View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by The University of Utah: J. Willard Marriott Digital Library

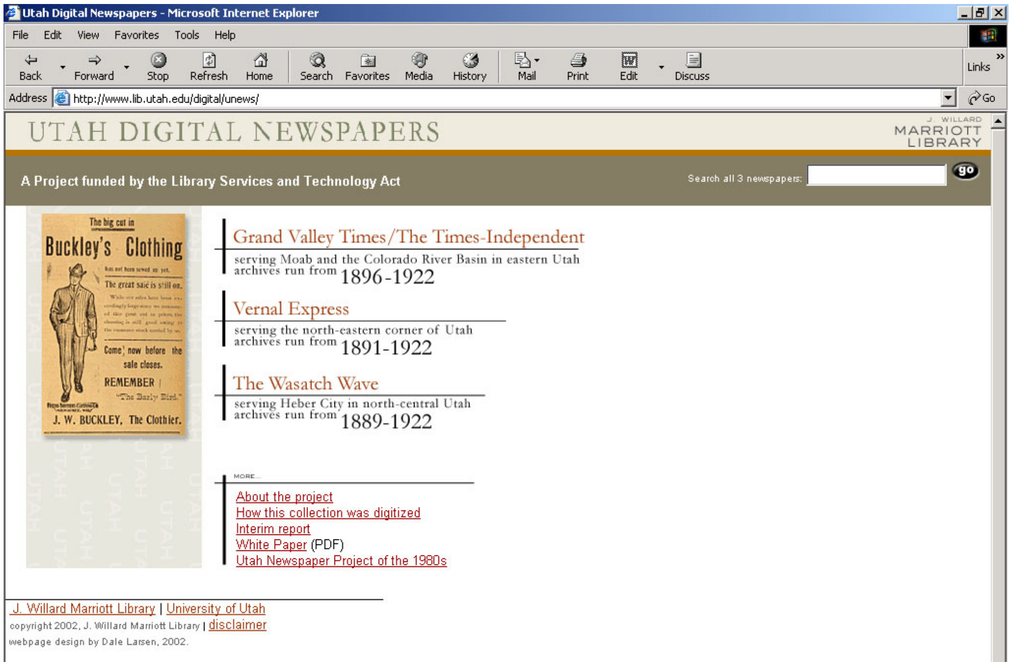
THE UTAH DIGITAL NEWSPAPERS PROJECT

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

BY KENNING ARLITSCH, L. YAPP, AND KAREN EDGE

ABSTRACT

The J. Willard Marriott Library at the University of Utah has digitized 30,000 pages from three weekly Utah newspapers from the period of 1889 – 1922 and made the collections freely available on the Internet. This article describes a new method for digitizing historic newspapers, developed in a partnership between the University and two commercial organizations. Utilizing OCR and newspaper processing technology from iArchives Inc. and the CONTENTdm digital collections software suite, the new method recently prototyped by the University of Utah presents a viable and affordable digitization method to cultural heritage institutions nationwide. In particular, the process can be implemented incrementally, making it affordable for both small and large collections, and the technology supports many different digital formats, not just newspapers. The digitized newspapers are publicly accessible and may be searched full text or browsed by issue. With the recent award of a new grant another 100,000 pages from an expanded selection of newspapers are slated for digitization in 2003.



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

UTAH DIGITAL NEWSPAPERS

A Project funded by the Library Services and Technology Act

Grand Valley Times/The Times-Independent
serving Moab and the Colorado River Basin in eastern Utah
archives run from 1896-1922

Vernal Express
serving the north-eastern corner of Utah
archives run from 1891-1922

The Wasatch Wave
serving Heber City in north-central Utah
archives run from 1889-1922

MORE...
About the project
How this collection was digitized
Interim report
White Paper (PDF)
Utah Newspaper Project of the 1980s

J. Willard Marriott Library | University of Utah
copyright 2002, J. Willard Marriott Library | disclaimer
webpage design by Dale Larsen, 2002.

Figure 1: Utah Digital Newspapers website.

BACKGROUND

Too often the histories of small-town newspapers, and the people who worked long hours to produce them, are forgotten. 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, and it is the broad appeal of these materials that garnered so much support in the academic and public library communities to drive the Utah Digital Newspapers project. In spite of their popularity, newspapers are also one of the most difficult and inefficient research materials, and they are often not consulted by researchers simply because they are so difficult to use. Regional historical newspapers are rarely indexed, and therefore cannot be searched, and are usually 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.[1] 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.[2] In general, most 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 one-year project. The three newspapers chosen for digitization have existed in Utah since the late 19th century and all are still published today:

The Vernal Express (Vernal, Utah)
Grand Valley Times/Times Independent (Moab, Utah)
Wasatch Wave (Heber City, Utah)

EARLY UTAH NEWSPAPER PROJECTS

Almost all of the microfilm of Utah historical newspapers dates from the 1950's. The filming was done by Universal Microfilming "which handled the collection and filming of the county papers." The company "would contact the local publisher and ask for the permission to film with the inducement that the Marriott Library would then subscribe to future issues." [3]

In the early 1980's the Marriott Library at the University of Utah was awarded a grant from the National Endowment for the Humanities (NEH) for planning the preservation of Utah newspapers, conducting an inventory of the papers, and filming them. The Utah Newspaper Project was designed to be part of a larger, nationwide effort for bibliographic control of newspapers - the NEH-funded United States Newspaper Project. The Marriott Library was ahead of the game on the filming, however, because most of the weekly newspapers in the state had been filmed since the 1950s.[4] The major job of the project, then, was to identify what state newspapers existed, verify the bibliographic records, and publish a checklist.[5]

In the late 1980's, the Uintah County Library in eastern Utah received a grant from the Utah Library Association (ULA) to begin indexing *The Vernal Express* newspaper. The library followed ULA indexing guidelines[6], and indexed only articles and advertisements that pertained to local news. Index fields included dates, page numbers, article titles, and subject headings. This index was used in 2001 when the University of Utah began to develop the newspaper digitization processes.

Syndicated News

In the late 19th century, the Western Newspaper Union of Omaha, Nebraska sold newsprint, printer's supplies and equipment to small town newspapers in Utah. It also sold syndicated news and other articles in the form of ready-print. The local newspaper editor would receive the newsprint, which was printed with national and international news on one side, and had space for local news on the other side. Western Newspaper Union also offered stereotype (metal) plates imprinted with articles. The plates could be fitted into the local newspaper's printing press.

Syndicated news was of great value because local newspapers could not afford to hire reporters to cover international and national events. Ready-print was a great bargain. Because the pre-printed advertisements paid most of the cost, ready-print was almost as cheap as blank newsprint. In turn, ready-print advertisers had a national market for their goods and services.

Besides national and international news, the Western Newspaper Union offered a wide range of content including articles on agriculture, fashion, and other topics as well as columns of interest to children, serialized fiction, poetry, and other features. Many established and fledgling writers found it worthwhile to have their work published through syndicated services like Western Newspaper Union. However, none of these fascinating articles or advertisements had been included in the index for *The Vernal Express*.

ADAPTABILITY

The newspaper digitization method developed at the University of Utah offers an affordable and incremental process that is adaptable to other institutions and other regions. It utilizes a two-pronged approach that may be adopted as a complete solution, or can be split to integrate with other systems.

1. A Utah-based service bureau, iArchives Inc., scans and processes the newspapers to produce open source image files and XML-tagged metadata. Other service bureaus across the nation could produce similar data for import into a variety of databases.

2. A relatively inexpensive digital collections database (CONTENTdm from DiMeMa Inc.), already in use at nearly 100 institutions across the nation, is utilized for managing and presenting the newspapers. CONTENTdm can be used for a variety of formats, including: photographs, books, maps[7], documents, 3D objects, and streaming media.

The newspaper digitization method is also adaptable in size. Libraries and other cultural heritage institutions generally do not have funding to digitize large runs of newspapers. The method developed in Utah can be implemented in an incremental manner, i.e., small institutions can begin digitizing only a few years of a local newspaper with a very small investment.

The Utah Digital Newspapers project will be integrated with the Mountain West Digital Library (MWDL), created by the Utah Academic Library Consortium (UALC). The MWDL is a cooperative digital library initiative designed to support the digital efforts of all cultural heritage institutions in Utah and Nevada. Implemented in early 2002, the MWDL currently consists of digitization centers at the libraries of the four largest universities in Utah:

University of Utah (Salt Lake City)	Brigham Young University (Provo)
Utah State University (Logan)	Southern Utah University (Cedar City)

The mission of the centers is to create their own digital collections and to support partner institutions by providing scanning and hosting services. The CONTENTdm Multi-Site Server installed at the University of Utah automatically harvests metadata from the four centers, providing searching capability across all the digital collections in the state from a single website. At this time more than 50 collections from academic libraries, public libraries, museums, and historical societies are aggregated for a total of nearly 90,000 objects. UALC membership extends to the state of Nevada, and during 2003 we expect to add the two main campuses of the University of Nevada (Las Vegas and Reno) as digitization centers of the MWDL and to aggregate their collections.

METHODOLOGY

Local contractors scanned the newspapers from microfilm, some of which was of very poor photographic quality. At first we considered purchasing our own microfilm scanner, but the automated type that was required for such volume would have used most of our grant funding. Contracting the scanning proved much less expensive.

Initially we anticipated some manual indexing, but thanks to the existing index for *The Vernal Express*, and the OCR engines at iArchives used for the other two newspapers, we managed to avoid any manual indexing. Newspapers were scanned at 300 dpi, and in both grayscale and bitonal bit depths because some images were unreadable as bitonal scans. After the first batch of digital files was returned to the University, we attempted to use off-the-shelf OCR packages, but they failed miserably.

Two distinct methods were employed to digitize the three newspapers. After scanning, *The Vernal Express* newspaper was processed and loaded entirely by the University of Utah. TIFF files were cropped and batch converted to MrSID® format, and imported into CONTENTdm. The Uintah County Library's index was imported simultaneously and matched with the images, and the software automatically generated display JPEG images and thumbnails for display. The CONTENTdm Full Resolution Manager was employed to provide a metadata link to the MrSID® file, allowing users to view the JPEG image first, and then retrieve the MrSID® file for zooming, if desired. Compound documents (XML wrapper files) were built manually to tie together the pages of a single issue. These compound documents were linked to the browse feature of the website, while the imported index provided search capability.

The *Grand Valley Times* and the *Wasatch Wave* were scanned and processed by iArchives Inc., and then delivered to DiMeMa Inc. for loading. iArchives applied their own OCR engines to the images to produce searchable text. They also applied zoning techniques to the images to create PDF files for each article and page, applied XML tagging to the metadata, and re-keyed headlines and other data requested by the University of Utah.

LOADING PROCESS

Grand Valley Times and *Wasatch Wave* articles were classed into the following types: articles, weddings/engagements, obituaries, and advertisements. PDF and XML files were delivered to DiMeMa Inc. for further processing and loading into CONTENTdm. The deliverables were compressed TAR files, and each newspaper issue was contained in one file. The *Wasatch Wave* had over 1,700 issues and the *Grand Valley Times* over 1,350 issues. Issues comprised 4-8 pages. The files were written onto DVDs, and the average size of the compressed file was approximately 12 MB.

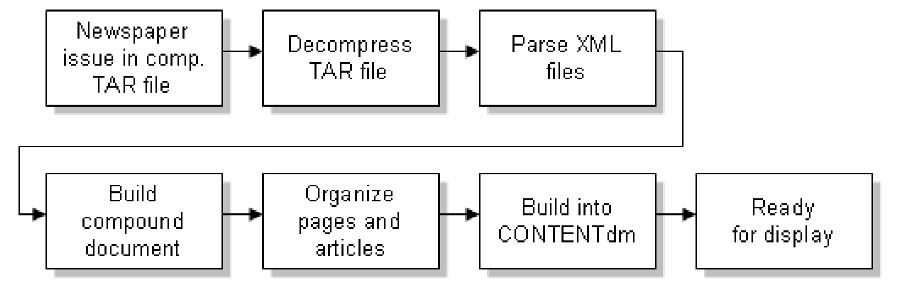
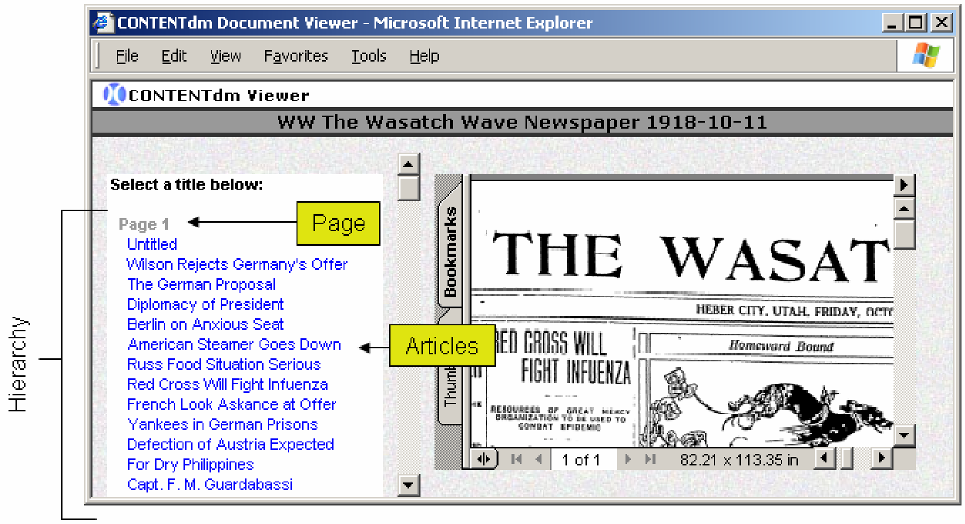


Figure 2: The steps to load one newspaper issue into the database for viewing.

For each issue, processing begins by decompressing the TAR file to extract the XML and PDF files that make up the newspaper. The XML files are used to describe the structure of the newspaper and contain a word transcription generated by several OCRs of the scanned pages. Then the XML files are parsed and a CONTENTdm compound document created. The compound document serves as a container to hold the pieces of the newspaper together and also to preserve the relationship between the pages and articles. For this project the PDF versions of the pages and articles were used, and they contain the OCR transcript text, making second-level searching possible from the PDF toolbar.



WW The Wasatch Wave Newspaper 1918-10-11

Select a title below:

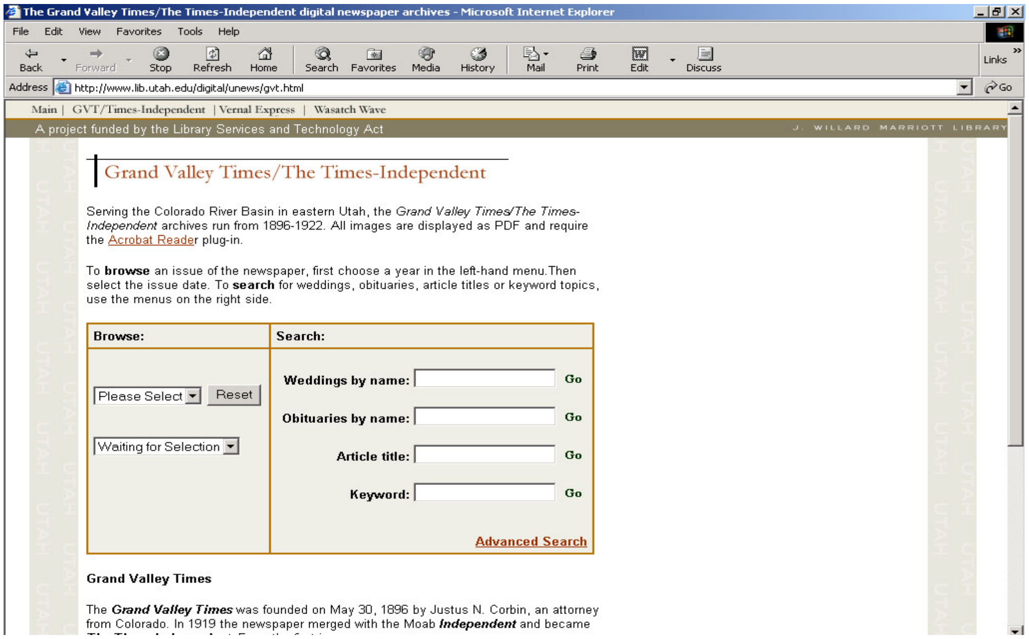
Page 1
- Untitled
- Wilson Rejects Germany's Offer
- The German Proposal
- Diplomacy of President
- Berlin on Anxious Seat
- American Steamer Goes Down
- Russ Food Situation Serious
- Red Cross Will Fight Infuenza
- French Look Askance at Offer
- Yankees in German Prisons
- Defection of Austria Expected
- For Dry Philippines
- Capt. F. M. Guardabassi

Figure 3: The monograph representation of a newspaper issue.

A monograph compound document type was used to model the newspaper. The monograph was chosen for its flexibility and ability to represent hierarchies. As a result, each of the newspaper pages can be displayed as its own link with all of its article titles displayed just below, as shown in Figure 3. Note that the monograph representation allows every page and article to be clickable. The PDF also holds all of the pieces of an article so that it can be viewed as one even if the article spans several newspaper pages.

The processing continues by extracting all of the transcription words from the XML files and other information such as dates, file size, article types, and title. A description file is created based on this information and is used to index the page, article, and newspaper. Generating a 160x120 pixel JPEG file from the PDF image creates a thumbnail image. The final step is uploading and building all of the items into the CONTENTdm database. The newly loaded newspaper issue can then be viewed using a web browser such as Internet Explorer.

The biggest issue that needed to be resolved was a way to automatically load the entire newspaper issue with minimal interaction. Significant computer processing power is required, especially in the XML parsing step. For example, it may require up to ten minutes to process one issue using a Dell PowerEdge 2600 with dual 2.2 GHz Xeon processor and 512 MB RAM. This process is further complicated if a user must periodically interact with the tools to load the issues one at a time. The solution was to create an asynchronous pipeline process with independent stages to perform tasks such as decompression, parsing, uploading, and building. With this process the entire loading of the newspaper issues is fully automated and can run with very little user interaction.


The Grand Valley Times/The Times-Independent digital newspaper archives - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Forward
Stop
Refresh
Home
Search
Favorites
Media
History
Mail
Print
Edit
Discuss
Links
Address
Go
Main | GVT/Times-Independent | Vernal Express | Wasatch Wave
J. WILLARD MARRIOTT LIBRARY
UTAH

Address: http://www.lib.utah.edu/digital/unews/gvt.html

A project funded by the Library Services and Technology Act

Grand Valley Times/The Times-Independent

Serving the Colorado River Basin in eastern Utah, the *Grand Valley Times/The Times-Independent* archives run from 1896-1922. All images are displayed as PDF and require the Acrobat Reader plug-in.

To **browse** an issue of the newspaper, first choose a year in the left-hand menu.Then select the issue date. To **search** for weddings, obituaries, article titles or keyword topics, use the menus on the right side.

Browse:

Please Select ▾ Reset

Waiting for Selection ▾

Search:

Weddings by name: Go

Obituaries by name: Go

Article title: Go

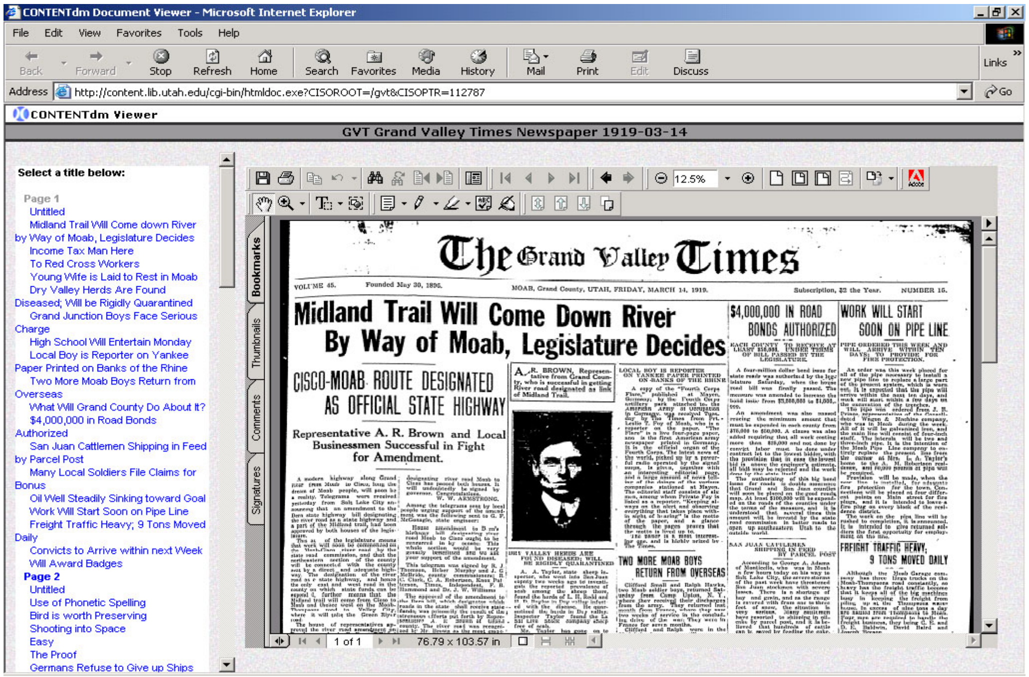
Keyword: Go

Advanced Search

Grand Valley Times

The ***Grand Valley Times*** was founded on May 30, 1896 by Justus N. Corbin, an attorney from Colorado. In 1919 the newspaper merged with the Moab ***Independent*** and became

Figure 4: The browse and search screen for the *Grand Valley Times*.



CONTENTdm Document Viewer - Microsoft Internet Explorer

Address http://content.lib.utah.edu/cgi-bin/htmldoc.exe?CISOROOT=/gvt&CISOPTR=112787

CONTENTdm Viewer

GVT Grand Valley Times Newspaper 1919-03-14

Select a title below:

Page 1
Untitled
Midland Trail Will Come down River by Way of Moab, Legislature Decides
Income Tax Man Here
To Red Cross Workers
Young Wife is Laid to Rest in Moab
Dry Valley Herds Are Found Diseased; Will be Rigidly Quarantined
Grand Junction Boys Face Serious Charge
High School Will Entertain Monday
Local Boy is Reporter on Yankee Paper Printed on Banks of the Rhine
Two More Moab Boys Return from Overseas
What Will Grand County Do About It?
$4,000,000 in Road Bonds Authorized
San Juan Cattlemen Shipping in Feed by Parcel Post
Many Local Soldiers File Claims for Bonus
Oil Well Steadily Sinking toward Goal
Work Will Start Soon on Pipe Line
Freight Traffic Heavy; 9 Tons Moved Daily
Convicts to Arrive within next Week
Will Award Badges
Page 2
Untitled
Use of Phonetic Spelling
Bird is worth Preserving
Shooting into Space
Easy
The Proof
Germans Refuse to Give up Ships

The Grand Valley Times

VOLUME 45. Founded May 30, 1896. MOAB, Grand County, UTAH, FRIDAY, MARCH 14, 1919. Subscription, $2 the Year. NUMBER 15.

Midland Trail Will Come Down River By Way of Moab, Legislature Decides

CISCO-MOAB ROUTE DESIGNATED AS OFFICIAL STATE HIGHWAY

Representative A. R. Brown and Local Businessmen Successful in Fight for Amendment.

A. R. BROWN, Representative from Grand County, who is successful in getting River road designated as link of Midland Trail.

LOCAL BOY IS REPORTER ON YANKEE PAPER PRINTED ON BANKS OF THE RHINE

DRY VALLEY HERDS ARE FOUND DISEASED; WILL BE RIGIDLY QUARANTINED

TWO MORE MOAB BOYS RETURN FROM OVERSEAS

$4,000,000 IN ROAD BONDS AUTHORIZED

EACH COUNTY TO RECEIVE AT LEAST $50,000, UNDER TERMS OF BILL PASSED BY THE LEGISLATURE.

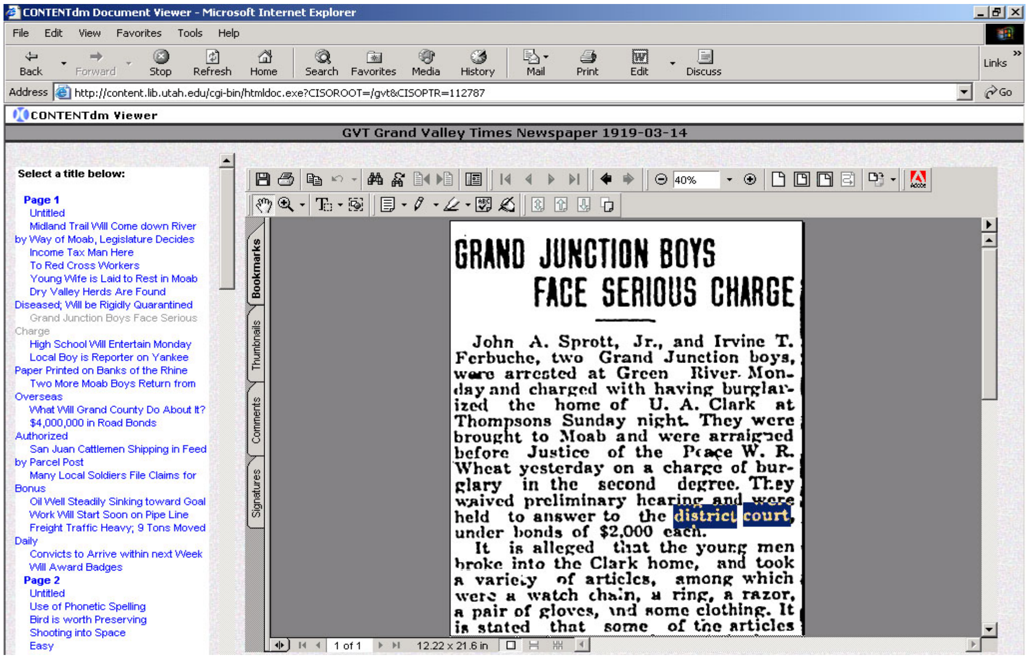
SAN JUAN CATTLEMEN SHIPPING IN FEED BY PARCEL POST

WORK WILL START SOON ON PIPE LINE

PIPE ORDERED THIS WEEK AND WILL ARRIVE WITHIN TEN DAYS; TO PROVIDE FOR FIRE PROTECTION.

FREIGHT TRAFFIC HEAVY; 9 TONS MOVED DAILY

Figure 5: Full page view. Links to other pages and articles for this issue are in left frame.



CONTENTdm Document Viewer - Microsoft Internet Explorer

Address http://content.lib.utah.edu/cgi-bin/htmldoc.exe?CISOROOT=/gvt&CISOPTR=112787

CONTENTdm Viewer

GVT Grand Valley Times Newspaper 1919-03-14

Select a title below:

Page 1
Untitled
Midland Trail Will Come down River by Way of Moab, Legislature Decides
Income Tax Man Here
To Red Cross Workers
Young Wife is Laid to Rest in Moab
Dry Valley Herds Are Found Diseased; Will be Rigidly Quarantined
Grand Junction Boys Face Serious Charge
High School Will Entertain Monday
Local Boy is Reporter on Yankee Paper Printed on Banks of the Rhine
Two More Moab Boys Return from Overseas
What Will Grand County Do About It?
$4,000,000 in Road Bonds Authorized
San Juan Cattlemen Shipping in Feed by Parcel Post
Many Local Soldiers File Claims for Bonus
Oil Well Steadily Sinking toward Goal
Work Will Start Soon on Pipe Line
Freight Traffic Heavy; 9 Tons Moved Daily
Convicts to Arrive within next Week
Will Award Badges
Page 2
Untitled
Use of Phonetic Spelling
Bird is worth Preserving
Shooting into Space
Easy

GRAND JUNCTION BOYS FACE SERIOUS CHARGE

John A. Sprott, Jr., and Irvine T. Ferbuche, two Grand Junction boys, were arrested at Green River Monday and charged with having burglarized the home of U. A. Clark at Thompsons Sunday night. They were brought to Moab and were arraigned before Justice of the Peace W. R. Wheat yesterday on a charge of burglary in the second degree. They waived preliminary hearing and were held to answer to the district court, under bonds of $2,000 each.

It is alleged that the young men broke into the Clark home, and took a variety of articles, among which were a watch chain, a ring, a razor, a pair of gloves, and some clothing. It is stated that some of the articles

Figure 6: Article view. Displayed article title is highlighted in left frame, and second level (hidden text) search result is shown in image.

EVALUATION

The method used to digitize *The Vernal Express* relied heavily on manual processes, and will not be used in the future. It grew out of the availability of the index provided by the Uintah County Library and because the processes with iArchives Inc. and DiMeMa Inc were still being conceived. The method used to digitize the *Wasatch Wave* and the *Grand Valley Times* has proven much more efficient and effective, and it 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:

THE VERNAL EXPRESS

Advantages

- The initial JPEG image loads quickly without need for a plugin, and headlines are legible.
- The high-resolution MrSID file is only loaded when requested.
- This method produces a 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

- Full-text searching is not possible.
- Individual articles are not zoned (full page view only).
- A free MrSID plug-in or viewer is required to see detailed text.
- MrSID files average 3MB each.
- Larger total server space is used than for the other two newspapers (21.5GB).
- Manual processes to create the collection were slow and inefficient.

THE WASATCH WAVE AND GRAND VALLEY TIMES

Advantages

- Full text searching is available.
- Articles are zoned.
- This method enables PDF navigation (ubiquitous viewer).
- Small file sizes (10KB-50KB for articles, 300KB for full pages) result from processing.
- Only a small total server space (approximately 10GB for each newspaper) is required.
- Automated processes from iArchives and DiMeMa greatly increases volume and efficiency.

Disadvantages

- This method results in a large number of files because of the zoned articles. Each of the two newspapers collections contains approximately 255,000 files (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. The iArchives processing fee includes OCR, article zoning, XML tagging, and re-keying of certain data such as headlines. DiMeMa's fee includes batch importing the images and metadata and delivering completed collections. 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 Utah Academic Library Consortium, led by 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 from 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 scanning newspapers from hard copy instead of only from microfilm. The hard copies are rare, but as word of the project has spread, several runs have been located and we expect better images and more accurate text searching as a result.

Other future developments may include a new generation viewer for the digital newspapers. While PDF is ubiquitous and the file sizes are quite small it still requires a plug-in, and new developments with JPEG2000 and other technologies may eliminate the need for any plug-in software. We will continue to work with DiMeMa Inc. to implement such a change.

CONCLUSION

The newspaper digitization method developed in this project is cost-effective and can be duplicated at other sites. Newspaper scanning and processing results are non-proprietary, and the database software used to present these and other collections on the Web utilizes open source images and metadata fields.

The Utah Digital Newspapers project has generated considerable excitement in the state, even though it has not yet been formally 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.”* [8]

REFERENCES

[1] Entlich, Richard. “Where are they now? Digitizing Microfilmed Newspapers.” *RLG DigiNews* June 15, 2002, Volume 6, Number 3

[2] Deegan, Marilyn. “Digitizing Historic Newspapers: Progress and Prospects.” *RLG DigiNews* August 15, 2002, Volume 6, Number 4

[3] Holley, Robert P. *The Utah Newspaper Project Final Report, Project Number PS-200010-85*. University of Utah Libraries, 1987. p. 5.

[4] Conversation with Yvonne Stroup, Serials Cataloging Librarian at the University of Utah. February 14, 2003.

[5] Holley, Robert P., ed. *Utah's newspapers, traces of her past : papers presented at the Utah newspaper project conference, University of Utah, November 18, 1983: With checklist of Utah newspapers* / compiled by Dennis McCargar ; edited by Yvonne Stroup. University of Utah, 1984.

[6] ULA Steering Committee on Utah Newspaper Indexing. *Utah Newspaper Indexing*, Salt Lake City, Utah: May 1990.

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

Deleted: D-Lib Magazine

[8] 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-Independent*.)

ACKNOWLEDGEMENTS

The authors would like to express their appreciation for the historical and institutional memories of the following people:

Yvonne Stroup	Cataloging Serials Librarian	University of Utah
Margaret Brady	Serials Staff	University of Utah
Doris Burton	Director	Uintah County Regional History Center