Submitted on: 12/08/2014

Texas Newspaper PDF Preservation: A Low-Cost Solution with Tremendous Value

Ana Krahmer
Digital Newspaper Unit, University of North Texas Libraries, Denton, Texas, USA.
E-mail address: ana.krahmer@unt.edu

Mark Phillips
Digital Libraries, University of North Texas Libraries, Denton, Texas, USA.
E-mail address: mark.phillips@unt.edu

Abstract:

In 2007, University of North Texas Libraries was the Texas institution selected to participate in the National Digital Newspaper Program, funded by the National Endowment for the Humanities. Shortly after announcement of the selection, multiple communities contacted UNT Libraries to ask if they could participate. Due to enormous demand for newspaper preservation in Texas, UNT Libraries established the Texas Digital Newspaper Program (TDNP), a digital newspaper initiative to host and preserve newspapers from any year, from any Texas community, via The Portal to Texas History (http://texashistory.unt.edu/explore/collections/TDNP/). Since establishing TDNP, UNT Libraries has preserved nearly 1.5 million pages of newspapers, dating from 1829 to present. In 2009, a publisher from Rusk, Texas, contacted UNT Libraries to discuss preservation of her PDF e-print editions, and this started UNT Libraries on the road to preserving newspapers of this additional media type.

Newspaper PDFs, or e-print editions, are rich targets for digital preservation because they represent a significant cost-benefit gain: large, full-color PDF e-print editions provide a relatively inexpensive means for digital conversion and preservation in comparison with fullpage physical newspaper or microfilm scanning. In communicating with libraries and newspaper publishers about scanning microfilmed newspapers, we have learned that, in cases especially of rural public libraries and publishers, microfilm is now too costly to create each month or year, and additionally that PDF editions are often not saved with a long-term strategy in mind. To assist with this problem, UNT Libraries offers PDF *preservation and hosting services to Texas publishers and public libraries. This paper will address the process UNT Libraries employs to collect, convert, and host newspaper PDFs on The Portal to Texas History; this paper will also address how UNT Libraries works with publishers to gain permission and content and to negotiate embargo periods for the preserved PDFs.*

Keywords: Preservation, repositories, newspapers, digital libraries

1 INTRODUCTION

Newspaper PDF e-print editions are a standard occurrence in the newspaper industry, created with the intent to be sent to the printer's office for paper distribution (CRL, 2012, p. 30). Another common format is microfilm, whose creation from physical newspaper pages has been a common, industry-wide preservation practice since the 1960s (Heritage Archives, 2010). With coordination between publishers and libraries, newspaper PDFs offer a significantly less expensive, higher utility preservation alternative to microfilm creation. Lisa Fox, in *Preservation Microfilming,* highlights the cost to film a book volume: "... the filming of a volume, one that has many physical features that may complicate filming, might be only $150. This, admittedly, is higher than the $100 typical cost for routine filming of general collections" (Fox, 1996, p. 28). Although Fox's numbers are from 1996, this pricing is close to today's pricing, though the cost of newspaper transport to and from the microfilm vendor is also a significant factor.

By contrast, PDF newspapers can be loaded onto an external hard drive or uploaded via a deposit architecture, and can then be processed digitally. Because PDF e-print editions are already digital, no overhead is needed to fund analog-to-digital conversion. In addition, PDF e-print editions do not lose color content, whereas microfilming in the most common black-and-white format causes color data loss.

University of North Texas Libraries' Digital Newspaper Team has begun to coordinate with newspaper publishers and the Texas Press Association to preserve current Texas PDF e-print editions. This entails such communication processes as facilitating communication, securing permissions, negotiating embargo periods, and arranging for transport of PDFs to and from publishers. Technologically, this process fits within the overall digital preservation and access infrastructure for the Texas Digital Newspaper Program, which is The Portal to Texas History (http://texashistory.unt.edu). Because PDF processing can fit neatly within the extant preservation framework, very little financial overhead is necessary to process the digital files.

2 UNIVERSITY OF NORTH TEXAS LIBRARIES' DIGITAL PRESERVATION FRAMEWORK

Much of the success of the Texas Digital Newspaper Program is tied directly to the digital preservation infrastructure at University of North Texas Libraries, upon which three access systems are hosted. In 2008 UNT Libraries made the decision to locally develop a digital asset management system that came to be called "Aubrey," which included development of a preservation repository, called Coda. The Coda preservation repository is operated in a single instance that is utilized by all of our digital access systems.

The system is built with the Django Web Framework, Solr for full-text indexing, and uses curation micro-services (Abrams, Kunze, & Loy, 2010) for storage of digital items on a file system. METS files encapsulate digital objects and provide structure for pagination. This system was built to support The Portal to Texas History and was later expanded to include the UNT Digital Library, as well. Then, in 2012, the Gateway to Oklahoma History was launched on this same access platform.

All UNT Libraries' digital object access systems support all types of image-based digital objects, including maps, books, photographs, archival materials, and, of course, newspapers. Of particular importance and especially relevant to newspaper digitization is that UNT Libraries staff took special consideration for allowing in-document searching, page-zooming functionality, and highlighting of terms on a page for improved usability of full-text search activities.

In addition to newspapers and other primary source materials, this system also hosts audio, video, and datasets. Because UNT Libraries locally developed the system, it has grown into a system that supports our primary model for digital objects and has a specific way of functioning. This system provides more functionality for our digital library operations than what most out-of-the-box, proprietary systems could provide. Because of the high customizability of the system, we are able to incrementally improve the access and metadata editing environment as time permits, which also enables us to perform usability testing and user experience design customizations (IOGENE, 2013). There is one primary data store and set of Solr indexes for all access platforms we operate. Currently three access systems sit on top of the preservation infrastructure, and The Portal to Texas History is the system designed to host Texas primary source materials, including all newspapers in the Texas Digital Newspaper Program.

3 ABOUT THE TEXAS DIGITAL NEWSPAPER PROGRAM

The Texas Digital Newspaper Program (TDNP) has an established mission to collect, preserve, and provide access to the newspaper output of the state of Texas. TDNP started in 2005, and then moved forward in 2007 when it became an extension of the National Digital Newspaper Program. At that time, UNT Libraries began working with institutions throughout Texas, including the Dolph Briscoe Center for American History at the University of Texas, the Texas State Library and Archives Commission, the Abilene Library Consortium, and numerous local and regional cultural heritage institutions to move forward digital preservation and access to the rich newspaper content created and held in the State of Texas.

Through word-of-mouth, TDNP staff have learned from publishers and libraries how they use their most recent, PDF e-print edition newspaper content. Texas publishers and libraries have reported that they no longer microfilm physical pages due to high cost and a lack of availability of microfilm reading machines. Instead, some publishers provide their libraries with DVDs of newspaper PDF content.

In 2010, the Texas Press Association established the Texas Press Archive, a depository system by which publishers could preserve their PDFs, without making them publicly accessible. However, even four years after its initiation, the Texas Press Archive only has 75% participation, according to its originator, Michael Hodges (personal communication, June 16, 2014). Local public libraries in Texas have told TDNP staff that they view this as a problem for long-term preservation of the early 21st-century newspaper content.

The Texas Digital Newspaper Program team regularly hears about this problem when the local public libraries call for advice on working with their publishers to continue to microfilm their newspapers on an annual basis. To respond to this problem, since 2010, the Texas Digital Newspaper Program offers a PDF e-print edition preservation service to publishers. The PDF e-print edition files can be converted, placed into the preservation infrastructure, and made publically accessible. Because these PDF editions are born digital, they are also relatively inexpensive for TDNP staff to preserve, as they require no analog-to-digital conversion. The process of preserving PDF e-print editions begins with acquisition, moves toward building issues into submission information packages (SIPs), and then moving them to ingest and long-term preservation via The Portal to Texas History; this paper explores how this process fits into University of North Texas Libraries' overall preservation infrastructure and strategy.

4 MISSION AND SCOPE OF THE TEXAS DIGITAL NEWSPAPER PROGRAM

The Portal to Texas History functions as both a preservation infrastructure and a content-access gateway. Hosted on the Portal are over 420,000 visible items and just over 17,000 hidden objects, of which are included all newspapers in the Texas Digital Newspaper Program. All newspapers hosted by TDNP are preserved for the long term via a preservation infrastructure based on curation micro-services model (Abrams, Kunze, & Loy, 2010). In addition to newspapers, the Portal also hosts and preserves other primary source object types from across Texas, including but not limited to: maps, photographs, diaries, yearbooks, letters, city directories, personal papers, patents, and manuscripts. The goal of the Portal is to house and make openly accessible primary source objects from across the State of Texas, to allow free access to Texas historical materials whose physical locations researchers would otherwise have to physically visit to see.

The Texas Digital Newspaper Program now serves nearly 2 million pages of Texas newspapers, contained in over 195,000 issues. There are more than 670 titles in the TDNP collection from across the state. At the beginning of 2014, Krahmer and Phillips analyzed the development and geographic distribution of the titles in the TDNP collection as a way of assessing the growth and coverage of the program. TDNP collection data showed that, as of early 2014, many of the TDNP newspapers ranged in date from between the late 1870s to the early 1920s. This wealth of pre-1923 content is easily explained by the fact that, in the United States, 1923 is the year that divides public-domain content from content that is either protected by copyright, or which requires further investigation to understand its rights restrictions. The Texas Digital Newspaper Program staff regularly work with publishers and cultural memory institutions throughout Texas to build access to newspaper content from later than the 1923 date.

The permissions-gathering process entails researching ownership and negotiating with publishers for the rights to preserve and make available content from after 1923. Table 1 displays the distribution of issues by decade held in the TDNP collection.

Decade	**Issues**
1830-1839	302
1840-1849	1,439
1850-1859	3,800
1860-1869	4,185
1870-1879	5,215
1880-1889	12,168
1890-1899	13,040
1900-1909	17,695

1910-1919	26,083
1920-1929	18,069
1930-1939	17,592
1940-1949	18,144
1950-1959	13,673
1960-1969	8,842
1970-1979	6,602
1980-1989	5,063
1990-1999	4,448
2000-2009	6,790
2010-2019	3,184

Table 1: Issues per decade in the TDNP collection.

Clearly, there are fewer issues after 1960, and this is explicable for a number of reasons. First, several institutions have had restrictions placed on their digitization projects by their respective granting agencies, who are currently focusing on pre-1960 issues of titles. Second, as we move into more recent years, publishers place a higher value on these publications because they perceive a potential market for their digitized titles. Finally, the most recent decades from 2000-2014 represent a shift in practice in how newspaper titles are created on the publishers' end, how the papers are printed, and, most significantly, in the digital curation approaches that we can take to prepare these newspapers for ingest and preservation.

5 STAGES FOR PREPARING NEWSPAPERS FOR PRESERVATION

The Texas Digital Newspaper Program has an established set of workflows designed for three distinct input formats: physical paper, microfilm, and born-digital content. Aside from the initial acquisition step in the workflow, the stages associated with the organization, preservation and access of newspaper content are the same, independent of format. (See Figure 1). Rights negotiations span the entire workflow, as the rights permissions can be revised and adjusted throughout the lifecycle with TDNP's goal being to provide the highest level of access for users. After Figure 1, we explore each stage in detail.

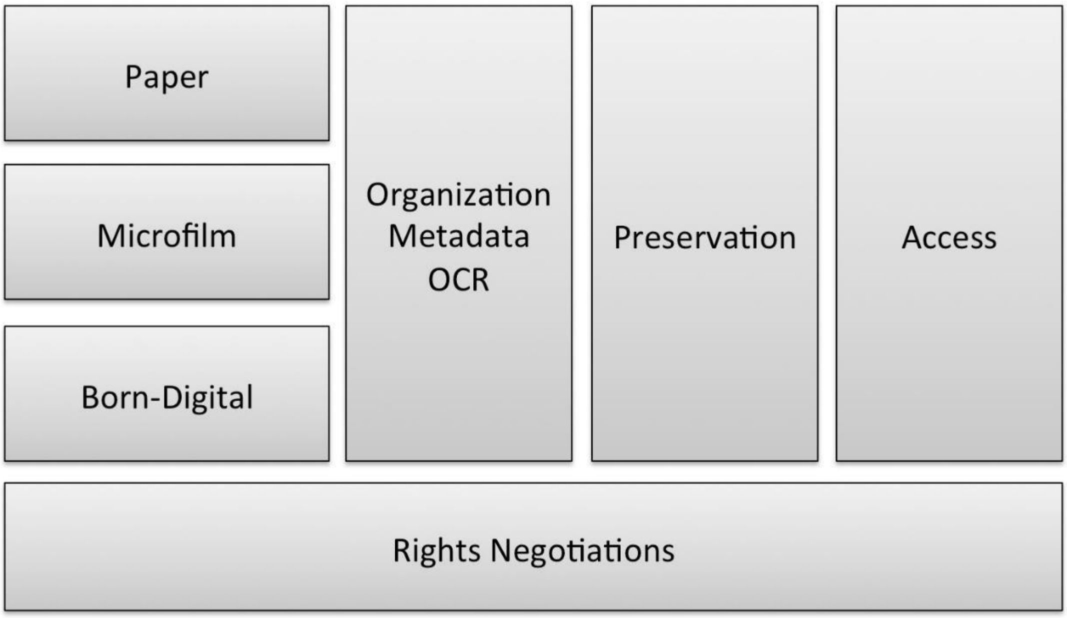


Figure 1: The overall workflow model for the TDNP program with the three separate input formats: paper, microfilm, and born digital, showing the activities of preparing acquired content, preservation, and access being the same, independent of the input content.

STAGE 1: ACQUIRE

As mentioned, the Texas Digital Newspaper Program has three primary ways of acquiring content: microfilm, paper, and born-digital. Because UNT Libraries is a National Digital Newspaper Program institution, microfilm processing for non-NDNP content is fairly standardized. Typically, we purchase 2nd-generation, duplicate negative film from one of the many holding institutions of Texas microfilm. These include both private companies, public institutions, and state agencies. As with NDNP, TDNP prefers a second-generation negative for digitization, but in some situations other microfilm formats are acceptable though not optimal. Because of their participation in NDNP, UNT Libraries can digitize microfilm in-house to standards established by the Library of Congress for NDNP (NDNP, 2013). Additionally, UNT Libraries can add content to TDNP even when external vendors have performed the digitization work, so long as vendor work meets the National Digital

Newspaper Program imaging standards.

Paper newspaper pages often come from publishers, cultural memory institutions, or private citizens who have collected and maintained these paper issues over time. Physical newspapers are scanned according to national standards established by Library of Congress. Thus, UNT Libraries scans all physical newspaper pages at 400 dpi, in 24-bit color, using an A0-sized, planetary scanner. While the process of physical newspaper digitization is time- and storage-intensive, the resulting files allow for high-quality Optical Character Recognition (OCR), and they retain the highest level of detail from the original, including color content that would be otherwise lost had the newspapers been digitized from microfilm.

Finally, TDNP works directly with publishers to acquire born-digital PDF e-print edition files because publishers are not necessarily in the process of depositing them with other local libraries or other cultural memory institutions as historically has been done with analog formats, as revealed through our informal discussions with publishers from across
Texas.

STAGE 2: ORGANIZE

Once acquired and prepared in a digital format, newspaper issues undergo a series of steps for ingest into The Portal to Texas History, which, as mentioned, is both an access and preservation system.

First, newspaper pages are sorted into issues. This step is currently done as an exercise in moving files into folders for organization within a traditional shared-network storage system. A folder for each issue of a title is then created, using the issue date and edition as the folder name. For example the pages from the issue of a newspaper from July 4th, 2013, go into a folder named 2013070401, thus following the format of yyyymmddee (y=year, m=month, d=day, e=edition). Most newspapers processed are the only edition published, and therefore are designated as edition 01 in the folder name. For either paper- or microfilm-based issues, the issue folder contains the image files and subsequent OCR, text and other bounding-box files. Born-digital issues are comprised of two sub-folders: 01_jpg, which holds the image files extracted from the PDF master files, OCR, text and other bounding-box files, and a folder named 02_pdf contains the master PDF files acquired from the publisher. Each issue also contains a metadata file which includes information related to the specific issue, such as volume, issue number, and unique information such as missing pages or incorrect information printed on the masthead. (See Figure 2.)

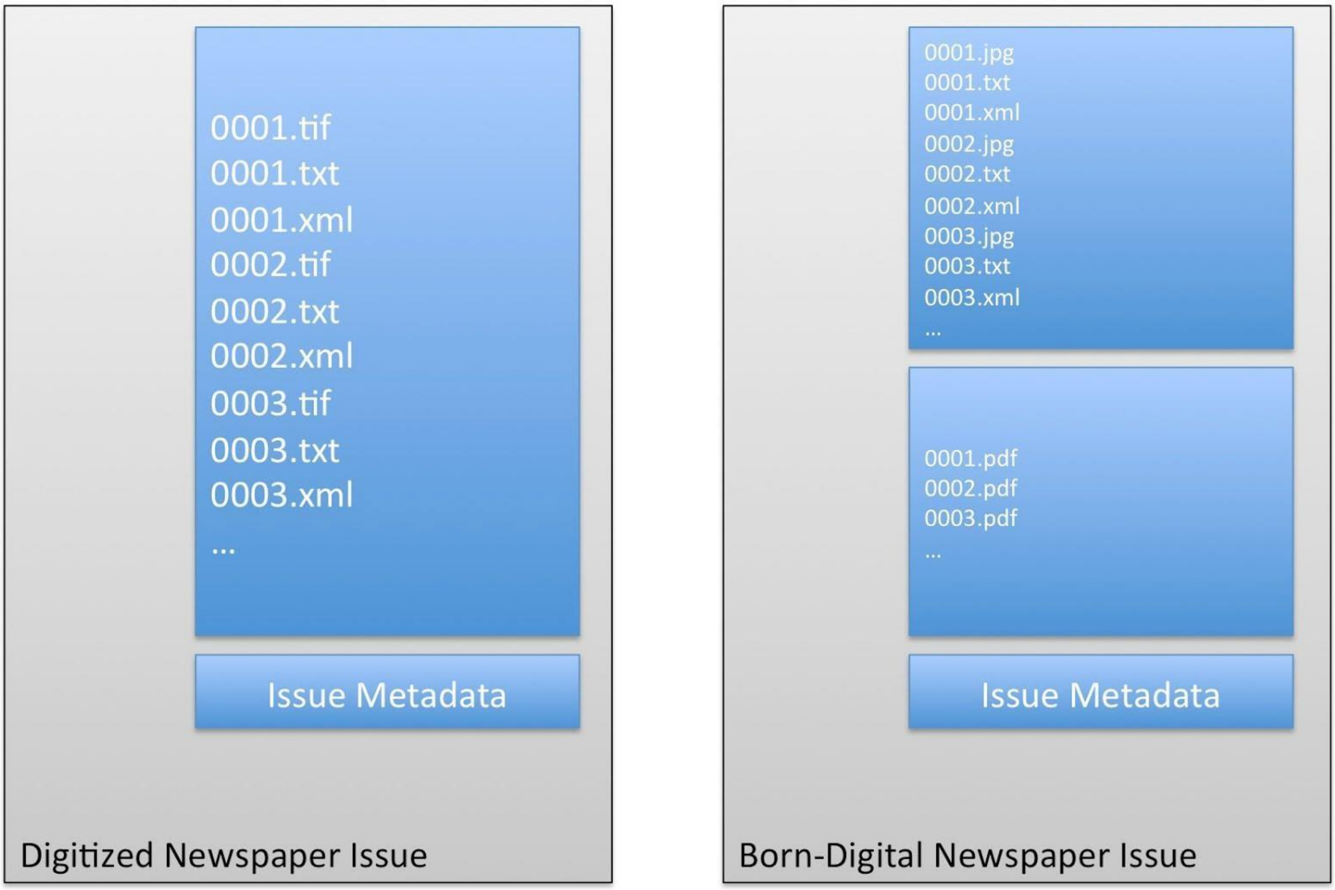


Figure 2: Pre-ingest issue structure for TDNP newspapers.

Once organized into the issue-level folder structure, each issue has Optical Character Recognition (OCR) applied to it to extract the text from the images. Even though PDF e-print edition files typically have text associated with them, we regularly perform OCR on the rasterized image to maintain a consistent workflow. Then, we save the output of the OCR engine into the issue folder. This output includes the raw OCR, a text representation of this output, and the bounding boxes denoting positions of words on the page. This enables full-text searchability with marked coordinates that allow highlighting of searched text.

Overall, it is during the first two stages that PDF e-print edition processing differs from analog conversion, from film or physical page to digital.

STAGE 3: PRESERVE

After initial preparation that normalizes the three different newspaper formats, all newspaper content is processed and loaded in the same way, into the UNT Libraries' digital repository. Accompanying each issue is a METS file containing structural information about the newspaper issue, including the correct ordering of pages within the issue. Preservation metadata is extracted from each file in the deposited package. This includes standard file information from the METS fileSec section and PREMIS Object records for each file in the issue package. A

JHOVE stream is generated and associated with each file in the issue package. A BagIt bag packages all submitted files and newly-created metadata files, and this provides a standardized way for transferring, verifying, and storing digital objects in the purpose-built repository infrastructure operated by UNT Libraries. Once registered in Coda, UNT Libraries' repository infrastructure, the digital newspapers are replicated to a remote storage facility using a local ResourceSync implementation (Open Archives Initiative, 2014) built into the repository infrastructure. All operations involved in the preservation process are performed as batch operations, with a typical batch of newspaper issues ranging from 100 to 1500 issues in size.

STAGE 4: BUILDING ACCESS

The Texas Digital Newspaper Program utilizes The Portal to Texas History as its access component. Derivative files are created from the high-resolution master TIFF or JPEG images, and this includes both JPEG2000 files and a Zoomify-based tile directory (Krahmer & Phillips, 2013, p. 10). The system tiles these derivatives in an OpenLayersbased viewer to provide full screen viewing of newspaper issues at multiple resolutions. For born-digital newspaper issues, the PDF e-print edition files are not included in the dissemination package added to the Portal because publishers generally do not want to provide download capability to complete, full-sized copies of their PDF e-print editions via the Portal.

More detailed information about each of these stages is available in "Laying the Groundwork for Newspaper Preservation through Collaboration and Communication: The Texas Digital Newspaper Program," by Krahmer and Phillips (2013), written for the World Library and Information Congress Satellite Meeting: *Newspapers to the People*.

6 WORKING WITH PUBLISHERS AND THEIR CONTENT

Prior to initiating a newspaper project, TDNP staff secures permissions from both the publisher and the cultural memory institution collaborating with UNT Libraries to digitize their community's newspaper. The Portal to Texas History has both a partnership agreement and an electronic permissions document and uses both for new projects with publishers and partner organizations. The Portal is a completely free and open access system and requires that items preserved and made available via the system are available freely to the public. Publishers of born-digital items have the opportunity to embargo their more recent content for a period of time, and this embargo functions as a moving access wall.

An example of this moving access wall is the *Rusk Cherokeen,* whose publisher embargoes the most recent three years of publications in order to drive traffic to the newspaper website but which makes issues older than three years freely available via The Portal to Texas History. At this point, these embargo periods are handled manually by Portal staff, but we are in the process of investigating the use of a more automated solution, which would allow the system to automatically open access to these resources once their embargo period has passed. This automated system is already in place for other collections in the UNT Digital Library, such as the UNT Scholarly Works Repository. Automating the moving wall process further would make it possible to work with a greater number of publishers of born-digital items which will have a variety of embargo periods associated with their newspaper titles.

7 BORN-DIGITAL PROCESSING STEPS

In 2012, the National Digital Stewardship Alliance prepared a case study report about preserving newspaper PDF, e-print editions. According to the NDSA, as newspaper microfilming practices have started to decline over the past decade, cultural memory institutions have reported preservation and access risks to the most recent newspapers published in the United States (NDSA, 2012). It is no accident that UNT Libraries associate dean, Cathy Hartman, and TDNP staff helped in preparing the NDSA case study, as TDNP staff have spoken with multiple publishers who report that they no longer microfilm their papers due to the high cost of filming and transporting, though they also do not ensure longterm integrity of the PDF e-print edition files.

As with all the other stages in TDNP newspaper processing, the steps for preserving PDF e-print editions begin with acquisition, move toward building issues into submission information packages (SIPs), and then to ingest and long-term preservation. Following is the in-depth process of preserving PDF-specific content, though the process for all content has already been discussed. Table 2 provides a more detailed outline of the processing steps the TDNP staff take to prepare PDF e-print editions for submission into the repository, and this more detailed view represents a slice of the actions that occur in the overall workflow model and stages described above.

Step	Action	Purpose
1	Transfer PDFs from publisher to UNT Libraries.	PDFs are moved on an external hard drive. Typically, UNT provides the drive to the publisher for transfer of PDF content.
2	Load PDFs on the digital newspaper fileshare server.	This allows Digital Newspaper Team staff and students to access the PDF files.
3	Move PDF issues into their own directories, with each directory named after the issue date plus edition.	Example: 2014050101; provides for easy location and organization of newspaper issues.
4	Within each issue directory, create two directories: One for JPEG derivative images, and one for PDF originals.	Allows for sorting of file types.

5A	If issues are divided into individual pages, combine files into one PDF using Adobe Acrobat Pro.	This step may or may not be necessary.
5B	Within Acrobat Pro, configure saving options to 400 dpi JPEG images.	Although we save to JPEG format for derivative access images, the PDF print masters are the preservation media types.
6	Save files as JPEG images, with no compression, with 400 dpi, at fullcolor, 24-bit depth.	JPEG files save into their respective issue directory->JPEG directory.
7	Perform quality control on saved images. Verify: Page order; pages are not missing, correct rotation on saved files, and correct color information level.	Quality control ensures that PDF content has successfully converted to JPG images.
8	At the issue directory level, create a text file for metadata, containing volume, issue, and (when necessary) explanatory aberration information.	The combination of volume and issue number within the text file ties to the issue directory, thus attaching a date to each volume/issue of the newspaper page files.
9	Perform a final quality control on issue-level metadata.	This step utilizes a local Python script, issueCheck.py, to verify that metadata fields are accurate and consistent. For example, if an issue is missing a page, the note in the metadata file note should read, “Missing one page.”

Table 2: Pre-ingest PDF preservation steps.

Utilitarian and pragmatic, this procedure is tremendously inexpensive because it relies on software already available at UNT Libraries, and it is an easy process to train new employees to do. While each step is important, the most time-consuming steps are where files are converted because these require a great deal of system resources. The most labor-intensive step is the image quality control step because a person must examine each page to verify that files have converted properly. Once files have been converted and described with metadata, they move into the optical character recognition (OCR) stage.

As already described, the OCR stage for newspaper PDF editions is no different from the stage for newspapers scanned from microfilm or from physical pages. As with all pages hosted in the Texas Digital Newspaper Program, once pages are OCRed, an xml and text version of the page is created, bounding boxes are made, and images are mapped to respective OCR files for full-text search. The PDF is the primary preservation file, which then is archived in the Portal. The JPEG file created of each page is a derivative that supports viewing.

8 PDF PRESERVATION: CURRENT PROJECTS, FUTURE DIRECTIONS

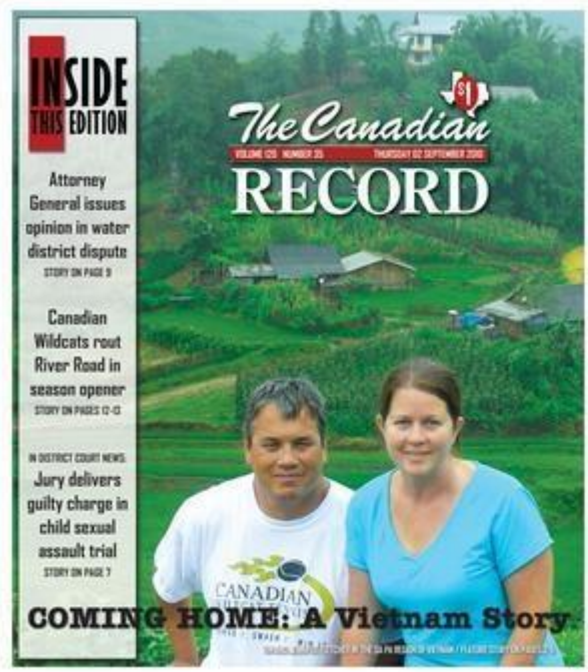
The Texas Digital Newspaper Program hosts multiple PDF projects. TDNP started preserving PDFs with one title, which represented a total of 5 years of PDF content. The Texas Digital Newspaper Program staff were first contacted by a publisher, Terrie Gonzales, from Rusk, Texas, in 2010, to preserve her PDF newspapers. *The Rusk Cherokeean Herald* currently embargoes its most recent three years, per contract with the publisher, who sells access to an archive those three years. Each year, the TDNP staff receive the prior most recent year, process and ingest it into the system, and then release the earliest embargoed year for public view. This embargo period is now an option that
TDNP can offer to all Texas publishers for the PDF content, modelled on the *Rusk Cherokeean* newspaper project.

Since this initial PDF project, libraries that work with TDNP to digitize earlier issues of their local newspaper runs also inquire about adding the most recent PDF content to the earlier content. Often, these libraries are motivated to preserve PDFs because their newspaper offices no longer provide a microfilm copy of the papers on an annual basis. Sometimes, libraries will report that their publishers have provided CD or DVD PDF files to them, but often, libraries do not receive any sort of newspapers for the more recent years, simply because filming can be a heavy monetary burden for either newspapers or libraries to handle. As a result, libraries closely coordinate with their publishers to accomplish the goal of preserving PDF content, and both libraries and publishers are often motivated to add the PDFs to previously-digitized issues because PDF conversion and preservation is relatively inexpensive compared to microfilming or scanning the physical pages.

Title	**PDF Start Dates**	**Newspaper Type**
Rusk Cherokeean Herald	2002	Weekly
Sweetwater Reporter	2010	Daily
Bastrop Advertiser	2007	Semi-Weekly
Canadian Record	2004	Weekly
Texas Jewish Post	2005	Weekly
Nocona News	2007	Weekly
Naples Monitor	2012	Weekly
The Dallas Voice	2004	Weekly
The Greensheet	2005	Weekly/Promotional
The NT Daily	2003	Daily/School
The University News	1998	Weekly/School
Texas Wesleyan Rambler	2008	Weekly/School
The J-TAC	2010	Weekly/School

Table 3: Current PDF titles accessible on The Portal to Texas History.

Most newspaper projects begin from analog digitization and move to PDF print master preservation, with the processes taking place one after the other. However, on occasion, publishers will begin with PDF conversion as their library simultaneously works to apply for grant funding to digitize and preserve analog content. In the case of student newspapers, school libraries often start by creating access to their PDF content, and use the newspapers that are accessible to generate financial support to digitize the remainder of analog content. Since first working with the PDF editions of the *Rusk Cherokeean*, TDNP has preserved and made accessible twelve more PDF newspaper projects. (See Table 3.) PDF preservation provides a low-cost alternative to microfilming from physical newspapers and scanning from film, which requires shipping newspapers to the microfilm vendor, having master film and duplicate film created, and then scanning from microfilm to reacquire into a digital format. Also, because the PDFs are full-color (see Figure 3), the object can be preserved as it was originally intended to appear. In contrast, microfilm loses the rich color detail that modern newspapers are printed with, thus losing significant information about the original primary source.



INSIDE THIS EDITION

Attorney General issues opinion in water district dispute
STORY ON PAGE 9

Canadian Wildcats rout River Road in season opener
STORY ON PAGES 12-13

IN DISTRICT COURT NEWS
Jury delivers guilty charge in child sexual assault trial
STORY ON PAGE 7

$1

The Canadian
VOLUME 120 NUMBER 35 THURSDAY 02 SEPTEMBER 2010
RECORD

COMING HOME: A Vietnam Story

Figure 3: The Canadian Reporter is one example of a PDF edition preserved on the Portal.

New Model for Born-Digital Newspapers in TDNP

A major challenge in the acquisition of digital newspaper content in the United States has been establishing the deposit of digital content as a systematic process that is efficient for publishers to use. Newspaper publishers have a wide variety of activities for which they are already responsible, and having another place to send their master PDF files increases an already busy workflow on the part of publishers. In order to work around this challenge, UNT Libraries began a partnership with the Texas Press Association and Newz Group in 2014 to streamline the depository process. Newz Group (http://newzgroup.com) provides a suite of media services to newspaper publishers throughout the U.S., including print monitoring, internet monitoring, and bids and leads. They provide press associations with services such as public and legal notices, archiving and electronic tear sheet services. The Texas Press Association partners with Newz Group to provide services for its members to build and maintain the Texas Press Archive, which boasts nearly 75% participation from newspapers in Texas in 2014.

Through this deposit process, newspaper publishers upload their PDF e-print editions to Newz Group as part of the Texas Press Association's arrangement for preservation services. In 2014, UNT Libraries, a Texas Press Association member, began working with the two organizations to investigate the possibility of obtaining these digital newspaper files to archive and preserve them within the Texas Digital Newspaper Program. Newz Group delivered the first set of digital newspapers to UNT Libraries in the Summer of 2014. This initial set amounted to more than 1.4 million pages of born-digital newspapers, representing over 500 titles in Texas. Now, TDNP staff has begun the process of contacting Texas publishers to secure permission to create access to these titles via The Portal to Texas History. In June 2014, TDNP's coordinator, Ana Krahmer, spoke at the Texas Press Association summer meeting, explaining how publishers could negotiate the embargo period (Texas Press Association, 2014). Multiple publishers expressed interest in creating access to their PDF issues, with only a short embargo period, and as a result, the embargo option has played an important part in the negotiation of rights with these publishers.

New workflows are also expected as part of this partnership. The high volume of digital content necessitates new approaches to the creation of archival packages and associated issue-level metadata. While there are a number of important issues to work out over the upcoming years with this project, publishers will benefit from PDF preservation, and TDNP staff foresees interest on behalf of publishers in digitizing analog content after they have worked out how TDNP will deal with their PDF content.

9 CONCLUSION

UNT Libraries has become heavily involved in preservation of PDF newspaper eprint editions for the simple fact that our partnering institutions have reported a crisis facing newspaper publishing at present: publishers have a lot of work and many financial burdens to deal with right now, and UNT Libraries has the capacity and infrastructure to ethically guarantee and engage in long-term preservation, for both analog and digital newspaper content. Because PDF preservation is relatively inexpensive, UNT Libraries can also save publishers some money and time, and thus assist in preserving community identity across Texas. The end goal of newspaper preservation via The Portal to Texas History is to benefit communities through strong preservation infrastructure and thorough long-term preservation planning. The Texas Digital Newspaper Program is committed to serving communities for the benefit of communities, no matter their size or geographic location, and no matter the size, date, or format of their newspaper collections. Preservation of PDF eprint editions is a natural next step for TDNP, with a goal being to make preservation easy and headache-free for

Texas publishers thus ensuring access to this important newspaper content for generations to come.

Acknowledgments

We would like to thank the National Endowment for the Humanities for their generous support of the National Digital Newspaper Program. In addition, we would like to thank the Tocker Foundation and the Ladd & Katherine Hancher Foundation, the Abilene Library Consortium, as well as the Texas publishers and libraries, without whom the Texas Digital Newspaper Program would not be possible.

10 REFERENCES

Abrams, S., Kunze, J., & Loy, D. (2010). An Emergent micro-services approach to digital curation. *International Journal of Digital Curation*, 5 (1). p. 172-186. Retrieved from http://ijdc.net/index.php/ijdc/article/view/154

Center for Research Libraries. (2011). Preserving news in the digital environment: Mapping the newspaper industry in transition. *A Report from the Center for Research Libraries.* Retrieved from http://www.digitalpreservation.gov

Heritage Archives. (2010). A Brief history of microfilm. Retrieved June 30, 2014, from http://www.heritagearchives.org/history.aspx.

Krahmer, A. & Phillips, M.E. (2013.) Laying the groundwork for newspaper preservation through collaboration and communication: The Texas digital newspaper program. *WLIC 2013 Satellite meeting.* Retrieved June 20, 2014, from http://digital.library.unt.edu/ark:/67531/metadc172339/m1/10/.

Krahmer, A. & Phillips, M.E. (2014). Research from many angles: Evaluating usage and impact of digital newspapers in the Texas Digital Newspaper Program. *IFLA International Newspapers Conference 2014.* Retrieved June 23, 2014, from http://digital.library.unt.edu/ark%3A/67531/metadc303227/

Library of Congress. (2014).The National digital newspaper program. (2014). Retrieved June 23, 2014, from http://www.loc.gov/ndnp

National Digital Stewardship Alliance. (2012.) Case Study: Newspaper e-prints. Retrieved June 21, 2014 from

http://www.digitalpreservation.gov/ndsa/working_groups/documents/NDSA_CaseStudy_NewspaperEPrints.pdf

Open Archives Initiative. (2014). ResourceSync framework specification. Retrieved June 22, 2014, from http://www.openarchives.org/rs/toc/

Texas Press Association. (2014). Krahmer to contact newspapers about archiving project. *Texas Press Messenger*. Retrieved July 8, 2014, from http://texaspress.com/index.php/messenger/2467-krahmer-to-contact-newspapersabout-archiving-project/

University of North Texas Libraries. (2013). IOGENE: The Interface optimization for genealogists project. Retrieved June 24, 2014, from http://digital.library.unt.edu/ark:/67531/metadc270771/

University of North Texas Libraries. (2014). The Portal to Texas history. Retrieved June 25, 2014, from http://texashistory.unt.edu/

University of North Texas Libraries (2014). The Texas digital newspaper program. Retrieved June 23, 2014, from http://texashistory.unt.edu/explore/collections/TDNP/