, and the Grand County Library.

This phase of the project, which was launched in January, 2003 and completed in September, 2003, had several goals, each of which was completed successfully:

- A project director was hired to run the day-to-day operation of the program and secure additional funding for the ongoing success of the program.
- 100,000 pages of historical Utah newspapers were digitized, effectively quadrupling the size of the collection.
- A publicity campaign was planned and implemented to insure broad knowledge of the program.

- Relationships were established with public libraries and current newspaper publishers throughout the state.
- Options were explored for including digital issues of current Utah weekly newspapers.

Microfilm vs. Hard Copy

The 2001-2002 project taught us that digitizing newspapers from microfilm was not always the best method. Early microfilming techniques sometimes created film with poorly focused images, uneven lighting, and dark spots. Poor film, in turn, produced poor digital images and usually decreased OCR accuracy. While most of the film at the University of Utah was of acceptable quality, some was not. This potential ongoing issue with poor film prompted us to consider scanning from original hard copies for the 2003 phase of the project.

Paper has its own problems, of course. First, original collections can be difficult to find. As Nicholson Baker pointed out several years ago[8], there is no national policy to retain newspapers in their original format, and as a result, complete historical runs are increasingly rare. If they do exist, it is largely because of local efforts of public libraries, historical societies, publishers and private citizens. Even within these groups, collection strategy is spotty and archival storage techniques are often non-existent.

Second, newsprint is notorious for its high acid content and over time can become yellow and brittle. The discoloration of the paper, of course, reduces the quality of the photographic image. Handling brittle paper for the scanning process, if not done carefully, can damage the material. But newsprint that is properly stored (i.e., shielded from air, light, humidity and heat) can remain in surprisingly good condition for decades. Bound volumes rarely opened and handled can look as good as when they were published more than a hundred years ago.

Third, newsprint is generally oversized and difficult to scan on standard equipment. While the Digital Technologies division at the Marriott Library owns several large flatbed scanners, they are not big enough to accommodate most newspapers. Flatbed scanners even of sufficient size can present other problems. Most newspapers are bound into books which cannot be opened wide enough to scan effectively on a flatbed. These bindings, especially of older volumes, are frequently too fragile to allow the handling (i.e., being turned upside-down and flattened) required with a flatbed. The Library owns a Leica S1 digital scanning camera, which has been a workhorse for over three years in scanning rare and fragile materials from a copy stand. But the Leica is an older generation scanner and scanning each newspaper page takes approximately three minutes – far too slow for the volumes we were considering. Recently the library purchased a Zeutschel 10000 book scanner that will be used to digitize rare books and newspapers in the next phase of our program. The Zeutschel has an overhead camera with a mechanized book cradle that scans two pages of newspaper in approximately five seconds at 300ppi, or in ten seconds at 600ppi.

All this notwithstanding, we were fortunate enough to uncover some substantial runs of original Utah papers. The Weber County Library recently acquired 77 years of the *Ogden Standard Examiner* and its predecessors (1879-1955) from the publisher. The Emery County Archives in Castle Dale held both loose and bound issues of the *Emery County Progress*. The publisher of the *Millard County Chronicle* had stored a full decade (1930-39) under tables in their office in Delta. A librarian at Dixie State College in St. George directed us to a complete run of the now-defunct *Washington County News* (1908-1988). Finally, a private citizen in Green River produced the only known copy of the *Green River Journal*, which was published for a short period (1955-56) during Utah's uranium boom.

In total, over 68,000 pages of the 106,000 digitized this year came from paper. They were scanned by two Provo-based contractors named e•prep and Access Imagery. e•prep did the bulk of the work

because they developed a high-speed process for large bound volumes. They used Hasselblad cameras with Imacon digital backs and were able to produce up to 2,000 images per day at an affordable price, discounted for our large volumes at thirty cents per page.

It should be noted that the original hard copies sometimes required repair work before they could be scanned. The Marriott Library's Preservation staff performed this function for the project. Most often, for the sake of time and lack of funding, they did not perform a complete restoration of the material, but rather repaired and stabilized it enough so that it could be scanned without additional damage.

How Titles Were Selected for Digitization

Given the statewide demand to have many newspapers added during 2003, the project team solicited input from several noted Utah historians in making the major selection decisions. Our 2002 experience showed us that adding any of the large daily newspapers from Salt Lake County would quickly exhaust our project funds because of their high page volume. So our guideline to the historians was that we were most interested in which of Utah's 29 counties, other than Salt Lake County and the three counties added in 2002, would have the most historical significance if added to our collection. When we tallied the results, the top five counties were Carbon, Summit, Tooele, Juab, and Sanpete. We examined the Marriott Library's catalog to see which early newspaper titles were available on microfilm, and several titles were selected.

- Carbon County – the *Eastern Utah Telegraph*, the *Eastern Utah Advocate*, the *Carbon County News*, and the *News-Advocate*.
- Summit County – the *Park Record*.
- Tooele County – the *Tooele County Chronicle*.
- Juab County – the *Eureka Reporter*.

- Sanpete County – the *Manti Messenger*.

We found original copies of three of these papers (the *Park Record,* the *Tooele County Chronicle,* and the *Eureka Reporter)* and scanned these originals rather than their microfilm.

We also added titles from the communities of the matching gifts we received.

- Weber County Library – four early predecessors of the *Ogden Standard Examiner*.
- Uintah County Library – the *Vernal Express*.
- Murray City Library – the *American Eagle* and the *Murray Eagle*.
- Grand County Library – the *Times Independent*.

Finally, during our travels around the state, we found several collections of unique historical newspapers that we decided should be added. As noted above, these titles were: the *Emery County Progress*, the *Green River Journal*, the *Millard County Chronicle*, and the *Washington County News*.

The Digitization Process

The process to create the digitized content involves two main commercial partners: iArchives Inc. of Lindon, Utah, and DiMeMa Inc., a University of Washington spin-off in Seattle. iArchives provides the zoning, metadata, and OCR processing outlined below. In addition, they scan microfilm and sub-contract the scanning of original hard copies to e•prep and Access Imagery. The company also has an open-architecture, XML-tagging process used for newspapers.

DiMeMa provides the CONTENTdm Digital Collection Management Software that has been used with other digital collections at the Marriott Library for several years. It manages practically all types of media: books, maps[9], photographs, documents, 3D objects, streaming media, and, as a result of this partnership with the U of U, newspaper collections. Its architecture includes server software

that runs on Windows, Linux, and UNIX Solaris operating systems, and workstation acquisition software that may be used within the same institution or remotely.

The digitization process involves seven high-level steps that are performed by these different groups, each one of which has its own Quality Assurance check.

1. SCANNING. The source materials, either original hard copy or microfilm, are digitally scanned at a minimum of 300 dpi. (400 dpi is recommended.) Images are cropped, de-skewed, de-speckled, and split if necessary. The result is a set of multi-functional tiff and jpeg images, some of which are used only in internal processes. The full-page tiff images are copied and shipped to the U of U, where they are kept as an archive of what was processed.
2. ZONING. Individual articles are zoned into separate images so they can later be viewed separately from the entire page. Articles are classified by type: news, births, deaths, marriages, advertisements, etc., by examining their headlines.
3. METADATA. Issue metadata are keyed, as well as masthead information and article headlines. All initial keying is verified by re-keying.
4. OCR. Sophisticated optical character recognition software, patented by iArchives, is run on the image to generate full text for each article.
5. PDF. The text output from the OCR is combined with the image to form a PDF file of each page and article, with the full text embedded in the images. XML-tagged metadata files for the pages and articles are also generated.
6. IMPORT. The PDF and XML files are batch imported into CONTENTdm using software developed by DiMeMa. XML wrapper files are automatically generated to tie together the pages of discrete newspaper issues. Because of the sheer volume of data associated with our newspaper collection, DiMeMa is currently providing this database importing service to us.

7. WEBSITE SERVER. The resulting CONTENTdm files are loaded onto the U of U server and are linked to the website by the Marriott Library staff.

Website Organization

The newspaper issues loaded into the digitalnewspapers.org website are fully searchable by keyword and by date, article title, and article type. Individual newspapers may also be browsed by year and issue date. Both the search and browse methods result in the articles or full pages shown in context to the other articles and pages of the issue. Because the PDF images contain hidden text, secondary searching is available from the Acrobat toolbar, and found words are highlighted.

The main page of the website offers entry into the site's features along with rotating newspaper front pages that we hope are intriguing to the reader (See Figure 1).

*** INSERT FIGURE 1 HERE ***

Here users have three main options for viewing the collection. The first will activate a keyword search across the entire newspaper collection with the additional option of an exact-phrase search. Results are displayed on a separate search-results page (See Figure 2).

*** INSERT FIGURE 2 HERE ***

Selecting a specific article from the search results opens a window displaying the PDF image of the article on the right side, along with the entire issue's table of contents on the left side (See Figure 3). By clicking on entries in the table of contents, users can easily browse through the entire issue, looking at whole pages or individual articles.

*** INSERT FIGURE 3 HERE ***

The second option from the main webpage is to pick a particular newspaper title from the pull-down menu. Selecting a title links to a secondary page for that newspaper. Here users see general

information about the paper and have the ability to select a specific issue from the "Browse" box, or search across all of that paper's issues from the "Search" box (See Figure 4).

To use the "Browse" box, users first select a year from the "Please Select" pull-down menu. When that is done, the dates of the issues from that year are displayed in the "Select an Issue" pull-down menu. Clicking a specific date opens a separate window displaying that issue's table of contents in the left frame and the PDF image of the front page in the right frame. Here, as described above, users can easily browse through the entire issue by clicking on entries in the table of contents.

*** INSERT FIGURE 4 HERE ***

The "Search" box provides several useful keyword searching options: within articles classified as births, marriages, or deaths; within article headlines; or across all issues of that newspaper title. Each keyword search also has the option of an exact-phrase search for multiple keywords. Another feature is clicking on the "Go" button next to the births, marriages, or deaths while leaving the keyword search box empty returns a list of all births, marriages, or death notices classified for that newspaper. (Note: birth, marriage, and death notice classifications are made by examining headlines. These notices sometimes exist in an issue but are not labeled as such by a headline. In this case, they may not be classified as a birth, marriage, or death notice and may not be retrieved through this search.)

The third option from the main webpage is to browse a state map that shows which counties have newspapers included in the collection. Moving the mouse over a county will display the titles/years from that county which are included in the collection (See Figure 5). Clicking on a title links to that paper's secondary page.

*** INSERT FIGURE 5 HERE ***

PDF and Macintosh

As described in the previous section, PDF images of the newspapers are presented in a frameset with text links to the other pages and articles of the issue in the left-hand frame, the image on the right, and additional banners and navigation across the top. Microsoft Windows users display this environment with a standard Adobe Acrobat Reader plug-in to their Web browsers. But the Apple Macintosh environment differs significantly in that no PDF plug-in is currently available for Mac browsers. Instead, when accessing PDF files, Mac users typically download them into the *stand-alone* Acrobat Reader, or other PDF-capable software, and view the image file separately. In the previous version of CONTENTdm, their viewer did not automatically determine the user's platform. Consequently, successfully downloading PDF files was dependent on the Mac browser recognizing that the file had to be displayed with separate software. The new CONTENTdm viewer solves this problem by recognizing the user's platform as Macintosh and offering the PDF file as a separate download. Until PDF plug-in software becomes available again for Mac users, this is the best solution, but users miss the context of the newspaper issue provided by the frameset.

Microsoft's June, 2003 announcement that they are halting development on the Macintosh version of Internet Explorer[10] also makes future Macintosh support difficult because the number of browsers vying to fill the gap left by Microsoft will continue to increase. One unofficial directory on the Web offers links to no fewer than 35 browsers currently available for Macintosh users.[11] Our own Web server statistics show that in 2003, 87% of visitors to the Marriott Library's website used some form of Windows and IE. Other sources put IE users as high as 94%.[12]

1-Bit vs. 4-Bit Image Files

Almost all the PDF files in the Utah Digital Newspapers website are displayed as 1-bit, black-and-white images; the pixels are either black or white with no gray tones. The benefit of this bi-tonal

approach is that the image files are relatively small and take far less time to format on screen. For instance, a full page of the *Emery County Progress* is approximately 325KB. With older newspapers, bi-tonal presentation is usually sufficient because there are few, if any, photographs. But in more recent newspapers, photographs increasingly become an important part of the publication. Early newspaper photographs are printed with a spectrum of gray tones, including black and white, and are best presented on screen as grayscale images, which use either four or eight bits to indicate 16 or 256 gray tones, respectively, for each pixel. Bi-tonal images represent photographs poorly, making them appear rather like cartoons.

The *Green River Journal* (1955-56) is the most recent newspaper we have digitized to date, and contains many more photographs than the other newspapers. Since it is also the smallest collection (only 31 issues), we have experimented by displaying it in 4-bit grayscale images. The result is much improved image quality over 1-bit images, but the size of a full page image has increased to nearly 2MB. This is a six-fold increase in file size and, if applied to our entire collection, represents a serious file storage problem for us and an equally serious downloading problem for users on narrow-band Internet connections. Since it will be a number of years before the majority of our users have broadband access, additional experiments with compression methods are needed to find the right balance between good image quality and relatively small file size.

The Future

As productive and rewarding as the last two years have been, the future of our newspaper program looks even brighter. As this article is being written, we have just received a two-year grant from the Institute of Museum and Library Services (IMLS), a federal granting agency under the Department of Health and Human Services, to continue our work. The total funding provided from the grant is

just over $1 million, with IMLS awarding $470,000 and the U of U and BYU providing matching funds of $450,000 and $100,000 respectively.

This grant represents a major step forward, allowing us to continue to develop and expand the program for another two years. Our goals for this grant are to:

- Expand our collection significantly by adding nearly a quarter-million pages of content.
- Expand our technology platform to house portions of the collection on servers at other members of the Mountain West Digital Library (MWDL[13]): BYU, Utah State University, and Southern Utah University.
- Harvest metadata from the four sites onto the existing MWDL server at the U of U by running the CONTENTdm Multi-Site Server software.
- Present a combined collection to our readers so they can perform searches on the entire collection at once, regardless of where the data's server is located.
- Administer a training program to other academic and historical institutions, with attendees receiving information on launching a digital newspapers program of their own, including how to manage the digitization process and how to write compelling grant proposals.

This dissemination and training goes to the heart of our long-term vision: propagating our program into many other similar programs in the West and, indeed, across the country. Should we be fortunate enough to have this come to pass, we can then say that broad-based online newspaper research has finally arrived.

---

[1] Richard Entlich, "Where are they now? Digitizing Microfilmed Newspapers." *RLG DigiNews* 6 no. 3, (2002).

[2] Three examples of Olive installations are:

Brooklyn Daily Eagle, 1841-1902. <http://eagle.brooklynpubliclibrary.org>.

The British Library Online Newspaper Archive. <http://www.uk.olivesoftware.com>.

The Ithacan; Ithaca College Library. <http://dpr.oclc.org/Archive/skins/Ithaca/navigator.asp?AW=1063646285593>.

[3] Marilyn Deegan, "Digitizing Historical Newspapers: Progress and Prospects." *RLG DigiNews* 6 no. 4, (2002).

[4] CONTENTdm offers different licenses and stepped pricing based on the volumes that will be loaded into the software. The lowest level of license accommodates up to 8,000 images and is priced at $6,000. See <http://contentdm.com/products/pricing.html>

[5] Robert P. Holley. *The Utah Newspaper Project Final Report, Project Number PS-200010-85*. (Salt Lake City: University of Utah Libraries, 1987). p. 5. <http://content.lib.utah.edu/cgi-bin/docviewer.exe?CISOROOT=/rare_books&CISOPTR=1624>.

[6] Kenning Arlitsch, Lawrence Yapp, and Karen Edge, "The Utah Digital Newspapers Project." *D-Lib Magazine* 9, no. 3 (2003).

[7] The Utah Academic Library Consortium (UALC) consists of all academic libraries in accredited institutions of higher education in Utah. Its purpose is to continually improve the availability and delivery of library and information services to the higher education community in Utah and to partner libraries in Nevada. For more information see <http://www.ualc.net>.

[8] Nicholson Baker, *Doublefold: Libraries and the Assault on Paper* (New York: Random House, 2001).

[9] Kenning Arlitsch, "Digitizing Sanborn Fire Insurance Maps for a Full Color, Publicly Accessible Collection." *D-Lib Magazine* 8, no. 7/8 (2002).

[10] Michael J. DeMaria, "FUDBusters." *Network Computing* 14 no. 14 (2003), p. 22.

[11] Darrel Knutson, Macintosh Web Browsers Past and Present. (2003) <http://darrel.knutson.com/mac/www/browsers.html>

[12] Browser Statistics. W3Schools. <http://www.w3schools.com/browsers/browsers_stats.asp>

[13] The Mountain West Digital Library is a consortium of digital collections from universities, colleges, public libraries, museums, and historical societies in Utah and Nevada. Six hosting institutions each run CONTENTdm servers supporting their own digital collections, and support partner institutions by providing scanning and hosting services. CONTENTdm's multi-site aggregating server automatically harvests metadata from the hosting institutions on a regular basis and provides the search engine. For more information, see <http://mwdl.org>.