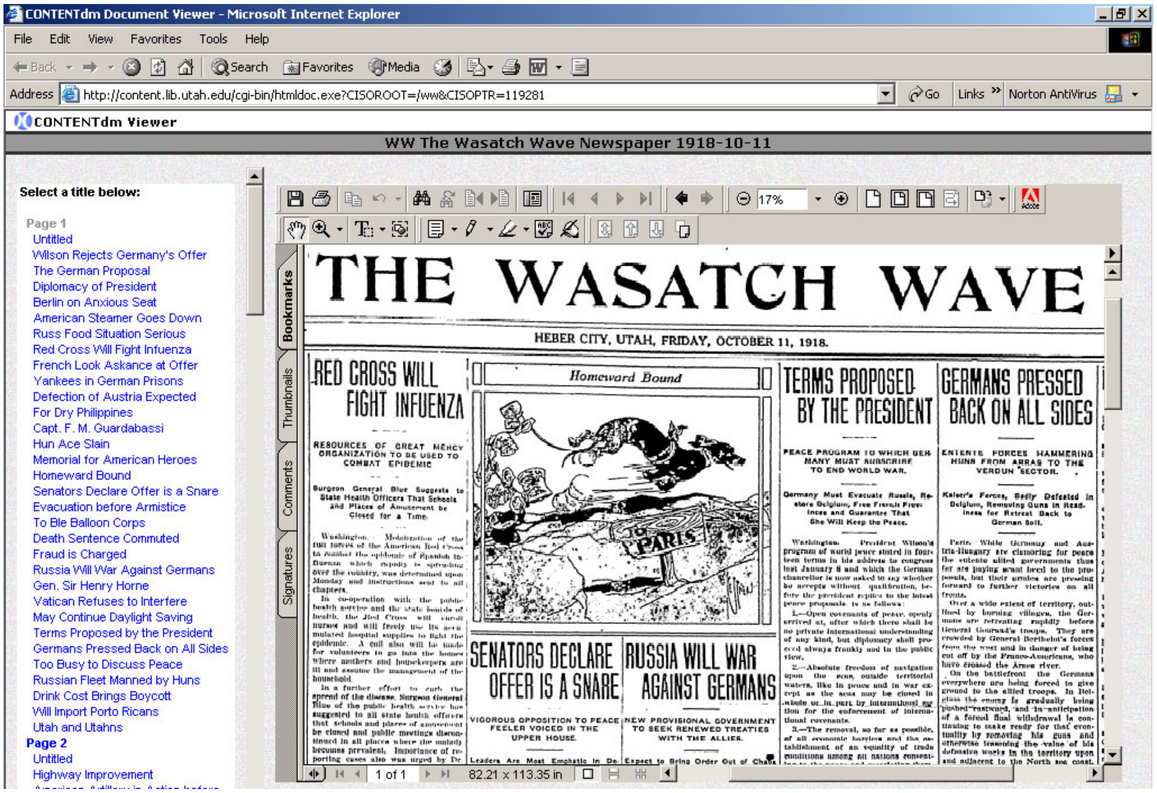
# THE UTAH DIGITAL NEWSPAPERS PROJECT

## A NEW DIGITIZATION METHOD DEVELOPED BY THE UNIVERSITY OF UTAH, IARCHIVES INC., AND DIMEMA INC

http://www.lib.utah.edu/digital/unews/



CONTENTdm Document Viewer - Microsoft Internet Explorer

Address http://content.lib.utah.edu/cgi-bin/htmldoc.exe?CISOROOT=/ww&CISOPTR=119281

CONTENTdm Viewer

WW The Wasatch Wave Newspaper 1918-10-11

**Select a title below:**

Page 1
- Untitled
- Wilson Rejects Germany's Offer
- The German Proposal
- Diplomacy of President
- Berlin on Anxious Seat
- American Steamer Goes Down
- Russ Food Situation Serious
- Red Cross Will Fight Infuenza
- French Look Askance at Offer
- Yankees in German Prisons
- Defection of Austria Expected
- For Dry Philippines
- Capt. F. M. Guardabassi
- Hun Ace Slain
- Memorial for American Heroes
- Homeward Bound
- Senators Declare Offer is a Snare
- Evacuation before Armistice
- To Ble Balloon Corps
- Death Sentence Commuted
- Fraud is Charged
- Russia Will War Against Germans
- Gen. Sir Henry Horne
- Vatican Refuses to Interfere
- May Continue Daylight Saving
- Terms Proposed by the President
- Germans Pressed Back on All Sides
- Too Busy to Discuss Peace
- Russian Fleet Manned by Huns
- Drink Cost Brings Boycott
- Will Import Porto Ricans
- Utah and Utahns

Page 2
- Untitled
- Highway Improvement

THE WASATCH WAVE

HEBER CITY, UTAH, FRIDAY, OCTOBER 11, 1918.

RED CROSS WILL FIGHT INFUENZA

RESOURCES OF GREAT MERCY ORGANIZATION TO BE USED TO COMBAT EPIDEMIC

Surgeon General Blue Suggests to State Health Officers That Schools and Places of Amusement be Closed for a Time.

Washington. — Mobilization of the full forces of the American Red Cross to combat the epidemic of Spanish influenza which rapidly is spreading over the country, was determined upon Monday and instructions sent to all chapters.

In co-operation with the public health service and the state boards of health, the Red Cross will enroll nurses and will freely use its accumulated hospital supplies to fight the epidemic. A call also will be made for volunteers to go into the homes where mothers and housekeepers are ill and assume the management of the household.

In a further effort to curb the spread of the disease, Surgeon General Blue of the public health service has suggested to all state health officers that schools and places of amusement be closed and public meetings discontinued in all places where the malady becomes prevalent. Importance of reporting cases also was urged by Dr.

Homeward Bound

TO PARIS

SENATORS DECLARE OFFER IS A SNARE

VIGOROUS OPPOSITION TO PEACE FEELER VOICED IN THE UPPER HOUSE.

Leaders Are Most Emphatic in De-

RUSSIA WILL WAR AGAINST GERMANS

NEW PROVISIONAL GOVERNMENT TO SEEK RENEWED TREATIES WITH THE ALLIES.

Expect to Bring Order Out of Chaos

TERMS PROPOSED BY THE PRESIDENT

PEACE PROGRAM TO WHICH GERMANY MUST SUBSCRIBE TO END WORLD WAR.

Germany Must Evacuate Russia, Restore Belgium, Free French Provinces and Guarantee That She Will Keep the Peace.

Washington. — President Wilson's program of world peace stated in fourteen terms in his address to congress last January 8 and which the German chancellor is now asked to say whether he accepts without qualification, before the president replies to the latest peace proposals, is as follows:

1.—Open covenants of peace, openly arrived at, after which there shall be no private international understanding of any kind, but diplomacy shall proceed always frankly and in the public view.

2.—Absolute freedom of navigation upon the seas, outside territorial waters, like in peace and in war except as the seas may be closed in whole or in part by international action for the enforcement of international covenants.

3.—The removal, so far as possible, of all economic barriers and the establishment of an equality of trade conditions among all nations consent-

GERMANS PRESSED BACK ON ALL SIDES

ENTENTE FORCES HAMMERING HUNS FROM ARRAS TO THE VERDUN SECTOR.

Kaiser's Forces, Badly Defeated in Belgium, Removing Guns in Readiness for Retreat Back to German Soil.

Paris. — While Germany and Austria-Hungary are clamoring for peace the entente allied governments thus far are paying scant heed to the proposals, but their armies are pressing forward to further victories on all fronts.

Over a wide extent of territory, outlined by burning villages, the Germans are retreating rapidly before General Gouraud's troops. They are crowded by General Berthelot's forces from the west and in danger of being cut off by the Franco-Americans, who have crossed the Aisne river.

On the battlefront the Germans everywhere are being forced to give ground to the allied troops. In Belgium the enemy is gradually being pushed eastward, and in anticipation of a forced final withdrawal is continuing to make ready for that eventuality by removing his guns and otherwise lessening the value of his defensive works in the territory upon and adjacent to the North sea coast.

KENNING ARLITSCH, HEAD OF DIGITAL TECHNOLOGIES, UNIVERSITY OF UTAH
295 S. 1500 EAST, SALT LAKE CITY, UT 84102
(801) 585-3721
kenning.arlitsch@library.utah.edu

# Utah Digital Newspapers Project

## ABSTRACT

This paper will describe a new method for digitizing historic newspapers, developed by a partnership between a university and two commercial organizations. Utilizing OCR and newspaper processing technology from iArchives Inc. and the CONTENTdm digital collections software suite, the University of Utah has posted on its website 30,000 pages spanning thirty years from three weekly newspapers. The OCR and newspaper processing technology from iArchives Inc. was used to format the digital files with XML. DiMeMa Inc. expanded its CONTENTdm Software Suite to batch import these files and automatically generate XML-wrapped newspaper issues. The newspapers are publicly accessible and may be browsed by issue or searched. With the recent award of a new grant another 100,000 pages from different newspapers are slated for digitization in 2003.

## BACKGROUND

Newspapers are a primary source of historical information, and are useful to scholarly researchers and laypeople alike. Historical newspapers are immensely popular with genealogists and historians.

> *"I cannot begin to tell you how many thousands of people have used this [Vernal Express] index all over the world… I have wished so many times the other newspapers in the state were indexed."* [1]

It is the broad appeal of newspapers that garnered so much support in both the academic and public library communities, and driven this project and others. Newspapers are also one of the most difficult and inefficient research materials; indeed it may be said that newspapers are often <u>not</u> consulted by researchers simply because they are so difficult to use. Regional, historical newspapers are rarely indexed and therefore cannot be searched. They most often can be found only in microform in centralized locations. Their use is therefore limited to one user at a time in one place and to non-electronic browsing.

Numerous attempts at digitizing newspapers have been made over the past ten years.[2] In most efforts the cost of digitization and file storage, and the lack of good optical character recognition (OCR) technology, outweighed the achievements. More recent efforts have required a specialized newspaper-specific software package that does not integrate into the larger digital collections offerings.[3] In general other approaches available today are costly, lack accuracy in the OCR results, or are stand-alone newspaper solutions.

In 2001, the Marriott Library at the University of Utah was the recipient of a $93,000 Library Services and Technology Act (LSTA) grant to digitize three weekly Utah newspapers. The goals of the grant were to develop a scalable and sustainable newspaper digitization method utilizing existing digital collections presentation and management technologies, and to post a significant portion of digital newspapers on a website. Issues of cost, server space, and file format were to be addressed during this year-long project. The three newspapers chosen for digitization have existed in Utah since the late 19th century and all are still published today:

1. *Vernal Express* (Vernal, Utah)
2. *Grand Valley Times/Times Independent* (Moab, Utah)
3. *Wasatch Wave* (Heber City, Utah)

## PARTNERS

### MARRIOTT LIBRARY, UNIVERSITY OF UTAH

The J. Willard Marriott Library is the main library of the University of Utah, in Salt Lake City. It is a member of the Association of Research Libraries with holdings of nearly 3 million volumes, including over 25,000 journals in electronic and print formats.

The Library began digitizing materials in 1998 and in early 2000 purchased the CONTENTdm digital software suite to manage and serve its digital collections over the World Wide Web. The digital library soon grew to include a variety of formats including rare books, photographs, maps[4], art prints, and documents. Prior to the digital newspapers project, the Digital Technologies department digitized most materials in-house.

In 2001 the Library began to support the digital aspirations of smaller cultural heritage institutions in Utah by providing fee-based scanning services and space on its CONTENTdm server. In 2002 the Library proposed the creation of the Mountain West Digital Library (MWDL) and secured support and funding from the Utah Academic Library Consortium. The goal of the MWDL is to support the digital collections of cultural heritage institutions throughout the states of Utah and Nevada in a manner that allows those institutions to retain control and identity of their collections. CONTENTdm multi-site server technology, which harvests metadata from dispersed servers, was developed by DiMeMa Inc. and allows a single point of search across all collections. The Utah digital newspapers are expected to become an integral part of the MWDL. More information may be found at http://www.lib.utah.edu/digital/mwdl/.

### IARCHIVES INC.

iArchives was selected to perform the task of converting the microfilm into a format that could efficiently be integrated into CONTENTdm. iArchives provides the technology and processing method that substantially reduces the time and cost it takes to convert the microfilm into a highly accurate searchable database. iArchives also provides a non-proprietary solution. The University of Utah worked closely with personnel at DiMeMa and iArchives to define the specifications of the project and provide a solution that can be duplicated easily in the future.

iArchives uses state-of-the-art technology and processes to produce searchable, high-resolution images in TIFF, PDF, or other file types. iArchives has developed a highly accurate and patented OCR software methodology that consistently produces higher accuracy results when compared to the "off the shelf" OCR engines.

iArchives was founded in 1994 and operated as a service bureau utilizing its internally developed OCR technology. In the year 2000, the company began to focus all of its resources on developing leading edge technology in the areas of image enhancement and OCR recognition of dirty documents. Because iArchives had already developed its own OCR technology, the characteristics that affect OCR accuracy were studied, and technology has been developed to enhance image quality and improve OCR accuracy. iArchives has developed and patented its OCR technology to maximize the results of its own OCR, as well as other top performing OCR engines. The iArchives OCR process utilizes multiple OCR engines that build upon each other. The results of the individual engines are not voted; rather the results of all engines are retained. This allows the iArchives OCR process to provide results that are substantially better than other OCR technologies currently on the market. More information about iArchives Inc. may be found at http://www.iarchives.com .

DIMEMA INC.

DiMeMa Inc. is the developer of the CONTENTdm Software Suite for digital collections. Originally created by the Center for Information Systems Optimization (CISO) at the University of Washington, CONTENTdm was spun off to DiMeMa Inc. in 2001. CONTENTdm has become the primary software for digital special collections development in over 100 sites in the United States alone. CONTENTdm handles virtually all media types and supports open standard file types and metadata fields. The CONTENTdm Software Suite includes multiple acquisition stations for capturing and indexing items, and a Query Building Tool for developing custom web interfaces for collections access. With the 3.4 release in February 2003 CONTENTdm will also be fully compatible with the Open Archives Initiative.

In May 2002 DiMeMa Inc. signed an agreement in which OCLC became the sole distributor and marketing agent of the CONTENTdm Software Suite to libraries, museums, historical societies, and non-profit archive organizations. More information on CONTENTdm and DiMeMa is available at http://www.contentdm.com

## METHODOLOGY

Before iArchives was brought into the project, newspapers were scanned from microfilm by a local vendor, and manual indexing was anticipated. Some of the microfilm was of an extremely poor photographic quality. Newspapers were scanned at 300 dpi, and in both grayscale and bitonal bit depths because some images were unreadable as bitonal scans. After the digital files were returned to the University an attempt was made to use off-the-shelf OCR packages, but they failed miserably. Adobe Acrobat® Capture® and ABBYY FineReader (which has performed admirably in other situations) were used and then abandoned.

Two distinct methods were employed to digitize the three newspapers. After scanning, the *Vernal Express* newspaper was processed almost entirely by the University of Utah. The *Grand Valley Times* and the *Wasatch Wave* were scanned and processed by iArchives and DiMeMa, and then delivered as complete collections to the University of Utah.

1. *Vernal Express (Vernal, Utah)*

   a. **Method:** Microfilm was scanned by FutureVision Inc. An index created manually for many years by the Uintah County Library (and generously contributed to this project) was imported into CONTENTdm and merged with the images. Compound documents (XML wrapper files) were created manually using CONTENTdm tools to display all pages of the newspaper issues.

   b. **Result:** Newspaper issues may be browsed by year and date. Searching is available by article title, subject headings, and date. Pages may be viewed as quick-loading

JPEG images and larger MrSID® files may be downloaded for detailed zooming. Only local-interest stories were indexed.

2. *Wasatch Wave (Heber City, Utah)* and *Grand Valley Times (Moab, ,Utah)*
   a. **Method:** Microfilm was scanned and processed by iArchives Inc. Articles were zoned into separate files, and headlines were re-keyed. Articles were classed into the following types: articles, weddings/engagements, obituaries, and advertisements. Files were delivered to DiMeMa Inc. where they were imported into CONTENTdm and compound documents were generated automatically.
   b. **Result**: Newspaper issues may be browsed by year and date, or searched. Articles are viewed in context to the pages they were published on, and full-text, article titles, and article types are searchable. PDF images with hidden text allow secondary searching in the Acrobat Reader.

## EVALUATION

The method used to digitize the *Vernal Express* relied heavily on manual processes. The method grew out of the availability of the index provided by the Uintah County Library and because the partnership with iArchives Inc. was still being conceived. The method used to digitize the *Wasatch Wave* and the *Grand Valley Times* has proven much more efficient and effective and will be used in future digital newspaper projects at the University of Utah. The advantages and disadvantages of each method are listed below. Note that approximately 10,000 pages were digitized for each newspaper, for a total of 30,000 pages:

VERNAL EXPRESS

**Advantages**

1) Initial JPEG image loads quickly; headlines are legible.
2) High-resolution MrSID file is only loaded when requested.
3) Relatively small total number of files (29,000 files – includes thumbnail, display JPEG, and high-resolution MrSID for each newspaper page, and an XML wrapper file for each newspaper issue)

**Disadvantages**

1) No full-text searching
2) Individual articles are not zoned
3) Free MrSID plug-in or viewer is required to see detailed text
4) Larger total server space used than for the other two newspapers (21.5Gb)
5) Manual processes to create the collection were slow and inefficient

*WASATCH WAVE* AND *GRAND VALLEY TIMES*

**Advantages**

1) Full text searching
2) Zoned articles
3) PDF navigation (ubiquitous viewer)
4) Small file sizes (10Kb-50Kb for articles, 300Kb for full pages)
5) Small total server space (approximately 10Gb for each newspaper)
6) Automated processes from iArchives and DiMeMa greatly increased volume

**Disadvantages**

1) Large number of files because of the zoned articles. Each of the two newspapers collections contains approximately 255,000 files. That figure includes a thumbnail image and PDF display image for each page and article, and an XML wrapper file for each newspaper issue.

## COSTS

### Contracted Service – Newspaper scanning and processing costs

The scanning and processing costs listed below include newspaper scanning from three different formats. They also include the OCR and post-processing, as well as DiMeMa's fee for batch importing the images and metadata. While newspapers for this project were only scanned from microfilm, the next phase will include hard copy (see Future Directions section). The average total cost per newspaper page for all three formats is $1.65.

| Format | iArchives Inc. scanning/page | iArchives Inc. processing/page | DiMeMa Inc. import CONTENTdm/page | Total per page |
|---|---|---|---|---|
| Microfilm | $.15 | $1.27 | $.15 | $1.57 |
| Paper – Unbound | $.22 | $1.27 | $.15 | $1.64 |
| Paper – Bound | $.32 | $1.27 | $.15 | $1.74 |

## FUTURE DIRECTIONS

In November 2002, with the first digital newspapers project nearly complete, the University of Utah was awarded a second LSTA grant in the amount of $282,000 to continue digitizing newspapers and to expand the scope of the project. Community support for the project was unprecedented. Thirty-five percent of the total grant was raised as matching funds from the Utah Academic Library Consortium and public libraries that were enthusiastic about seeing their community newspapers on the Web.

The funded proposal will digitize 100,000 pages of an expanded selection of Utah newspapers by December 2003. It will also break new ground by including newspapers scanned from hard copy instead of only microfilm. The hard copies are difficult to find but several runs have been located and we expect better images and more accurate searchable text as a result.

The new grant includes funding for a Project Director whose job it will be to write additional grant proposals to keep the project moving into 2004 and beyond. The Project Director will also be responsible for developing and implementing a plan to address the following issues:

1. **Newspaper collection** – Are hard copies of all currently published newspapers in the state being collected? Who is collecting them? How many copies are being collected, where are they stored, and which dates have been collected? What is the condition of the hard copies, are they accessible for digitization, and for the public to browse? Can missing issues be found?

2. **Storage** – Is a centralized storage facility for hard copies warranted, and if so, how would this be established and funded? Or, if a distributed storage system is adequate, how will it be coordinated to ensure that copies are being collected and stored in an acceptable manner?
3. **Microfilming** – Microfilm is still considered to be the archival storage media of choice, and will be for the foreseeable future. Is our microfilm being created and stored to archival standards, as mandated by the Research Libraries Group (RLG)? Where are the master reels of microfilm being stored? Do we have appropriate contracts with our microfilm provider?
4. **Digitization** – How will newspapers be prioritized for digitization? Do we have permission to digitize all years, or are we limited to those years in the public domain? Can we get permission? Can we establish relationships with the largest newspapers (*SL Tribune, Deseret News*) to digitize their papers? Can the digitization process established by the previous LSTA grant be improved? Will digital newspaper files be stored on a centralized server, or distributed at sites around the state? What additional hardware will be required to support the digital files as they reach large numbers? Is it possible for Utah to participate in the LOCKSS (Lots of Copies Keep Stuff Safe) program?
5. **Electronic copy collection and storage** – Newspaper publishers currently use software to produce their newspapers. In our surveys we have found that some publishers simply discard their electronic files after publication. This is an alarming condition and must be changed so that the files are collected and archived by libraries. Digitizing these files would eliminate the need for scanning and would produce more accurate search results. What software are the publishers using? Can files be exported? How can we store these files? Are publishers amenable to participating in an archival storage project?

## CONCLUSION

The newspaper digitization method developed in this project is cost-effective and easily duplicated at other sites already running CONTENTdm. Because CONTENTdm effectively manages and presents a variety of collection formats, no additional server or interface is required for digital newspapers. The low cost of entry and the ability to scale from small to extremely large collections with CONTENTdm makes an easy pathway to get started in the digital library business. Having proven its OCR and processing capabilities, iArchives has again been chosen as the contractor to process all 100,000 newspapers pages in the new LSTA grant for 2003.

The Utah Digital Newspapers project has generated considerable excitement in the state, even though it has not yet been publicized. For the first time historic regional newspapers in Utah are available and accessible free of charge to anyone with an Internet connection.

> *"We've never met before, but I sit at the front desk down here at the T-I [Times-Independent] I just wanted to thank you for forwarding that link to the T-I archives online. Answering the editor's email, you wouldn't believe how many inquiries I field from people searching for long-lost information about their Moab ancestors. Now I can give the people who don't have access to the T-I on microfilm at the GC Library or at the U of U a place to go to do research.".*[5]

## REFERENCES

[1] Burton, Doris Karren. Quote from letter of support for LSTA grant. September 23, 2002.

[2] Entlich, Richard. "Where are they now? Digitizing Microfilmed Newspapers." RLG DigiNews June 15, 2002, Volume 6, Number 3

[3] Deegan, Marilyn. "Digitizing Historic Newspapers: Progress and Prospects." RLG DigiNews August 15, 2002, Volume 6, Number 4

[4] Arlitsch, Kenning. "Digitizing Sanborn Fire Insurance Maps for a Full Color, Publicly Accessible Collection." D-Lib Magazine. vol 8 no. 7/8, July-August 2002.

[5] Warner, Sadie. Assistant Editor at the Times-Independent in Moab, Utah, in an email to Eve Tallman, director of the Grand County Library. December 30, 2002. (*The Grand Valley Times* merged with the *Moab Independent* in 1919 and became the *Times-Indpendent.*)