Submitted on: 19.01.2016

Preserving Kentucky's Newspapers: Analogue Beginnings to Digital Frontier

Kopana Terry
Curator of Newspapers & Oral History Archivist at University of Kentucky Libraries Special Collections Research Center, Lexington, Kentucky, USA
klterr0@uky.edu

Eric Weig
Digital Library Architect, at University of Kentucky Libraries Special Collections Research Center, Lexington, Kentucky, USA
eweig@uky.edu

Abstract:

Over fifty years ago an historian and a library director traveled the back roads of Kentucky (USA) with a portable microfilm camera, two lights, and a dream of preserving Kentucky's newspapers. From their ambitions arose a successful newspaper preservation program at the University of Kentucky Libraries (UKL). Now in its sixth decade, the program has developed a new way of preserving contemporary born-digital newspapers. This paper explores some of the people and events behind the early success of UKL's program, as well as an in-depth look at the development and functionality of Paper Vault: a largely automated, in-house process delivering and preserving Kentucky's born-digital contemporary newspapers.

Keywords: newspapers, preservation, born-digital, NEH, NDNP, microfilm, Paper Vault, NDNP, curate, Kentucky, harvest, automation, KDNP, meta | morphosis

Our Story Begins

Setting The Stage

The use of 35mm microfilm as a long-term, stable medium for newspaper preservation was first adopted in the 1930's by the New York Times. (15) It quickly proved a viable medium for preservation and access. Before long, commercial microfilm companies opened for business, and newspaper microfilming was off and running. Sitting on pallets of crumbling newspapers, libraries were quick to jump on the microfilm bandwagon. UKL - the state's flagship University Library - was one of the earliest adopters of microfilm for preservation.

Pioneers

It was the 1940's. Dr. Thomas D. Clark (1904-2005), and Dr. Lawrence S. Thompson (1916-1986), began traveling the backroads of the Commonwealth in search of publishers willing to let the pair photograph their collections. Their setup wasn't fancy. They used a compact, collapsible microfilm camera and copy stand with two household incandescent bulbs for lighting. The images were terrible by today's standards, but they were good enough to save something of the newsprint.

Clark and Thompson were very different men, and their approach to collecting history differed as well, save for newspapers.

Thomas Dionysius Clark lived to be 101 years old. He taught history at the University of Kentucky from 1931-1965 and was named Kentucky's Historian Laureate for life in 1990. (18) He authored 40 books during his life, including "The Southern Country Editor," an exquisite tome about newspaper publishing in the American South. Dr. Clark was, some would say and quite by accident, almost solely responsible for the creation of the Kentucky Department of Libraries and Archives (KDLA). The story goes that, one night, Clark saw valuable state records being thrown into a dumpster outside a government building. He borrowed a friend's truck and pulled the records out of the garbage that same night. Thus, he began to lobby for a permanent state supported archive. He was successful, and those records he saved became the nucleus of KDLA's collection.

Lawrence Thompson loved books, but he also loved adventure. He traveled extensively through Europe, Asia Minor, and the Caribbean as librarian and professor. (21) It has been suggested that, of the two men, Thompson was the driving force behind the microfilm excursions through Kentucky. His passion for microformats was something more than simple adoration. He found the preservation properties of microfilm compelling, and not just for newspapers. He created Lost Cause Press; a publishing company of books, of course, though the press became best known for its microfilming of rare books. (17)

Set in Stone... sort of

Evidence suggests a microfilming facility was on UK's campus as early as 1949. (23) By 1950, the University was accepting historic newspapers from the state law library specifically for preservation microfilming. (25) Around 1954, Clark and Thompson collaborated with the Kentucky Press Association (KPA) and the University of Kentucky (UK) School of Journalism to build a formal center for newspaper microfilming on UK's campus. Housed in the School of Journalism, the Microfilm Center managed the collection and preservation of more than 200 of the Commonwealth's contemporary newspapers. It is believed the program moved physically and administratively to UKL facilities in the 1960's, where it has remained.

Around 1956, KDLA, the Kentucky Historical Society (KHS), and UKL agreed that UKL would assume custodial responsibility for the state's newspaper preservation. Though no formal record of this agreement can be found at the time of this writing, the University's Board of Trustees minutes from 1956 imply such a mandate, if not between these exact agencies, then between the University and the Library.

> "*But over and above the requirements of teaching and research, the university has an obligation to join with other agencies in the preservation of the books and other material evidences of the achievements of civilization...State and regional materials will receive special emphasis...to make the fullest use of microfilming and other*

> *money saving devices*…[under 'Obligations to Society'] *to maintain our cultural heritage.*" (24)

UKL was the State's logical choice for newspaper preservation because of its experienced cataloguing staff and conservation tools. Most importantly, UKL had over a decade of newspaper microfilm experience at this point. For the next fifty years this mandate, implied or otherwise, held true, and the state's publishers, public libraries, archives, historical societies, genealogists, researchers, and the general public came to rely on UKL for access to, and preservation of, Kentucky's newspapers.

Up and Running

UKL's newspaper microfilm program was by no means perfect. For that matter, microfilm production wasn't perfect anywhere. There were no standards until the National Micrographics Association (NMA) developed guidelines in the early 1970's. (19) Microfilm shops made up their own rules for targeting, content, and quality. It was the Wild West of microfilm: everybody had a camera, and nobody was afraid to use it. This would, of course, come back to haunt everyone when newspaper digitization came along in the 2000's.

In 1981, preservation microfilm took a giant leap forward when the Association of Information and Image Management (AIIM) released MS23, "*Standard Recommended Practice: Production, Inspection, Quality Assurance of First Generation Microforms of Documents*." These guidelines became a major force in modernizing microfilm production and remains relevant today. Nevertheless, its widespread adoption was far from universal.

Also in 1981, UKL's microfilm operation, now known as The Reformatting Center, became one of the first five participants in the National Endowment for the Humanities' (NEH) United States Newspaper Program (USNP). During USNP, UKL catalogued 5,000 Kentucky newspaper titles, microfilmed approximately 1.5 million pages, and expanded the microfilm collection to nearly 20,000 reels. USNP bolstered newspaper preservation at UKL, and laid significant groundwork for future federal grant funding.

Also during USNP, additional microfilming standards were developed and older guidelines refined. Though some best practices existed before USNP and others came along during the program, not all awardees followed them. Microfilm shops had used their homegrown guidelines for so long that they preferred to keep using them. And so they did, including UKL. That's not to say guidelines were completely ignored, they weren't entirely, but changing behaviors often need motivation to do so. Enter Rebecca "Becky" Ryder.

A Brand New Day

Becky Ryder started at UKL as a Graduate Assistant in 1991 with a focus on book conservation. She was hired in 1992 as UKL's first full-time Preservation Librarian for its relatively new Preservation Department.

Ryder knew nothing about making microfilm. So, she enlisted the help of Bob Mottice, master microfilmer at Bell and Howell, which is today ProQuest. Mottice confirmed there were few microfilm shops following standards. His answer to the problem was a workshop called "Preservation Microfilm: The Silver Standard." Like Clark and Thompson, Mottice found a kindred spirit in Ryder. The pair teamed up, and for more than a decade, they taught hundreds of microfilmers around the country the quality production, care, and preservation of microfilm.

Naturally, the Reformatting Center, too, began to create microfilm by the book. They phased out acetate film, used consistent targeting, and proper development of the film was stressed. The result of these changes were immediate, and UKL's Preservation Reformatting Center became one of the most well respected microfilm shops in the U.S. Ryder, too, became a leader in the national preservation community, and was the first awardee of the George Cunha and Susan Swartzburg Preservation Award in 2008.

Incidentally, Dr. Clark was still active on campus and in archival circles when Ryder joined UKL. He championed the modernization of the Reformatting Center, and Ryder drew on his pioneering spirit more than once.

By 1998 the Preservation Reformatting Center was a well-oiled machine. With a small army of student workers, and a host of well-trained filming technicians, Preservation Reformatting was preserving thousands of newspaper pages each year. Kentucky's modern newspapers were the heart of the Center's operation, though historic issues trickled in after USNP ended in 1991. Students would split the pages, iron out creases using common household irons (low heat, of course), and then collate each title for filming. Every effort was made to photograph an issue in its entirety, including supplementary inserts.

With UKL's on-site cold storage, microfilm produced under Ryder's tutelage is expected to last up to 500 years. Such microfilm expertise would prove priceless when USNP's successor, the National Digital Newspaper Program (NDNP), came to be in 2005.

Diving Into Digital

At the same time Preservation Reformatting was going gangbusters in 1998, 42% of American households had personal computers, and 26% of those had Internet access. (16) The demand for online access of library collections was on the rise, and cultural institutions around the country were feeling the push. To say newspaper digitization was in its infancy at this time would be an understatement. Newspapers, with their columnar structure and large physical size with small type, were extremely complicated to digitize *and* make searchable in a meaningful way. Their digitization was more a point of research than anything tangible for the larger preservation community. For digitization novices, newspapers definitely weren't a prudent place to break ground. Enter Eric Weig.

You Have To Start Somewhere

UKL dipped its toe into the digitization pool for the first time in 1998. Eric Weig was hired out of Library School to lead an endeavor known simply as "E-text." Things would change fairly soon after Weig came aboard, however. The Kentucky Virtual Library (KYVL), lead by the State-Assisted Academic Library Council of Kentucky (SAALCK), comprised of deans from the eight state-assisted university libraries, called for online access to special collections materials held by Kentucky academic libraries. KYVL partnered with UKL, with Weig in the directorial chair, to build and maintain the Electronic Information Access and Management Center (EIAMC). They (Eric, actually) would digitize and make accessible online historical materials from around the state. It wasn't long before EIAMC became the Kentucky Digital Library (KDL) (and Eric got a staff). For the next fifteen years Weig and UKL managed the KDL infrastructure, digitizing archival material from around the commonwealth.

To build the KDL, Weig recruited some of the nation's leading digital library experts; Cornell's Peter Hirtle (copyright), Duke's Stephen Miller (encoding), Digital Library Federation's David Seaman (standards and imaging), and the California Digital Library's Roy Tennent (library structure and metadata). By 2001, the first photographic collections and archival finding aids were available online and things began to move quickly.

In 2002, Weig and Ryder teamed up for a grant from the Institute for Museum and Library Studies (IMLS) to digitize from microfilm rare and imperiled Kentuckiana books for a project called "Beyond the Shelf". This project gave them valuable experience in microfilm-to-digital methodologies that would prove fortunate two years later when, in 2004, the National Endowment for the Humanities (NEH) called for proposals for NDNP.

NDNP and meta|morphosis

With Weig and Ryder at the helm of NDNP, the pair assembled a team of specialists from within UKL to assist: Enter Kopana Terry. After a career in the music business, Terry earned her BA in studio art photography. In 2001, she started with UKL in reformatting, and played a key role on the KY-NDNP team before taking over as Program Manager in 2007.

Kentucky and NDNP

In 2005, NEH chose UKL as one of only six participant states for NDNP. In 2013, we completed our fourth and final grant (each grant cycle is two years) with 78 historic Kentucky newspapers represented in Chronicling America, NDNP's online database.

NDNP prefers that each state's newspaper of record be among the first titles digitized. The unanimous choice among the KY-NDNP advisory board - made up of historians, librarians, and journalists - was Louisville's *Courier Journal* (CJ). NDNP specifications require image scans be made from a second generation (print master) negative only. UKL had not microfilmed the CJ. The commercial company that created the master microfilm, and subsequently held the right to duplicate from it, quoted UKL a six figure sum for 80 print master reels comprising only 1900-1910, the date range for the first NDNP grant cycle (2005-2007). Needless to say, that purchase was out of the question.

With over 450 other titles to choose from, we selected titles to represent the six geographic regions of the Commonwealth. (20) Among the choices were labor union, temperance, African-American, and orphaned titles. Thirty-seven (37) Kentucky titles were digitized that first grant cycle; more than any other awardee at that time.

Ironically, just a year before NDNP began, UKL, citing budget re-evaluation, announced to Kentucky's newspaper publishers a one-year suspension of microfilming operations. The 'hiatus', as it was termed, was "met with outrage", and letters from publishers across the Commonwealth poured into UK President Lee Todd's office.

> "*If newspapers have served no other purpose, they have all served as the historical record of Kentucky's history. From the smallest communities and counties to the Commonwealth as a whole, newspapers have, do and will record Kentucky's history.*" wrote KPA President David Thompson. "*The Kentucky Press Association worked closely with the University of Kentucky library in the mid-1980s on the Newspaper Preservation Act* [USNP]...*to record more of Kentucky's history. We would welcome the opportunity to work with you, the administration or personnel at the library to ensure the microfilm service continues, at least for the immediate future if not long-term.*" (22)

The pushback was so severe that the Reformatting Center's hiatus, which could have resulted in permanent closure, reopened as promised, and just in time to provide a critical service to NDNP. The temporary closure, however, was a harbinger of things to come.

It Could Have Been a Disaster

What set Kentucky apart in NDNP wasn’t the large number of digitized titles, but our choosing an entirely in-house production methodology. Every page image was created on the University of Kentucky campus — from microfilm duplication to the deliverable data package. It was a time and labor intensive approach that no other awardee could have done at the time. Armed with a heap of that pioneering spirit, we had no doubt we could pull it off. Everyone else thought we were crazy.

We licensed software from digitization vendor iArchives, and duplicated their infrastructure in our digital lab. We then hired a small army of student workers to perform column zoning, a task since automated. By doing the work ourselves we learned where our librarian skills were most useful in the process. It was the metadata.

Newspapers on microfilm can be complicated. Remember those non-existent standards? The lack of unstandardized microfilm reared its ugly head during NDNP. Even the USNP film proved challenging at times. No institution’s USNP microfilm could meet NDNP expectations, and this was but one of the program’s growing pains.

By the end of our second award cycle, we had proven librarians had it over the vendors concerning newspaper metadata. Meanwhile, many new awardees had come aboard, and all of the awardees, including the Library of Congress, were interested in our in-house advantage. We had not suffered the same qualities control issues they had, although we had our own problems such as a lack of IT support and enormous administrative undertaking with that army of students.

Working with iArchives during this time, we helped develop a new production methodology known as the *hosted-hybrid* model. This production model took advantage of iArchives' technical infrastructure in Utah as well as their offshore workforce. We continued the microfilm duplication, image capture, and metadata entry from the UK campus while iArchives did the heavy lifting. It was a rousing success, and other awardees soon adopted the method. Shipping and labor costs dropped significantly. The advantage for iArchives was no longer reworking errored batches; a major cost savings for them. Other digitization vendors were eager to offer a similar service, and they, too, developed systems to allow awardees to perform tasks from their office desktops. It transformed the quality of the final data packages, workflows, and budgets.

Simultaneously, we were experimenting with color imaging. Through a partnership with the University of Louisville, which held a large cache of hard copy CJ, we digitized nearly 20,000 pages of the title in full color. Many turn of the 20th Century issues carried four-color advertisements and line art. We worked with the Library of Congress to include this color content, but the image file size virtually tripled storage space needs. For a program to include all fifty states and U.S. territories, space was a valuable commodity. Chronicling America simply couldn’t contain color page images. Concurrent with this, we teamed with the Lexington Public Library to digitize the Kentucke Gazette, the first newspaper published west of the Allegheny Mountains. The earliest issue printed in 1787 fell outside NDNP’s 1836-1922 date range. It could not be included in Chronicling America. These titles, and several others, were instead made available on the KDL, and today they’re available in the *Kentucky Digital Newspaper Program* http://kdnp.uky.edu.

When Crazy Proved Completely Sane

Doing everything in-house taught us how to digitize newspapers top to bottom. The first year in NDNP we were getting calls almost daily asking how to do it all in-house. Rather than answer the same questions time and again, we developed a two-day seminar that would take participants through the process beginning to end. NEH and the Library of Congress were keen to the idea, and thus *meta |morphosis: a university of kentucky microfilm-to-digital institute* was born. From 2006 through 2010 participants came from all over the world to the University of Kentucky. Our KY-NDNP team developed lectures and hands-on activities for imaging, metadata, microfilm inspection, infrastructure

- even how to write an RFP. During our last NDNP award (2011-2013), Kopana turned all of those lectures into a set of self paced online video tutorials that are today freely available from NDNP home page and UKL http://www.uky.edu/Libraries/NDNP/metamorphosis/index.html .

The End of an Era

In 2008, the U.S. experienced a devastating recession, the effects of which were keenly felt in Kentucky, and the University, the following year. The Preservation Reformatting Center was down to filming 152 Kentucky newspapers by this time. Newspapers had been suffering circulation losses for many years, due almost entirely it seems to online news outlets. (25) The economic collapse didn't help.

From Victory to Collapse

It wasn't a single event that caused the demise of the newspaper preservation program at UKL. Rather, it was a perfect, deadly storm; equal parts economics, administrative anxieties, and programmatic over-extension. In some ways, you might say we were a victim of our own success.

The beginning of the end started with the final closure of the Preservation Reformatting Center in 2010, which by this point had been rolled in with UKL's digital initiatives to become Preservation and Digital Programs. By University rules, the center couldn't make too much money from its microfilm duplication sales. It also couldn't show a loss. It was a slippery slope in the best of times. The center reopened after its yearlong hiatus in 2004, but with the economic collapse, pressures to save every dime across campus had built to fever pitch. Microfilm was seen by an administration grasping for every dollar as an expendable extravagance. This time, the closure was met with minimal defiance from publishers, in part because they were suffering financially, too, but also because we were developing an alternative to microfilm called Paper Vault.

At the same time Preservation was being dismantled, the demands on Digital Programs were growing. The success of NDNP had proven that we could successfully manage remarkably large projects. Pressure was building from institutions around the state to have their content included in the KDL. UKL's Special Collections was equally eager to digitize their collection. Several national digitization grants were awarded throughout UK's Special Collections division based, in no small part, on the success of KDL and NDNP. A team of programmers were hired as more digitization projects and proposals were added to the Digital Programs plate.

In late 2010, the same year the Preservation Reformatting Center closed, NEH announced that NDNP was to be a limited program for the states. To this point, we understood the program to be a twenty year program for all parties. Naturally, the collapse of 2008 left NEH with tough choices to make, too. And as they took more states into the program, demands on the Chronicling America infrastructure were rising, making sustenance difficult. Though we were given a rare fourth grant (2011-2013), the loss of NDNP funding was devastating to UKL's newspaper program.

By 2013 the NDNP funding was spent. The Preservation Reformatting Center had closed. The state had continued to suffer from the 2008 meltdown, and the KDL funding was drying up. The team of programmers were being stretched so thin by new and unfinished projects that they couldn't keep up with new data, least of all large data sets like newspapers. Kopana, who had managed all of the newspaper curation and production over the last eight years, was transferred to the Louie B. Nunn Center for Oral History, where she had helped develop audio digitization with Weig back in 2005. That was essentially the end of newspaper preservation at UKL. Or was it?

The Comeback

So much energy had been spent over the decades to preserve Kentucky's historic newspapers that it was disastrous to the historical record to be without an active newspaper preservation mechanism. By the same token, Weig and Terry had invested a decade of their professional lives to the preservation of UKL's newspaper collections. It became personal, perhaps in the same way it had been personal for Clark and Thompson sixty years prior.

In 2010, when Weig and programmer Michael Slone first built Paper Vault, we turned once more to the KPA for help in harvesting the print-ready PDFs being deposited with NewzGroup; an agency that acts a bit like the old fashioned clipping services for legal notices. Once the agreement for harvest from NewzGroup was in hand in 2012, Weig and Slone began harvesting Kentucky's newspapers. UKL has been harvesting Kentucky's newspapers (approximately 135 as of this writing) from NewzGroup ever since.

The problem then became two-fold: how to provide access to the pages, and do so without money.

First, we did not have rights to display the newspapers. Preservation aside, instant global access to news content is vastly different from making microfilm for access: a significantly slower process. Copyright in the Internet age becomes a whole new animal. Newspapers were already struggling with circulation problems, and financial failings. They didn't need another problem, nor did we.

Second, UKL had zero money for storage, and zero staff to manage the collection even if we did. Both Weig and Terry had been redirected elsewhere in the library. But, library users, and the reference librarians who served them, were increasingly frustrated by the lack of access to Kentucky's current newspapers. It was a growing problem in UKL's fundamental mission to provide free and open access to Kentucky's historical documents.

The stars aligned in the fall of 2014. UKL Special Collections Associate Dean Deirdre Scaggs knew there was a growing need for newspaper curation. She tapped Terry's curatorial expertise, naming her Curator of Newspapers. This allowed for more active engagement with patrons and the newspaper community at home and abroad. Serendipitously, during the first *Dodging the Memory Hole: An Action Assembly* meeting in Missouri, Weig met with representatives of the Internet Archive and hatched an idea that would realistically get the Preservation Newspaper Program back up on its feet. It starts with Paper Vault and utilizes the Internet Archive for storing newspaper page images.

Paper Vault

The idea for Paper Vault was sparked by a desire to save time and money but not sacrifice quality with the processing of born digital newspapers for digital preservation and online access. A sustainable model for digital preservation and access was needed, one that could replace the costly microfilm process and the even more costly NDNP process. The solution would be to automate as much as possible: to use computer time rather than human time to do the majority of the heavy lifting.

Paper Vault is a framework and workflow of strategies, standards, and open source software tools for automating the processing of born digital newspapers for preservation and access. It was developed in-house at UKL by Eric Weig and Michael Slone and continues to evolve. It relies on the emerging NDNP-Lite metadata standard, and is meant to be adaptable to local constraints concerning staffing and technological acuity, allowing for a little or a lot of metadata, and simple or sophisticated technological infrastructure.

Content Harvesting

When we began, there were only a handful of newspaper publishers contributing content. We tested harvesting from some of these publishers directly with mixed success, but then we were able to take advantage of an arrangement the publishers had with the NewzGroup media technology company, to which the publishers were already contributing their content for the purposes of preserving legal notices. This worked much better. Publishers submitted their content once instead of multiple times. As a result, the number of publishers contributing to Paper Vault quickly scaled to over 100 titles. These contributions were once again supported through the cooperation of the KPA. They opened channels of communication with the publishers and established agreements for the harvesting.

How We Do It

Paper Vault content harvesting is automated. Each night during established windows of time, one of our UKL servers connects to the NewzGroup server holding the newspaper files. Night harvesting limits technological drag on the network, which could adversely affect other work being done on the systems at either end.

Files are gathered and arranged in a directory structure that organizes the content first by publisher, title, and then specific issue date. Both the publishers and the titles in the system have unique identifiers that are five digit numbers. For example, the *Licking Valley Courier* is 70196. These numbers are used for high-level directory naming conventions. The issue dates are then used on subdirectories in a YYYY-MM-DD format to identify specific issues.

Our mission is both preservation and access. Much of our approach can be credited to our experience working within the NDNP. Considering that the NDNP Guidelines are often regarded as the standard for newspaper preservation, our goal is to provide as similar a set of deliverables as possible just using the born-digital PDFs as the master format instead of microfilm. [4, 11]

The parts of the NDNP package that we centered on emulating within our born-digital processing primarily deal with the file set. For NDNP, these are comprised of a TIF image, Publisher, Title and Issue Level metadata, and full-text with bounding box data. All of this content is contained within a Bagit structure. [2]

The following derivative files are produced during our processing of the PDF content.

- An uncompressed TIF image in 8-bit RGB color space, keeping the DPI as it is set within the PDF file. [9, 14]
- A METS file constructed by extracting metadata for Publisher, Title and Issue from the Paper Vault MySQL database.
- Bounding box data derived. Here, we developed two methods.
    - One is handled as a part of the ingest process within the Internet Archive.
    - The second we developed locally to extract full-text and bounding box data from the PDF. We do this using Apache PDFBox and a custom PDF2ALTO converter written in-house to form an ALTO XML file for each newspaper page image. [1, 3, 12]

This last point, the format of the locally derived ALTO file, is one significant difference between the born-digital data set vs. the NDNP data set for historic newspapers. It is a conversion to ALTO from Adobe Bounding Box data extracted from the PDF using Apache PDFBox vs. an ALTO file created during Optical Character Recognition as with the historic issues. [1] This way, with the born-digital content, we offer better searching of the full-text that is not riddled with the inaccuracies of OCR technology. [5]

It is important to note, however, that there are some major differences between the NDNP ALTO and Paper Vault ALTO files. These differences are due to the limited amount of bounding box data encapsulated within our PDFs. They do not provide bounding boxes for the Page, PrintSpace, TextBlock, or TextLine, but only one for each page and for individual words, no matter how the individual Strings on the Page are arranged. [3, 12] Still, use case trials proved that word-level bounding box data was ‘good enough’ for search hit highlighting within an access system. [3]

Serial Metadata Store

The serial metadata store is comprised of a simple database holding serial level metadata for newspaper titles. It allows metadata extraction via an API so that the metadata can be gathered in an automated fashion. For UKL’s implementation of Paper Vault, we use a simple Omeka database hosted on omeka.net.

The metadata kept in the serial metadata store minimum description set as outlined for the NDNP-Lite draft metadata application profile. [23]

Issue Level Metadata Store

The Issue Level Metadata store is a store of static metadata files or simple database holding issue level metadata. It allows metadata extraction via an API or HTTP. Storage format for static metadata files can adhere to local preference. At UKL, we use JSON formatted issue level metadata files.

The metadata kept in the issue level metadata stores minimum description as outlined for the NDNP-Lite draft metadata application profile. [23]

The following directory structure is utilized at UKL and is described here as a model for others storing issue level metadata files.

```
project
¦
¦
+---ada
    ¦
    ¦
    +---collections/
    ¦    ¦     metadata.json
    ¦
    +---lccn1/
    ¦    ¦
    +     issues/
    +            ¦
    +            YYYY/
    ¦           ¦     adaYYYYMMDDED.json
    ¦           ¦     adaYYYYMMDDED.json
    +            YYYY/
    ¦           ¦     adaYYYYMMDDED.json
    ¦           ¦     adaYYYYMMDDED.json
+---ame
    ¦
    ¦
    +---collections/
    ¦    ¦   metadata.json
    ¦
    +---lccn1/
    ¦    ¦
```

```
+     issues/
+          ¦
+          YYYY/
¦          ¦    ameYYYYMMDDED.json
¦          ¦    ameYYYYMMDDED.json
+          YYYY/
¦          ¦    ameYYYYMMDDED.json
¦          ¦    ameYYYYMMDDED.json
+---lccn2/
¦    ¦
+     issues/
+          ¦
+          YYYY/
¦          ¦    ameYYYYMMDDED.json
¦          ¦    ameYYYYMMDDED.json
+          YYYY/
¦          ¦   ameYYYYMMDDED.json
¦          ¦   ameYYYYMMDDED.json
 ...
```

Page Image Host

The page image host is an HTTP accessible store for page images concatenated as newspaper issues. It can be a local storage space or a hosted storage space. It can be simple, for example holding only the PDF versions of the newspaper issues, or it can be more sophisticated utilizing a local JP2 server, for instance, or image tiling utilities such as OpenSeadragon or OpenLayers. [21, 22] These choices are left up to local managers. At UKL, we utilize the Internet Archive as a page image host. Images are served via a JP2 server and offer much of the functionality available within the Chronicling America site. The Internet Archive’s approach to digital preservation and open access to information complements our own. [24]

Metadata Wrangler

The metadata wrangler is an application utilized to gather metadata from the serial and issue level metadata stores, format it for preservation and access, and then index it within a content management system. At UKL, our metadata wrangler for Paper Vault is PHP and formats as JSON for online indexing, and METS for preservation storage.

Content Management

Indexing the content does not require any specific software application. At UKL, we use an in-house implementation of the Blacklight system, which utilizes Apache Solr for indexing. However, the data set produced through the Paper Vault process could be indexed in any number of content management systems.

Quality Control

As with any digitization activity, quality control (QC) is an important aspect of the process. Particularly important for our QC workflow is the use of a secure development version of our interface. This allows for timely QC of issues under an embargo period, so that we can identify missing pages or other anomalies early.

We automate much of the processing of the pages and then verify their correctness before releasing issues for general online access. Many of the issues coming in are individual page images vs. compound issue level PDFs. We analyze file names for proper page order while creating the compound PDF. This is done by creating a text list of the filenames, running a rename on that list to ensure that the part of each filename containing a sequence number is expressed with the same number of digits. If not, corrections are made to the list and it is then used to feed the PDF merge utility that creates the issue level PDF.

After processing, correct metadata, corrupt pages, accurate bounding boxes, and proper image display are also checked. Although problems arise in all of these areas, we have found that the predominant snags have been page order correctness due to file naming conventions that don't allow proper sorting by the file system. We continue to improve the Paper Vault system's abilities to identify and deal with this issue on a title by title basis.

Some Statistics on Gathering the Content

We began gathering content in June of 2011 and began reviewing numbers concerning our harvests periodically after that. The following chart describes content harvested within a window from June of 2011 through August of 2013.

Total number of pages	24,902
Total PDF	21.58 GB
Total TIF	121.35 GB
Mean PDF size	.89 MB
Mean TIF size	4.99 MB

Moving Forward

Meanwhile, the practice of news outlets is rapidly changing. We're no longer dealing with newspapers alone, or even born-digital PDFs, but rapidly changing, multi-platform, highly stylized and proprietary web news delivery. Even social media like Facebook and Twitter, are by some considered the first draft of history. One need only refer to the Arab Spring uprising fueled almost entirely through social media to understand this. We're living in a new era of news media. Paper Vault needs to be malleable enough to adapt and encompass these new formats. This is why Paper Vault is more of a framework to build upon rather than just one box to put things in. Since it's creation, Paper Vault has continued to evolve and continues to do so as we refine and enhance our process with an emphasis on automation and efficiency. There is a lot of data to process, and we are moving toward a goal to get all of our titles up to date online. Research has also started in regard to born digital newspapers in hypertext format. One of our core goals is to have ALL Kentucky newspapers searchable in one interface. Right now, https://kdnp.uky.edu offers access to page image based historic newspapers derived from analog as well as born digital sources. Soon, hypertext based news will have to be a part of Paper Vault as well.

Conclusion

UKL's newspaper preservation program has survived trials and tribulations for over sixty years through the determination of dedicated individuals. The Library has pulled itself up by the bootstraps to keep the program alive, providing a service to researchers, genealogists, and the greater public by preserving Kentucky's first records of history for future generations. Today, UKL has made a way to create a largely automated process for born-digital newspapers. We were able to successfully employ technology already well established within the digitization of historic newspapers, specifically within the NDNP. There are still aspects to improve upon with our process, but, we continue to move forward as we make improvements.

Acknowledgments

The authors wish to thank the following folks for their wisdom, guidance, assistance, brain trust, and general awesomeness in preserving Kentucky Newspapers: KPA President David Thompson, Deirdre Scaggs, Michael Slone, Judy Sackett, Digital Library Services, and the entire KY-NDNP team.

References

1. "Apache PDFBox - A Java PDF Library." *Apache PDFBox*. N.p., n.d. Web. 05 Sept. 2014. <http://pdfbox.apache.org/>.
2. "Bagit: Transferring Content for Digital Preservation." - Multimedia. N.p., n.d. Web. 08 Sept. 2014. <http://www.digitalpreservation.gov/multimedia/videos/bagit0609.html>.
3. "Cokernel/pdf2alto." *GitHub*. N.p., n.d. Web. 05 Sept. 2014. <https://github.com/cokernel/pdf2alto>.
4. Congress, Library Of. *National Digital Newspaper Program 2013 Technical Guidelines for Applicants* (2013): n. pag. Web. <http://www.loc.gov/ndnp/guidelines/NDNP_201315TechNotes.pdf>.
5. "Digitizing Microfilm and Optical Character Recognition (OCR)." - *National Digital Newspaper Program (Library of Congress)*. N.p., n.d. Web. 05 Sept. 2014. <http://www.loc.gov/ndnp/guidelines/digitizing.html>.
6. LaFrance, Adrienne. "Why the British Library Is Spending $55 Million on News Archives." *The Atlantic*. Atlantic Media Company, 29 Apr. 2014. Web. 05 Sept. 2014. <http://www.theatlantic.com/technology/archive/2014/04/the-british-library-is-spending-55m-on-news-archives/361346/>.
7. National Digital Stewardship Alliance Working Group (2013). "Case Study: Born-Digital Community and Hyperlocal News" 2013. <http://www.digitalpreservation.gov/ndsa/working_groups/documents/NDSA_CaseStudy_CommunityNews.pdf>.
8. Nick Krabbenhoeft, Katherine Skinner, Matt Shultz, and Frederick Zarndt. "Chronicles in Preservation: Preserving Digital News and Newspapers" 30 July 2013 <http://library.ifla.org/242/1/146-krabbenhoeft-en.pdf>.
9. "PDF Standards." - *PDF Reference*. N.p., n.d. Web. 05 Sept. 2014. <http://pdf.editme.com/PDFREF>.
10. "Purdue E-Pubs." - *Charleston Library Conference: The Future of Online Newspapers*. N.p., n.d. Web. 05 Sept. 2014. <http://dx.doi.org/10.5703/1288284314878>.
11. Skinner, Katherine, and Matt Schultz. *Guidelines for Digital Newspaper Preservation Readiness* (n.d.): n. pag. Web. <http://metaarchive.org/public/publishing/Guidelines_for_Digital_Newspaper_Preservation_Readiness.pdf>.

12. "The PDF Page Boxes: MediaBox, CropBox, BleedBox, TrimBox & ArtBox."*Prepressure The PDF Page Boxes MediaBox CropBox BleedBox TrimBox ArtBox Comments*. N.p., n.d. Web. 05 Sept. 2014. <http://www.prepressure.com/pdf/basics/page-boxes>.

13. Thomas, Deborah. "The National Digital Newspaper Program: Enhancing Access to American Newspapers" - *American Libraries Association Mid-Winter*, January 2005. <http://www.loc.gov/ndnp/guidelines/docs/NDNP_ala0105.ppt>.

14. "TIFF." *Developer Resources*. N.p., n.d. Web. 05 Sept. 2014. <http://partners.adobe.com/public/developer/tiff/index.html>.

15. Daavid, Joel. "History of Microfilm" - University of California Southern Regional Library Facility, March 2005. http://www.srlf.ucla.edu/exhibit/text/BriefHistory.htm

16. U.S. Census Bureau. "Home Computers and Internet Use in the United States: August 2000" - U.S. Government, September 2001. https://www.census.gov/prod/2001pubs/p23-207.pdf

17. Allen Kent, Harold Lancour, Jay E. Daily. Encyclopedia of Library and Information Science: Volume 13. CRC Press, 1975.

18. Saxon, Wolfgang. "Thomas Clark, 101, Chronicler of the History of Kentucky, Dies" - New York Times, June 30, 2005. <http://www.nytimes.com/2005/06/30/us/thomas-clark-101-chronicler-of-the-history-of-kentucky-dies.html?_r=0>

19. "MS23-1998, Standard Recommended Practice: Production, Inspection, Quality Assurance of First Generation Microforms of Documents" - Association for Information and Image Management International. <https://law.resource.org/pub/us/cfr/ibr/001/aimm.ms23.1998.pdf>

20. Terry, Kopana. "The Kentucky Edition Is Born" - NDNP, The Kentucky Edition, 2011. <http://www.uky.edu/Libraries/NDNP/introduction.html>

21. Herald-Leader Staff Report. "Lawrence Thompson, Ex-Director of UK Libraries, Dies At Age 69." Lexington Herald-Leader (KY), April 21, 1986.

22. Thompson, David. Letter to UK President Dr. Lee T. Todd, Jr. June 11, 2004.

23. Minutes of the University of Kentucky Board of Trustees, 1949, page 25, section R. <http://exploreuk.uky.edu/catalog/xt7w6m332r4h_24>

24. Minutes of the University of Kentucky Board of Trustees, 1956, page 22, "The Library and Museum Function". <http://exploreuk.uky.edu/catalog/xt7tdz02zv3x_22>

25. Perez-Pena, Richard. "Newspaper Circulation Continues To Decline Rapidly." New York Times, October 27, 2008. <http://www.nytimes.com/2008/10/28/business/media/28circ.html?_r=0>

21. OpenSeadragon <https://openseadragon.github.io/>

22. OpenLayers <http://openlayers.org/>

23. "Draft Newspaper Metadata Applicaton Profile", June 2015, Newspaper Digitization Interest Group (NDIG). <https://sites.google.com/site/digitalnewspaperspractices/technical-specifications/metadata-specfication>

24. "Internet Archive: About IA", July 20, 2015, <https://archive.org/about/>