Digital Preservation of Newspapers: Findings of the *Chronicles in Preservation* Project

Katherine Skinner
Educopia Institute
1230 Peachtree St., Su 1900
Atlanta, GA 30309
404-783-2534
katherine@metaarchive.org

Matt Schultz
MetaArchive Cooperative
1230 Peachtree St., Su 1900
Atlanta, GA 30309
616-566-3204
matt.schultz@metaarchive.org

Martin Halbert
University of North Texas
1155 Union Circle #305190
Denton, TX, 76203
940-565-3025
martin.halbert@unt.edu

Mark Phillips
University of North Texas
1155 Union Circle #305190
Denton, TX, 76203
940-565-2415
mark.phillips@unt.edu

ABSTRACT

In this paper, we describe research led by Educopia Institute regarding the preservation needs for digitized and born-digital newspapers. The *Chronicles in Preservation* project, builds upon previous efforts (e.g. the U.S. National Digital Newspaper Program) to look more broadly at the needs of digital newspapers in all of their diverse and challenging forms. This paper conveys the findings of the first research phase, including substantive survey results regarding digital newspaper curation practices.

Categories and Subject Descriptors

E.1 [**Data Structures**]: *distributed data structures*. H.3.2 [Digital Libraries]: Information Storage, *file organization.* H.3.4 [Systems and Software]: *distributed systems*. H.3.6 [**Library Automation**]: *large text archives.* H.3.7 **[Digital Libraries**]: *collection, dissemination, standards, systems issues*.

General Terms

Management, Documentation, Performance, Design, Reliability, Standardization, Languages, Theory, Legal Aspects, Verification.

Keywords

Archival Information Packages, Data Management, Digital Archives, Digital Curation, Digital Libraries, Digital Newspapers, Digital Objects, Digital Preservation, Distributed Digital Preservation, Ingest, Interoperability, Micro-Services, Repository Software, Submission Information Packages.

1. INTRODUCTION

U.S. libraries and archives have digitized newspapers since the mid-1990s using highly diverse and ever-evolving encoding practices, metadata schemas, formats, and file structures. Increasingly, they are also acquiring born-digital newspapers in an array of non-standardized formats, including websites, production masters, and e-prints. This content genre is of great value to scholars and researchers, and it is in critical need of preservation attention. The diversity of file types, formats, metadata, and structures that constitute this genre raises two major concerns: How can curators ready these collections for preservation? How may they conduct efficient repository-to-repository transfers from their local systems into digital preservation repositories?

The US National Endowment for the Humanities (NEH)-sponsored "Chronicles in Preservation" project is enabling the Educopia Institute, in collaboration with the MetaArchive Cooperative, the San Diego Supercomputer Center, and the libraries of University of North Texas, Penn State, Virginia Tech, University of Utah, Georgia Tech, Boston College, Clemson University, and the University of Kentucky, to investigate these issues through the following research questions:

1. **How can curators effectively and efficiently prepare their current digitized and born-digital newspaper collections for preservation?** We are documenting guidelines and available tools for the evaluation and preparation of a diverse set of newspaper collections for preservation. We are analyzing the costs and benefits of data preparation and studying how best to lower obstacles to preservation.
2. **How can curators ingest preservation-ready newspaper content into existing digital preservation solutions?** The project team is studying existing mechanisms for repository exchange. We are building software bridges to facilitate the exchange of newspaper collections between partners' local repository systems and distributed digital preservation (DDP) frameworks

This paper conveys the findings of the first phase of our project work, including substantive survey results we have gathered and analyzed regarding digital newspaper curation practices. In it, we begin by exploring the range of issues that born-digital and digitized newspaper content raises for curation and preservation practices. We then share information regarding our project findings and recommendations for near-future work.

2. THE CALF-PATH SYNDROME

...A hundred thousand men were led
By one calf near three centuries dead.
They follow still his crooked way,
And lose one hundred years a day,
For thus such reverence is lent
To well-established precedent.
-Sam Walter Foss, "The Calf-Path"

The story that the nineteenth century librarian and poet Sam Walter Foss tells in his poem entitled "The Calf-Path" is the story of a calf that perambulates through a wilderness, leaving behind a crooked trail that is gradually built up by subsequent animals and then humans. Over the course of a century the twisted trail becomes a road and eventually a highway through the center of a great metropolis. The poem is a humorous cautionary tale about the dangers of blindly following unexamined precedents.

iPRESS2012, Oct 1-5, 2011, Toronto, ON, Canada. Copyright 2012, Digital Curation Institute, iSchool, University of Toronto.

The poem is a useful allegory concerning the problems that digitization and digital preservation programs may encounter when growing over time. Many such programs have humble origins in underfunded libraries and other cultural memory organizations, and are begun informally by a small number of staff who often "make it up as they go along." As such programs blossom and achieve larger scale they often unwittingly preserve unexamined workflow precedents, much like the humans following the crooked trail of the calf in the poem. Often, these "calf-path" workflow problems are not evident to the individuals following the pre-established precedents. Rather, staff members are so busy trying to move more digital content through these well-established but inefficient practices that they never have the opportunity to step back and assess the overall efficacy of established workflows. The authors have examined the calf-path syndrome in digital preservation programs previously. [1] The calf-path syndrome is evident in most existing digital preservation programs for newspapers. We will occasionally invoke the calf-path syndrome in critiquing programs examined in this paper.

3. SIGNIFICANCE

The curation and long-term preservation of digital newspaper content presents unique challenges that are not fully understood and that demand additional research to ensure the survival of today's digital newspaper collections for tomorrow's researchers.

3.1 Newspapers as a Preservation Problem

Libraries and archives provide researchers with access to millions of digitized pages of historic newspapers. Some of these newspapers were scanned from print copies; others from microfilm. Some were digitized in-house; some outsourced to vendors. The scanning and encoding processes used in the digitization of historical newspapers vary wildly, as do the repository structures and storage media in which they are held.

Further complicating this digital genre, most newspaper producers shifted their operations to digital production by the beginning of this century. Increasingly, these born-digital print-production files are being acquired by libraries and archives. Many news groups also maintain websites that include non-AP wire materials of great value to researchers. As with digitized newspaper files, these born-digital files represent a range of format types (including websites, production masters, and e-prints) and are arranged in a wide variety of file structures and repository systems.

Digital newspaper files, then, are of increasing cultural and historical importance to researchers served by libraries, archives, and other memory organizations. One quality shared by nearly all of these diverse digital newspaper collections is that they are not yet preserved. [2] The lack of standard or normalized practices for the curation of these digital newspaper collections both within individual institutions (where practices have changed over time and remediation of earlier collections has not been pursued) and across the nation makes digital newspaper collections a high-risk genre of content that presents significant preservation challenges

Research has demonstrated clearly that content preparation and ingest are the most time-consuming and costly parts of preservation (creating SIPs and AIPs, in OAIS terminology). [3] The steps involved in preparing content include properly documenting a collection (ascribing descriptive, technical, and structural metadata to files and collections), ensuring its current and future viability (establishing that the files will render on current and future media), and organizing the files so that they can be managed over time (attending to file naming conventions and file structures such as folder and sub-folder designations).

The more normalized a collection is, the easier (and thus less time intensive and expensive) the process becomes of creating SIPs and, upon ingest, AIPs. In the case of digital newspapers, our research demonstrates that news content held within one institution is likely to include multiple digitized collections with different encoding levels, metadata treatment, file naming conventions, file types, and file structures because these collections were digitized at different times according to different standards, often by different teams (including external vendors). Also, these collections often are held in different repository systems.

For those institutions that are collecting born-digital newspapers, there are additional "calf-path" concerns. These collections are acquired in a wide range of ways, from hard-drive hand-offs of the master print-ready PDFs to Web crawls conducted upon newspaper Web sites. Because publishers vary widely in their own practices, the file types and file structures in these collections also include much variability. According to such factors, each of an institution's digital newspaper collections may need individualized analysis to ready it for ingest into a preservation environment.

Unsurprisingly, curators cite grave concerns about how they will be able to prepare such problematic collections for preservation, both from practical and fiscal perspectives. [4] With limited resources, how can institutions prepare their content for preservation, and how much data preparation is "enough" to suffice? To address this question, our research team has explored the applicability of the NDNP's existing set of recommendations for digitization efforts to the diverse body of legacy and born-digital newspaper content curated by libraries and archives.

3.2 NDNP Standards

The goal of the NEH and Library of Congress-supported National Digital Newspaper Program (NDNP) has been to develop an Internet-based, searchable database of U.S. newspapers that explicitly addresses the long-term content management and preservation needs of these collections.

The foremost set of technical parameters defined by the program relates specifically to scanning resolutions and establishing standard, high-quality file formats for NDNP digitization (TIFF 6.0). The majority of the additional technical parameters developed by the program seek to establish quality requirements for uniform metadata (CONSER-derived), encoding levels (METS/ALTO), and derivative file formats (JPEG2000 and PDF w/Hidden Text). Each of these requirements is in keeping with current high standards for archival-quality digitization for image-based items, and prepares the collections for successful repository management as defined by the OAIS Model. [5] The NDNP, then, is establishing best practices with implications far beyond the "Chronicling America" collection. Other institutions that are beginning or continuing digitization of newspapers benefit greatly from these standards, which help to ensure standard levels of encoding, file types, and uniform metadata that are geared for inter-repository sharing and long-term data management.

However, a wealth of digitized and born-digital newspaper collections exists in libraries, archives and other institutions that has been produced and obtained over the past two decades in a broad range of format types. [6] These "calf-path" collections

have been encoded at varied levels, use a diverse array of metadata schemas, and are arranged in highly irregular file structures and repository systems. The NDNP technical guidelines do not currently provide explicit recommendations for readying such "legacy" and born-digital collections for preservation.

Our research explicitly seeks to fill this gap, building on the stable foundation of the NDNP guidelines to address additional content within the broader "newspaper" genre. Rather than taking a "one-size-should-fit-all" approach, we differentiate between two tiers of preservation preparation: the *essential* and the *optimal*. If data preparation guidelines aim only for the "optimal," curators at institutions with limited resources will be unable to implement them. This would be detrimental to our main goal, which is to enable curators at institutions with a wide range of resources and collection types to begin preserving their digital newspaper collections. We seek to ensure that guidelines enable curators of various resource levels to preserve collections (again, defined as "ensuring that they may be accessed for as long as they are needed"), and that the standards and guidelines for the field do not themselves become preservation obstacles by making overly high demands that curators lack the resources to implement.

4. WHY DDP?

Recent studies and national initiatives (i.e., US NDIIPP) have urged the digital library community to explore collaborative technical and organizational solutions to "help spread the burden of preservation, create economies of scale needed to support it, and mitigate the risks of data loss." [7] The library community has concluded "the task of preserving our digital heritage for future generations far exceeds the capacity of any government or institution. Responsibility must be distributed across a number of stewardship organizations running heterogeneous and geographically dispersed digital preservation repositories." [8] Some early answers to this call embed collaborative practices in their technical and organizational infrastructures. For example, in distributed preservation repositories (e.g. Chronopolis, MetaArchive, CLOCKSS, Data-PASS), preservation activities occur within a dispersed network environment that is administered by multiple institutions. This approach combines geographic distribution with strong security of individual caches to create secure networks in which preservation activities may take place.

Such *Distributed Digital Preservation* (DDP) networks leverage inter-institutional commitments and infrastructures to support the requisite server infrastructures and to conduct necessary preservation activities in a local manner. In so doing, they capitalize on the existing infrastructures of libraries and archives (and in some cases, their parent institutions), simultaneously reducing costs and ensuring that digital preservation expertise is community-sourced, or built within the cultural memory community, not outsourced to third-party service providers.

Though the digital medium is relatively new, the conceptual approach taken by DDP practitioners is not. In the scribal era, this combination of approaches—geographic dispersal of content and secure storage environments—maximized the survivability of content over millennia. [9] Secure distribution helps content to withstand large-scale disasters (e.g., wars, hurricanes power grid failures) and more isolated, local-level events (e.g., media failures, human errors, hacking, fires).

In the last decade, many programs have developed using collaborative and distributed methodologies, and still others are in pilot phases of their research and development work. Examples of proven approaches include MetaArchive (Private LOCKSS Network (PLN)), Chronopolis (SDSC's iRODS-based service), and the Data-PASS Network (ICPSR/Roper Institute/Odum Institute partnership to preserve social science datasets using a PLN). Other experimental approaches show great promise, including Digital Preservation Network (DPN, bridging heterogeneous preservation environments), DuraCloud (DuraSpace's cloud-storage-based environment) and LuKII (a German program that bridges LOCKSS's cost-effective preservation with KOPAL's usability and curation tools).

The demand for community-based initiatives hosted and managed by libraries and archives is strong. Surveys conducted by the MetaArchive Cooperative in 2009 and 2010 reveal that curators of digital newspaper content both need and actively seek implementable digital preservation solutions and models. Most institutions (80%) report that they do not aspire to build their own preservation repository due to the expense, technical expertise, and infrastructure required. Fully 73% of 2009 and 2010 respondents reported that they were interested in using community-based preservation networks, while only 30% reported interest in third-party vendor solutions. [10]

The Chronicles research project focuses on three approaches to preservation—MetaArchive, Chronopolis, and CODA—which share certain common characteristics, but use very different technologies to accomplish their goals. The three most salient similarities between these approaches are 1) they all use open-source technologies; 2) these are library-run, community-sourced ventures; and 3) these are *Distributed* Digital Preservation (DDP) approaches. Each of these approaches varies in other key areas such as ingest mechanisms, data management practices, organizational model, and recovery options.

4.1 MetaArchive Cooperative

The MetaArchive Cooperative is a community-sourcing network that preserves digital collections for more than 50 member libraries, archives, and other digital memory organizations in four countries. The Cooperative was founded in 2003-2004 to develop a collaborative digital preservation solution for special collections materials, including digitized and born digital collections. Working cooperatively with the Library of Congress through the NDIIPP Program, the founders sought to embed both the knowledge and the technical infrastructure of preservation within MetaArchive's member institutions. They selected the LOCKSS software as a technical framework that matched the Cooperative's principles, and built additional curatorial tools that layer with LOCKSS to promote the curation and preservation of digital special collections, including newspapers, Electronic Theses and Dissertations, photographs, audio, video, and datasets. In doing so, they created a secure, cost-effective repository solution that fosters ownership rather than outsourcing of this core library/archive mission. The Cooperative moved to an open membership model in 2007, and has expanded in five years from a small group of six southeastern academic libraries to an extended community of more than 50 international academic libraries, public libraries, archives, and research centers.

4.2 Chronopolis

The Chronopolis digital preservation network has the capacity to preserve hundreds of terabytes of digital data—data of any type or size, with minimal requirements on the data provider. Chronopolis comprises several partner organizations that provide a wide range

of services: San Diego Supercomputer Center (SDSC) at UC San Diego; UC San Diego Libraries (UCSDL); National Center for Atmospheric Research (NCAR); and University of Maryland Institute for Advanced Computer Studies (UMIACS). The project leverages high-speed networks, mass-scale storage capabilities, and the expertise of the partners in order to provide a geographically distributed, heterogeneous, and highly redundant archive system. It uses iRODS (Integrated Rule-Oriented Data System) to federate three partner sites and replicate data, BagIt to transfer data into the storage locations, and ACE (Audit Control Environment) to monitor content for integrity.

4.3 University of North Texas

The University of North Texas has constructed a robust and loosely integrated set of in-house archiving infrastructures to manage their digital collections, including a delivery system (Aubrey) and a Linux-based repository structure (CODA). The underlying file system organization of digital objects is tied to a UNT-specific data modeling process that relies on locally developed scripts and micro-services to generate and define all master, derivative, related objects, metadata, and other information that may be tied to a single digital object in order to effect archival management and access retrieval. This archival repository solution has been designed with open source software and relies on loosely bundled specifications to ensure on-going flexibility. UNT's archival repository implemented its integrated offsite replication in 2010. The micro-services that support the current instance of CODA are being experimented with for optimizing workflows across both instances of the repository.

5. SURVEYING DIGITAL NEWSPAPERS

The Chronicles in Preservation project has investigated a diverse array of digital newspaper content and its associated preservation needs across a broad stratum of institutions. This took the form of an extensive survey and set of interviews that were carried out beginning in October 2011. [11] The eight academic libraries participating in the project were asked for detailed information about the range of digital newspaper collections they curate (e.g., file formats, encoding practices, etc); the repository infrastructures they use to support this content; and their requirements for archival ingest and long-term distributed digital preservation. A summary of the survey findings follows.

5.1 Preservation Formats, OCR & Metadata.

The surveyed content curators cited divergent needs and practices regarding what image formats they produce, manage, and intend to preserve. Most surveyed libraries report using TIFF as their primary master image format (the exception, Virginia Tech, works exclusively with born-digital content—HTML and PDF). The respondents also reported using a range of derivative file types, including PDF (7 libraries), JPEG2000 (6 libraries), JPEG (3 libraries), xml (2 libraries), and HTML (1 library).

Preservation ambitions vary across the surveyed libraries. Some locations intend to preserve only their master TIFF images (Clemson, University of Kentucky, University of Utah, and UNT). Others also focused on their derivative JPEG and PDF images (Georgia Tech), and JPEG2000 images (Boston College). All respondent libraries report that no file format used in their newspaper curation practices has become obsolete to date. All likewise report that they have only normalized and migrated files for the purposes of producing derivatives for access. Four of the respondent libraries report using JHOVE for file format identification and/or validation purposes.

In addition to the array of target image formats mentioned above, all of the content curators are creating & maintaining a range of OCR formats (XML, PDF, ABBYY, METS/ALTO, ALTO, PrimeOCR, etc.) and metadata (Fedora Core, METS, MIX, MODS, customized Dublin Core, etc.) formats. In some cases, the collection/object-to-metadata relationships remain somewhat opaque to the content curators due to their reliance upon their repository software for metadata creation and maintenance. In several other cases, content curators are making use of METS to encapsulate their digital objects and various associated metadata. In most cases, the content curators were confident that their metadata could be exported from their repository systems in some form of XML for external processing.

5.2 Repository Systems & Features.

Content curators are using a diverse array of repository software solutions to manage their digital newspaper collections. These include licensed open-source solutions such as Fedora (Clemson) & DSpace (GA Tech), as well as licensed proprietary solutions such as CONTENTdm (Penn State; University of Utah), Olive ActivePaper (Penn State) & Veridian (Boston College). Other implementations range from University of Kentucky's (UKY) and University North Texas's homegrown infrastructures modeled on a micro-services architecture, all the way to the use of simple web servers (Penn State; Virginia Tech). It should be noted that with the exception of UKY and UNT, none of the repository solutions indicated above are aiming to be fully supported preservation systems. The systems reported are generally prioritized to support access. Only Georgia Tech is storing their master TIFF images in their DSpace repository instance (with backup support on-location). In most cases, master TIFFs or JPEG2000s are typically stored and backed-up in on- or off-site SAN or tape systems.

In order to prepare the content stored in these access-oriented systems for ingest into preservation systems, SIPs may need to be staged externally. It should also be noted that some dependencies exist at the level of metadata and object/collection identifier creation and export, as these systems provide custom-built or proprietary modules with varying degrees of flexibility for open- and user-defined conventions. Export utilities and HTML/XML parsers may need to be identified or developed to support their harvest and retention at ingest.

5.3 Data Management Practices.

Collection and/or object identifier schemes for content curators' repository environments spanned a wide range of implementations. Most of these schemes employ user- or system-generated persistent identifiers (e.g., Fedora PID at Clemson, DSpace Handles at Georgia Tech; Veridian custom URLs at Boston College; NOID and CDL Identity Service at UKY; CONTENTdm Reference URLs at University of Utah; Coda ARKs at UNT). Only three of these content curators have developed formal digital object identifier schemes external to these repository systems (Boston College and UNT). Boston College uses a standard code for a newspaper title, a CCYYMMDD date, and 3-digit image/page sequence number (e.g., bcheights/1921/05/21/ bcheights_19210521_001.jp2). UNT assigns a unique identifier at the digital object level according to CDL's ARK specification. UKY makes use of NOID in conjunction with a locally developed identifier scheme. All content curators have indicated that the retention of any collection

and/or object identifiers is crucial for recovering their current repository environments. However, this warrants further investigation into the ramifications of decisions regarding what forms of the content are preserved (e.g., preserving master images and not derivatives) as this may hinder the recovery of an access-based repository environment.

5.4 Collection Sizes, Growth Rates & Change.

Reported collection size aggregations follow a number of models—some by title, some by issue, others by originating institution. Some aggregations are no more than 60 megabytes, others can reach as much as seven terabytes. The majority of collection aggregations that were surveyed stay well below half a terabyte. Content curators are systematically acquiring and adding new digital newspaper content according to a variety of schedules. University of Utah, University of Kentucky, and University of North Texas reported the most dynamic rates of acquisition—20,000 pages per month, 20,000 pages per quarter, and 40,000 issues per year respectively. Penn State also reported a robust rate of acquisition at approximately 75,000 pages annually. The majority of content curators however have relatively static or only mildly growing digital newspaper collections. Georgia Tech reported ten issues of growth per month, and Clemson University only one or two titles per year. Boston College could only speculate on future growth with two potential titles under negotiation, and Virginia Tech suggesting no future growth.

Content curators were surveyed for any existing change management policies or practices in the midst of such rates of growth. This was intended to account for image or metadata files that may have undergone repair or refreshment— tracking or associating versions of files through identifier or naming conventions for example. This was also intended to account for any changes to underlying technical infrastructure supporting local archival management—perhaps recording technical and administrative metadata through METS or PREMIS. None of the content curators, with the exception of UNT, had formal change management policies or could clearly identify repository or other system features that were accomplishing version management. UNT has a robust set of data management workflows that account for all events that take place on a digital object (files and metadata). They are also moving towards establishing workflows that track changes to technical infrastructure (hardware refreshment, system updates, etc.). Knowing the state of such local policies and practices can help institutions understand the degree to which such meaningful preservation activities may need to be accommodated or similarly maintained external to the content curator.

5.5 Preservation Preparedness

As detailed above, content curators are currently managing a range of well-supported digital formats for their digital newspaper collections. In most cases, content has been digitized to high archival standards. Master images are in TIFF format, and derivative access copies are in high-resolution JPEGs, PDFs, or JPEG2000s. Exceptions to these standards include a small subset of very early versions of HTML-encoded websites, and lower-resolution PDF master images.

As previously mentioned, none of the content curators we surveyed have performed format migration or normalization for the purposes of preservation. Among the surveyed libraries, file format identification tools like JHOVE, JHOVE2 or DROID are in moderate use (4 of the 8 institutions). None of the surveyed content curators currently subscribe to format registry services such as the Unified Digital Formats Registry (UDFR). With the exception of one content curator, the use of PREMIS is not yet routine or programmatic. However, as also noted above several content curators are gathering administrative, technical, structural, and provenance metadata for the digital objects that comprise their digital newspaper collections. In some cases this metadata is being systematically generated at ingest through the use of JHOVE, and other system utilities, and being related to corresponding digital objects through use of METS, MIX & MODS—which can be bridged to PREMIS. When asked about near- to long-term capacity for creating and managing preservation metadata most content curators stated a current lack of familiarity with PREMIS, but noted their awareness of it and their potential staff capacity for integrating PREMIS in their local workflows in the future.

Beginning in Fall 2012, the Chronicles in Preservation project will enter the Transition and Documentation Phases, in which project staff will document the necessary preservation readiness steps that the project partners need to apply to their own very diverse holdings—both digitized and born-digital—for the purposes of experimenting with more robust preservation. These individualized "preservation preparedness plans" will be derived from the more general *Guidelines to Digital Newspaper Preservation Readiness* that we are currently producing. Like the *Guidelines,* these preservation preparedness plans will seek to document preservation readiness strategies for each institutional partner along a spectrum of the *essential* to the *optimal*.

This "spectrum" approach enables the content curators at our partner institution sites (as with the larger field addressed in the *Guidelines*) to understand the acceptable range of activities they may undertake in their preservation readiness practices. By documenting the *essential* and the *optimal*, we invite and encourage institutions to engage responsibly with preservation at the level they can currently handle without delay. We also make it possible for those with lower resources to understand the difference between their current activities (essential) and those to which they should aspire in the future (optimal). The *essential* recommended readiness steps to be taken may be achieved even given the limited resources and expertise that are typically available to the average content curator. These are what we consider non-negotiable activities, because to neglect them would undermine the long-term preservation of their content. The *optimal* workflows will ensure the highest standards in long-term preservation for those that do have the resources to pursue them now, and they will provide those institutions that can only aspire to the "essential" level today with benchmarks for later success.

We believe that taking this flexible approach to documenting preservation measures for digital newspapers will enable content curators to understand what they can begin doing in the short-term in the absence of high levels of resources and expertise, and will provide them with a foundation for the "optimal" curation practices to enhance their preservation capacity going forward.

5.6 Preservation Pathways

Each of the project's three DDP sites has its own unique mechanisms for handling ingest, packaging AIPs, and effecting long-term preservation. During the surveys, content curators were asked a series of questions about their experience concerning digital newspapers with the general types of ingest-related technologies that each of the preservation sites use (e.g., web harvesting mechanisms, use of the BagIt specification, and the use

of micro-services). Aside from Virginia Tech's previous development work to ingest digital newspaper content into MetaArchive, and UNT's use of BagIt and various micro-services, none of the respondents have pursued these technologies for managing their digital newspapers.

Similarly, but with a different emphasis, content curators were surveyed for their preferences for ingest strategies. Suggested options included shipping hard drives, performing server-to-server copies, performing BagIt based transfers, or triggering web harvests on staged content. Half of the content curators (4 of 8) indicated a strong preference for shipping their hard-drives to preservation sites or allowing a preservation site to perform remote copying of data from a secure server connection, and half also showed a preference for the use of BagIt. Web-crawl strategies fared somewhat lower in terms of preference, with only two content curators listing this strategy as a first option.

6. DIGITAL NEWSPAPER CASE STUDIES

Following the survey, we conducted in-depth interviews with our partners. Below, we share information from University of North Texas (UNT) and Virginia Tech drawn from the focused interviews we have conducted. The UNT case study provides one possible pathway for rectifying the calf-path syndrome by carefully balancing the needs associated with inherited precedents against local needs for achieving scale and efficiency. The Virginia Tech case study illuminates the kind of meandering workflows that can arise when a preservation program inherits content streams from many pre-existing sources.

6.1 University of North Texas Case Study

The University of North Texas Libraries (hereafter UNT) are actively involved in a number of newspaper digitization and preservation activities. Beginning in the same year as its first NDNP award, UNT developed a comprehensive program to identify, collect, digitize and preserve newspapers from around the state of Texas with a program called the Texas Digital Newspaper Program [12]. The team at UNT leveraged the technical specifications of the NDNP program in all but one area for use in non-NDNP newspaper digitization as well as identifying several new workflows for the acquisition and processing of born-digital print masters from publishers around the state. All digitized and born-digital newspaper content is added to The Portal to Texas History [13] for end user access and also to the UNT developed CODA preservation infrastructure for long-term storage and management. To date UNT has made freely available over 750,000 pages (95,000+ issues) from 409 different titles via The Portal to Texas History.

6.1.1 Standards and Workflow

The UNT workflow for newspaper digitization and born-digital processing is heavily influenced by the *NDNP Technical Guidelines and Specifications* [14] that is comprised of a number of technical sub-specifications, all of which are important when trying to organize a large-scale newspaper digitization program like the NEH NDNP program or UNT's Texas Digital Newspaper Program. UNT found that these specifications provided a good starting point for refining its internal workflows and standards.

Source Material Selection: The NDNP specification advises use of second-generation negative film on a silver halide substrate. The specification also allows use of born digital images or images scanned from paper. UNT found it very important to use second-generation negatives for the best results in the digitization process. For titles only available in print format UNT contracted with vendors to microfilm the title before the digitization process. Born-digital files are also collected from a number of publishers around the state. Typically these are the production print masters sent to the printers that are then delivered to the UNT team. The goal in each content stream is to ensure that the highest quality, most complete version of the title is being use for later processing.

Scanning: The NDNP specification describes the resolution and color space that is optimal for scanning content: 300-400 DPI using 8 bit grayscale. UNT views this as a minimum resolution, whether the scanning is performed by outsourced services or internally within the UNT Libraries. Born-digital print masters are converted from their delivered formats (usually pdf) into 400dpi, 24bit JPEG images which are used for subsequent processing. The delivered pdf masters are retained and stored with the object in the final archival package ingested into the CODA repository.

File processing: UNT aligns with the NDNP specification with regard to processing on the master files created in the digitization process. Scanned images are de-skewing to within 3% skew and cropping with a slight edge around the physical piece of paper, not just the text on the page. Born digital items are left unaltered other than occasional 90-degree rotation to properly align the text.

OCR: UNT utilizes the ABBYY Recognition Server for the optical character recognition (OCR) process when items are digitized in-house. The ABBYY software is operated in a cluster configuration with six nodes (52 cores) dedicated to the OCR process. UNT has found this tool to provide an appropriate tradeoff between quality, convenience and costs of OCR.

Serializing a newspaper issue to files: The NDNP specification describes the use of the METS and ALTO specifications to represent a newspaper issue on a file system. This is an area that UNT begins to depart from the NDNP specifications to allow for integration into local systems. OCR files from the ABBYY Recognition Server are converted into several legacy formats for representing bounding box information and indexed text. The master ABBYY XML file is also saved with the output files for later processing if the need arises. All pages associated with an issue are placed in placed in a folder named with the following convention, yyyymmddee (y=year, m=month, d=day, e=edition). Descriptive metadata is collected for each issue and stored alongside the page images in the issue folder and is used at a later point in the ingest process. A future area of development is the conversion of the proprietary ABBYY format into the standard ALTO format used by our NDNP projects to allow for a greater use of ALTO enabled workflows and tools.

Derivatives: The NDNP specification calls for creating a JPEG2000 and PDF for each page of newspaper. UNT currently creates JPEG2000 derivatives on ingest into its Aubrey content delivery system. In addition to JPEG2000 files, traditional JPEG images are created in a number of resolutions such as square, thumbnail, medium and large to provide a variety of viewing strategies for end users. UNT also pre-tiles each image loaded into The Portal to Texas History with the Zoomify tile format and stores these tiles in WARC [15] files.

Ingest: The UNT Libraries' ingests all digitized and born-digital newspapers into a locally developed system called CODA, which provides archival file management for digital content under its management. Each item ingested is assigned a globally unique ARK identifier that is used to request the item from CODA.

Summary: The UNT internal workflow is heavily influenced by the NDNP technical specifications, which constitutes an excellent set of specifications for libraries and vendors to use in digitizing and delivering newspaper content. These specifications can be used as a starting point for developing local workflows that take into account new content acquisition strategies and formats not covered completely by the NDNP program. One key aspect missing in the NDNP specifications that might be useful to the newspaper digitization community is an extension to allow for article level data to be encoded into the METS/ALTO format.

6.1.2 Avoiding the Calf-Path

The UNT case study demonstrates ways of avoiding the calf path by carefully comparing and analyzing competing requirements that derive from external precedents and internal optimization needs. This is possible when setting up a new or relatively new program at scale, but may not be possible when a program has long-standing inherited precedents. It may be very difficult to get off the calf path in some situations, as the following case study from Virginia Tech illustrates.

6.2 Virginia Tech Case Study

The digital newspaper collections of Virginia tech represent a diverse and un-normalized legacy of digital content. Within the *Chronicles in Preservation* project, Virginia Tech is a good case study in dilemmas associated with born-digital content, since the university has not engaged in digitization but has hosted born-digital newspaper content for almost two decades. Virginia Tech began accepting web pages and PDFs from various local, regional, international news agencies in 1992. More than 19 gigabytes of news content has now accumulated at the university, which was received directly from the publishers in digital formats.

In 1992, the Virginia Tech library began receiving online news feeds from the two major newspapers in Southwest Virginia, ultimately resulting in over 400,000 text files documenting life in this region. In 1994 the library began capturing the university's newspapers, and in 1997 international news began arriving in PDF format. The 2,600 PDF files collected provide a context for studying Lebanon, Iran, and France in the local languages—Arabic, Farsi, and French.

6.2.1 Problems with Metadata

Metadata was not systematically collected for this body of content for many years, since the Virginia Tech staff working on these projects was quite limited and in the early search engines of the 1990's ignored metadata. Staff members to create metadata were gradually added with the intent of implementing a better practice for organizing the digital content being gathered.

The first step taken was to begin adding very basic article-level info derived from the text files comprising individual newspaper articles. An example newspaper for which this practice was implemented is the *Roanoke Times*, which began including date, author, edition, location, and title information in the text file headers circa 1996. These metadata elements could be parsed and used for indexing, access, and organization purposes.

Various ad hoc parsing scripts were developed over time to extract metadata from the news content feeds received at Virginia Tech, and normalize this metadata into Dublin Core elements. This practice was fragile, however, and prone to malfunction if the format of the feeds changed over time. Virginia Tech is still considering how to effectively automate the generation of metadata for these content feeds. This is an example of the most difficult kind of calf-path to escape, a long-standing set of uncontrollable external data feeds that cannot be remediated.

7. PRESERVING DIGITAL NEWSPAPERS

Though the range of content needs for the various digital newspaper holdings are highly diverse, even within a single curatorial location, the concept of "standardizing" requires us to pursue uniform approaches and recommendations, both broadly through the *Guidelines*, but also within the individualized "preservation readiness plans." This applies not only to such tasks as exporting and compiling metadata or forward migrating to de-facto standard OCR formats such as ALTO, but also attempting to achieve common packaging and ingest measures.

7.1 Repository-to-Repository Exchanges

Data exchange challenges are complex and as yet unresolved, both within and well beyond the library and archives communities. The most successful data exchange models address issues that arise in specific genres of content, from emergency alert systems (OASIS) to social science data sets (DDI). [16] Most data exchange models to date—including those created for newspapers— have been used primarily to address the integration and federation of content for access purposes. How might the genre of interest here—newspaper data—be exchanged for preservation purposes? The issues involved in data exchange in the preservation context are twofold, involving both data structures (the way that the collections' constituent parts are stored and how the repository system uses those stored components to assemble an access view) and repository system export and ingest options (ways of moving content in or out of repository environments). Libraries and archives, as mentioned above, use many different types of repository systems to store their digital newspaper content. Each of these repository systems has expectations about how data is structured. The mismatch of these expectations between repository systems makes it difficult to move collections from one system to another while maintaining each collection's integrity and set of relationships. [17]

We are currently studying existing specifications for transfer to assess their applicability to the genre of digital newspaper content, including UIUC's HandS Project, TIPR, and BagIt. [18] To date, much of the interoperability and exchange work between access-oriented repositories and preservation repositories for collaborative frameworks, like those chosen for evaluation in this project, have happened in one-off fashion. For example, the MetaArchive Cooperative has successfully exchanged content with Chronopolis, and has also ingested content from DSpace, CONTENTdm, Fedora, Digital Commons, and ETDb repositories by creating "plugins" specific to each content contributor's collections. Likewise, there have been projects that have explored the use of DSpace with SRB/iRODS and Fedora with iRODS. These have been largely geared toward addressing an individual institution's collections and have been mapped in a straightforward pathway from DSpace to iRODS and Fedora to iRODS. Such work may help individual institutions, but it does not efficiently streamline the ingest process in a way that is relevant to the larger digital library and archives community when preserving their content in various collaborative solutions.

7.2 Towards Interoperability Tools

We are currently documenting the complexities involved in streamlining such access-to-preservation repository exchanges. We are encountering a range of issues, exemplified here by our preliminary research. As detailed above, during these investigations a number of questions have arisen regarding compatibilities between partner institutions' collections and both the access-oriented systems and the preservation systems being evaluated. For example, what data management components must be implemented in the MetaArchive and Chronopolis environments to facilitate, create, and update the administrative, preservation, and technical metadata that accompanies a potential exchange profile? Is UNT-CODA's micro-services based approach for preparing SIPs to become AIPs extensible to the MetaArchive and Chronopolis environments and could this approach provide flexible alternatives to requiring well-formed and standardized exchange profiles? Conversely, how do the UNT workflows for enhancing SIPs through micro-services interact with exchange packages that already include this information (e.g., Penn State's NDNP collections)?

To study these and other issues, the project's technical team is analyzing the applicability of existing efforts to move content between systems for meeting our project goals. We are also experimenting with BagIt to determine whether that transfer mechanism will accommodate the full range of digital newspaper packaging requirements as documented in the *Guidelines* and "preservation readiness plans." In conjunction with our Chronicles Committee and Advisory Board, the project team is also studying the benefits of and barriers to implementing PREMIS and METS for our partners' collections and for these preservation environments. All of these findings will be documented in a white paper that will be released in early 2013 via the project site: http://metaarchive.org/neh.

8. CONCLUSIONS

The first phase of the project facilitated our understanding of the current practices and workflow needs of newspaper content curators. It also substantiated our theory that a single unified workflow is not \an optimal approach for engaging institutions in the process of readying their content for preservation. To encourage broad participation, we should not seek to establish a single workflow or exchange mechanism for preparing a collection for ingest across all three preservation systems explored in this project. Rather, we will aim to reduce barriers by establishing a range of guidelines and workflows and by building systematic approaches for exchanging content between common access-oriented repositories and mature preservation solutions.

9. ACKNOWLEDGMENTS

We greatly appreciate the generous support of the National Endowment for the Humanities through Award PR-50134.

10. REFERENCES

[1] Halbert, M., Skinner, K., and McMillan, G. 2009. "Avoiding the Calf-Path: Digital Preservation Readiness" *Archiving 2009 Proceedings.* pp. 86-91.

[2] Skinner, K. and McMillan, G. 2009. "Surveys of Digital Preservation Practices and Priorities in Cultural Memory Organizations." NDIIPP Partners Meeting, Wash. DC,. DOI= www.digitalpreservation.gov/meetings/documents/ndiipp09/NDIIPP_Partners_2009_finalRev2.ppt; Skinner, K. and McMillan, G. 2010. *Survey of Newspaper Curators.* Educopia Institute; Skinner, K. 2012. Survey on Digital Newspaper Archiving Practices. Educopia (*forthcoming*).

[3] Beagrie, N., Lavoie, B., and Woollard, M. 2010. Keeping Research Data Safe 2 Final Report. JISC/OCLC. DOI= http://www.jisc.ac.uk/publications/reports/2010/keepingresearchdatasafe2.aspx#downloads.

[4] Angevaare, I. 2009. Taking Care of Digital Collections and Data 'Curation' and Organisational Choices for Research Libraries. *Liber Quarterly*, 19:1. DOI= http://liber.library.uu.nl/index.php/lq/article/view/7948.

[5] Library of Congress. 2009. NDNP: Technical Guidelines. http://www.loc.gov/ndnp/guidelines/archive/NDNP_201113TechNotes.pdf.

[6] As the Library of Congress underscored in a Broad Agency Announcement (BAA) as a *Draft Statement of Objectives on Ingest for Digital Content* (June 2010): "Some digital content types have remained relatively stable in format over time (such as digital versions of academic journals), while others (such as digital versions of newspapers and other news sources) have become increasingly complex, evolving with the Internet environment…. Some digital content types are relatively self-contained… while others …contain (and/or are linked to) multiple digital content objects."

[7] Fran Berman and Brian Schottlaender, "The Need for Formalized Trust in Digital Repository Collaborative Infrastructure." *NSF/JISC Workshop*, April 16, 2007: http://www.sis.pitt.edu/~repwkshop/papers/berman_schottlaender.html (last accessed 06/07/2012); Please also see the following reports: American Council of Learned Societies. (2006) "Our Cultural Commonwealth: The Report of the ACLS Commission on Cyberinfrastructure for the Humanities and Social Sciences" *American Council of Learned Societies*: http://www.acls.org/cyberinfrastructure/ourculturalcommonwealth.pdf (last accessed 06/07/2012); Blue Ribbon Task Force on Sustainable Digital Preservation and Access. "Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information" February, 2010. Available at: http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf (last accessed 06/07/2012), and the JISC/OCLC "Keeping Research Data Safe 2 Final Report" (previously cited), which have pointed to the economic challenges inherent in "silo"-based development and maintenance in the area of preservation.

[8] Priscilla Caplan, "IMLS Funds TIPR Demonstration Project." *Digital Preservation Matters*, 2008. Available at: http://preservationmatters.blogspot.com/2008/09/imls-funds-tipr-demonstration-project.html (last accessed 06/07/2012).

[9] Katherine Skinner and Matt Schultz, Eds., *A Guide to Distributed Digital Preservation*, Educopia, 2010. Available: http://www.metaarchive.org/sites/default/files/GDDP_Educopia.pdf (last accessed 06/07/2012).

[10] For more on the surveys, please see Skinner and McMillan.

[11] Schultz, M., Skinner, K., 2011. Chronicles in Preservation: Collections Assessment Survey. Educopia Institute.

[12] The Texas Digital Newspaper Program (TDNP). Available at: http://tdnp.unt.edu

[13] The Portal to Texas History. Available at: http://texashistory.unt.edu

[14] NDNP Technical Guidelines and Specifications. Available at: http://www.loc.gov/ndnp/guidelines/

[15] ISO. Information and documentation – WARC file format (ISO 28ûı0:2009), 2009.

[16] OASIS Emergency Interoperability. OASIS. DOI= http://www.oasis-emergency.org; Data Documentation Initiative (DDI). Data Documentation Initiative Alliance. DOI= http://www.ddialliance.org/

[17] See for example, Clay Shirkey's "NDIIPP-Library of Congress Archive and Ingest Handling Test Report" (2005) http://www.digitalpreservation.gov/partners/.../**ndiipp**_aiht_final_report.pdf

[18] Hub and Spoke Project (HandS). UIUC. DOI= http://dli.grainger.uiuc.edu/echodep/hands/index.html;Towards Interoperable Preservation Repositories (TIPR). FCLA. DOI= http://wiki.fcla.edu:8000/TIPR/; BagIt. CDL. DOI= https://confluence.ucop.edu/display/Curation/BagIt