NATIONAL ENDOWMENT FOR THE
Humanities

DIVISION OF **PRESERVATION AND ACCESS**

Narrative Section of a Successful Application

The attached document contains the grant narrative and selected portions of a previously funded grant application. It is not intended to serve as a model, but to give you a sense of how a successful application may be crafted. Every successful application is different, and each applicant is urged to prepare a proposal that reflects its unique project and aspirations. Prospective applicants should consult the Preservation and Access Programs application guidelines at http://www.neh.gov/grants/guidelines/PARD.html for instructions. Applicants are also strongly encouraged to consult with the NEH Division of Preservation and Access Programs staff well before a grant deadline.

Note: The attachment only contains the grant narrative and selected portions, not the entire funded application. In addition, certain portions may have been redacted to protect the privacy interests of an individual and/or to protect confidential commercial and financial information and/or to protect copyrighted materials.

Project Title: Distributed Preservation of Born-Digital and Digitized Newspaper Collections

Institution: Educopia Institute

Project Director: Katherine Elizabeth Skinner

Grant Program: Preservation and Access Research and Development

INTRODUCTION

The Educopia Institute, with the San Diego Supercomputer Center and the libraries of University of North Texas, Penn State, Virginia Tech, University of Utah, Georgia Tech, Boston College, and Clemson University, proposes to study, document, and model the use of data preparation and distributed digital preservation frameworks to collaboratively preserve digitized and born-digital newspaper collections.

U.S. libraries and archives have been digitizing newspapers since the mid-1990s using a highly diverse and ever-evolving set of 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 in the humanities, 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 distributed preservation repositories?

The foundation for addressing the first of these issues is provided by the NEH- and Library of Congress-sponsored National Digital Newspaper Program's (NDNP) recommendations for digitizing newspaper content. The NDNP has developed preservation-oriented standards for current newspaper *digitization* practices. This project will explore how these standards may be applied and elaborated upon to foster the preservation readiness of collections from the last two decades that were digitized according to evolving standards, as well as the born-digital content that institutions are steadily acquiring..

Once curators successfully prepare their collections for preservation, how can they effectively exchange it across repository systems? This project will study and document for newspaper preservation the use of distributed digital preservation (DDP), a collaborative approach in which content is exchanged and replicated across multiple sites, and actively monitored using various network-driven technologies.

We propose to investigate these issues through the following series of research questions:

1. How can curators effectively and efficiently prepare their current digitized and born-digital newspaper collections for preservation? We will study and document guidelines and available tools for the evaluation and preparation of a diverse set of digitized and born-digital newspaper collections for preservation. We will analyze the costs and benefits of data preparation and study how best to lower the obstacles to preservation that are presented by this often-expensive process.

2. How can curators ingest preservation-ready newspaper content into existing DDP solutions? The project team will study existing mechanisms for repository exchange. We will build open source software bridges to facilitate the export of newspaper collections from partners' local repository systems (including Olive, CONTENTdm, DSpace, and DigiTool) and their ingest into DDP frameworks (iRODS, LOCKSS, California Digital Library microservices).

3. What are the strengths and challenges of three leading DDP solutions when used to preserve digital newspaper content? The project team will conduct a comparative analysis across three U.S.-based DDP environments (Chronopolis-iRODS, MetaArchive-LOCKSS, UNT's CDL microservices-based CODA) to document the strengths and challenges curators face when using them to ingest and preserve this content genre.

This research will result in *guidelines* for preparing digital newspaper collections for preservation, *interoperability tools* to facilitate the exchange of these newspaper collections between repositories, and a *comparative analysis* of the strengths and challenges of three distinct DDP frameworks when they are used for the preservation of digital newspaper content. In so doing, it will facilitate the long-term sustainability of this essential content genre for tomorrow's humanities scholars and researchers.

SIGNIFICANCE

Researchers rely heavily on newspaper content—from national dailies and local weeklies to subject- and community-oriented publications—to understand the context for events, to analyze local environments, and to compare the variety of perspectives that emerge therein. Yet the 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 and scholars.

Why are Newspapers a Preservation Problem?

Libraries and archives provide 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 digital newspaper files are being acquired and managed by libraries and archives. Today, with few exceptions, even those newspapers that maintain print copies are creating them from digital files, and 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 scholars.[1] The one quality that is 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[3] of earlier collections has not been pursued) and across the nation as a whole makes digital newspaper collections a high-risk genre of content that presents significant preservation challenges. The resulting body of digitized and born-digital newspaper content is in critical need of preservation attention.[4]

Preservation has been defined as the "series of managed activities necessary to ensure continued access to digital materials for as long as necessary."[5] Research has demonstrated that content preparation and ingest is the most time-consuming and costly part of preservation (creating Submission Ingest Packages, or SIPs and AIPs, in OAIS terminology).[6] 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 render on current media and are likely to do so in the future), 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, Archival Information Packages (AIPs). In the case of digital

[1] Indeed, as Advisory Board member Bob Horton has pointed out, digital *news* files (including blogs and social media commentary) are beginning to be prioritized for acquisition by libraries and archives and present major preservation challenges, well beyond those posited by text files.
[2] Katherine Skinner and Gail McMillan. "Surveys of Digital Preservation Practices and Priorities in Cultural Memory Organizations." NDIIPP Partners Meeting, Washington, DC, June 24, 2009. Available at: http://www.digitalpreservation.gov/news/events/ndiipp_meetings/ndiipp09/index.html, (last accessed 05/10/2010); 2010 Survey of Newspaper Curators (Educopia, 2010).
[3] Defined as the re-presentation of "old" media content in "new" media forms, the term "remediation" is often used by the library community to denote the process of correcting/updating digital content and its associated information to keep up with changing user needs and delivery options.
[4] A notable exception to this general rule is the NDNP's Chronicling America collection, which is preserved in a centralized repository at the Library of Congress. Most other news content is not yet being programmatically preserved.
[5] Digital Preservation Coalition. "Introduction: Definitions and Concepts." Digital Preservation Handbook. Available at: http://www.dpconline.org/advice/introduction-definitions-and-concepts.html, (last accessed 05/15/2010).
[6] Neal Beagrie, Brian Lavoie, and Matthew Woollard, "Keeping Research Data Safe 2 Final Report." *JISC/OCLC*, 2010. Available at: http://www.jisc.ac.uk/publications/reports/2010/keepingresearchdatasafe2.aspx#downloads, (last accessed 05/10/2010).

newspapers, even newspaper content held within one institution is likely to include different encoding levels, metadata treatment, file naming conventions, and file structures because collections were digitized at different times according to different standards. Also, these collections often are held in different repository systems. According to such factors, each of an institution's digital newspaper collections may need individualized analysis and treatment to ready it for ingest into a preservation environment.[7] Unsurprisingly, curators cite grave concerns about how they will be able to prepare such problematic collections for preservation, both from practical and fiscal perspectives.[8]

With limited resources, how may institutions prepare their content for preservation, and how much data preparation is "enough" to suffice? This project will begin to answer this question by examining the applicability of the NDNP's existing set of recommendations for digitization efforts to the diverse body of legacy and born-digital digital newspaper content curated by libraries and archives (i.e., content not digitized according to the NDNP's standards). The project will also expand upon these recommendations as necessary to address the additional issues raised by legacy collections and born-digital acquisitions.

NDNP Standards and Legacy and Born-Digital Collections

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.[9]

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 technical requirements is in keeping with current accepted high standards for archival-quality digitization for image-based items, and prepares the collections for successful repository management as defined by the OAIS Model.[10]

The NDNP, then, is establishing best practices that have implications well 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.[11] They have been encoded at varied levels, use a diverse array of metadata schemas, and are arranged in a wide variety of file structures and repository systems. The NDNP technical guidelines do not currently provide explicit recommendations for readying such "legacy" and born-digital collections

[7] Based on analysis of our Case Studies (see Appendix B), our partners' newspaper collections will require focused and individualized attention, rather than cookie-cutter processes when preparing for ingest and long-term management.

[8] Inge Angevaare, "Taking Care of Digital Collections and Data 'Curation' and Organisational Choices for Research Libraries", *Liber Quarterly*, Vol.19, No. 1, April 2009. Available at: http://liber.library.uu.nl/publish/articles/000278/article.pdf, (last accessed 05/15/2010). Angevaare candidly addresses the difficulties of singlehandedly marshalling resources and expertise toward digital curation that even well-established research libraries are prone to face for preparing their collections for digital preservation.

[9] As identified in this proposal's accompanying example case studies, several of the Chronicles project's participating sites and Advisory Panel members have worked with newspaper collections that have been contributed to the first phase of the NDNP, and some are preparing to contribute titles to the second phase.

[10] Library of Congress, "NDNP: Technical Guidelines for Applicants." 2009. Available at: http://www.loc.gov/ndnp/pdf/NDNP_201012TechNotes.pdf, (last accessed 06/04/2010).

[11] As the Library of Congress has 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 (such as an electronic book), while others (such as electronic serials) contain (and/or are linked to) multiple digital content objects."

for preservation. Can institutions with legacy content and born-digital newspapers use the NDNP guidelines to help them to prepare their collections for preservation? Are there other ways that institutions can achieve the proper level of documentation and normalization of collections to facilitate preservation?

This project will study how best to prepare legacy collections and born-digital newspaper acquisitions for preservation.[12] We will also investigate the following question: for preservation purposes, what type and level of preparation is *essential*, and what type and level is *optimal*? We will document this range of preservation readiness options so that institutions with variable resources (funding and technical infrastructures) may ensure that they achieve the essential and aim for the optimal.

If data preparation guidelines aim only for the "perfect," curators at institutions with limited resources will be unable to implement them.[13] 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 must ensure that guidelines enable curators 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 *obsticles* to preservation by making demands that are higher than necessary and that curators lack the resources to implement.

Enabling the Exchange of Data Between Repositories

Once newspaper collections have been readied for preservation, how can we best exchange these collections between repositories? 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)[14] to social science data sets (DDI).[15] 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). Both have implications for the integrity of collections as they are moved from one repository to another.

Because libraries and archives use a diverse range of *data structures*, questions abound regarding how to ensure the preservation readiness of the resulting collections. The naming conventions used, the folder sizes, and the ways that the folders and sub-folders are stored all may have an impact on the exchange of that content between the local repository and a preservation repository.[16] An institution must assess the

[12] This work will also have implications for federating content for access purposes. Our team hopes to undertake new research following this project to study how to most effectively combine access and preservation in DDP networks for a more streamlined solution, but such work is outside of the scope of this project proposal.

[13] In "Taking Care of Digital Collections and Data 'Curation' and Organisational Choices for Research Libraries" (previously cited), Angevaare encourages institutions to take realistic measures and pursue a range of options that respect institutional capacity while simultaneously helping them to preserve their digital assets according to best practices, including leveraging collaborative partnerships with similar institutions. The Library of Congress, from its own experiences has endorsed similar approaches in their recent Board Agency Announcement (BAA—June 2010) as a Draft Statement of Objectives on Open Source Software for Digital Content Delivery, stating: "At no other time has the emergence of technology so directly affected how the Library acquires, catalogs, preserves, serves, and secures cultural records for its vast collections and holdings. The communities in which the Library participates are increasingly interested in collaboration and cooperation for more cost-effective management and distribution of the exponentially-expanding volume of content, both digital and analog. Shared responsibilities for stewardship and more complex roles and responsibilities are emerging as key factors for the future of cultural heritage institutions and government agencies."

[14] OASIS Emergency Interoperability: http://www.oasis-emergency.org/cap.

[15] Data Documentation Initiative (DDI): http://www.ddialliance.org/.

[16] For example, we have discovered at our partner sites that there are key differences in collections that are organized such that each issue of a newspaper is stored within its own folder that contains all of the files that are part of this newspaper issue (master scans, encoded files, metadata files) vs. collections that are organized such that each element of each issue are stored separately (all master scans in one folder, all encoded files in another, and all metadata files in another).

viability of the data structures it has used and ensure that all of the files that constitute a collection will remain both intact and identifiable when exchanged with another repository.

Likewise, libraries and archives use many different types of *repository systems* (e.g., Olive, CONTENTdm, DSpace, DigiTool, and home-grown 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.

The project team will study existing specifications for transfer (e.g., TIPR-RXP[17] and BagIt[18]) to assess their applicability to the genre of digital newspaper content. We will demonstrate successful transfers of content between the set of repository infrastructures represented by our partners (including Olive, CONTENTdm, DSpace, DigiTool, iRODS, LOCKSS, CDL's microservices). We will identify and resolve issues that impede such transfers, and will develop interoperability tools that interface with the repositories studied by this project team. These tools will be released under an open source license and broadly disseminated to the library and archives community.

As the project team conducts these data exchanges between repositories, it will document its findings in a comparative analysis of the three DDP solutions considered here. This analysis will facilitate better understandings of both DDP practices and the inherent strengths and challenges of each framework.

Distributed Digital Preservation (DDP) Repositories

Recent studies and national initiatives (i.e., 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."[19] The library community has concluded that "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."[20]

Some of the 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[21], Data-PASS[22]), 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.

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 built within

[17] http://wiki.fcla.edu:8000/TIPR/21.
[18] https://confluence.ucop.edu/display/Curation/BagIt.
[19] Fran Berman and Brian Schottlaender, "The Need for Formalized Trust in Digital Repository Collaborative Infrastructure." *NSF/JISC Repositories Workshop*, April 16, 2007. Available at: http://www.sis.pitt.edu/~repwkshop/papers/berman_schottlaender.html, (last accessed 05/10/2010); Please also see the following reports: American Council of Learned Societies. "Our Cultural Commonwealth: The Report of the ACLS Commission on Cyberinfrastructure for the Humanities and Social Sciences" *American Council of Learned Societies*, 2006. Available at: http://www.acls.org/cyberinfrastructure/ourculturalcommonwealth.pdf, (last accessed 05/10/2010); 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 05/10/2010), 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 digital preservation. Countless panels, presentations, and meetings in the digital library community have likewise addressed this topic.
[20] 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 05/10/2010).
[21] Controlled LOCKSS (CLOCKSS): http://www.clockss.org/clockss/Home
[22] Data Preservation Alliance for the Social Sciences (Data-PASS): http://www.icpsr.umich.edu/icpsrweb/DATAPASS/

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.[23] The collaborative strategy likewise should help content to withstand the myriad threats to its integrity, including large-scale disasters (e.g., wars, hurricanes Katrina and Rita, the 2003 power grid failure) and more isolated, local-level events (media failures, human errors, hacker activities, and smaller-scale floods and 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/Odem Institute partnership to preserve social science datasets using a PLN). Other experimental approaches show great promise, including DuraCloud[24] (DuraSpace's experimental cloud-storage-based environment) and LOCKSS-KOPAL[25] (a project to bridge 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 if the pricing was consistent across these options.[26]

Several open source technical frameworks currently enable institutions to preserve their content in such repository environments, including Chronopolis (iRODS-based), MetaArchive Cooperative (LOCKSS-based, with additional layered data management tools), and CODA (UNT's system, based on CDL's microservices). Each of these approaches varies in key areas such as ingest mechanisms, data management practices, and recovery options. However, most institutions do not have the information they need in order to evaluate the appropriateness of these three environments for the preservation needs of their collections. Are specific collection types (in terms of data, file, and repository structures) better suited for ingest into particular DDP frameworks? Are there different barriers to preservation that arise in each environment, and likewise, different strengths for preservation displayed in each? This project will study this issue by performing a comparative analysis of three production environments that use these frameworks. This analysis provide a systematic evaluation of the strengths of each system, as well as the challenges each system presents for particular types of collections.

Project Research and Development Outcomes

This research will result in *guidelines* for preparing digital newspaper collections for preservation, *interoperability tools* to facilitate the exchange of these newspaper collections between repositories, and a *comparative analysis* of the strengths and challenges posed by three distinct DDP frameworks for the

[23] Katherine Skinner and Matt Schultz, Eds., *A Guide to Distributed Digital Preservation*, Educopia, 2010. Available at: http://www.metaarchive.org/sites/default/files/GDDP_Educopia.pdf, (last accessed 05/10/2010).
[24] DuraCloud: http://www.duraspace.org/duracloud.php.
[25] LOCKSS-KOPAL: http://www.ibi.hu-berlin.de/forschung/digibib/forschung/projekte/LuKII.
[26] Pricing is not, of course, consistent across these options. Consider quotes provided to one institution with 5 TB of content in 2009. A leading vendor's quote to this institution was for $55,650 for the first year of preservation, and approximately $35K for each year in the following three years, for a total of $163,200 over a four-year period. The iRODs-based Chronopolis service, in contrast, quoted this institution a charge of $1000/TB/year, or $20,000 for a four-year period, and the LOCKSS-based MetaArchive Cooperative quoted this institution a charge of $1,000/year for membership, and an additional charge of $670/TB/year for storage, for a total of $14,050 for a four-year period. For more on the surveys, please see Skinner and McMillan.

preservation of digital newspaper content. In so doing, it will facilitate the long-term sustainability of this essential content genre for tomorrow's humanities scholars and researchers.

This project has the capacity to alter the digital preservation landscape for newspaper content. It also has implications beyond this content type. By fostering cultural memory organizations' capabilities to forge together to create community-owned and community-governed preservation activities such as the three DDP services in this project, we ensure that institutions have options beyond those offered to them by vendors at prices they often cannot afford *and* with "black box" types of restrictions that we as a culture cannot afford. Having options that include non-vendor based offerings helps to keep vendors' prices and offerings reasonable and thus increases the health of the field. We depend greatly upon the health of this field as we preserve the historical newspapers, both big and small, that chronicle our culture's history.

BACKGROUND OF APPLICANTS

This project team recognizes that the level and amount of work that we are proposing in this project is ambitious by any measure. We are overachievers and have a proven track record of accomplishing great things with small(ish) amounts of funding. The deliverables of this project are desperately needed, not just by the field at large, but also by the project participants. We are confident in our ability to conduct this research and finish the project deliverables within the proposed project timeframe.

Project Lead: Educopia Institute

The Educopia Institute is a 501(c)(3) organization founded in 2006 to serve and advance the wellbeing of libraries, research centers, and museums by catalyzing the advancement of shared information systems and infrastructures. Educopia assists and advises organizations in the creation of new digital means of preserving and providing access to scholarship and the cultural record. The strength of the Institute's approach comes from its decentralized goal of fostering the creation of successful cyberinfrastructure elements *in the cultural memory community*, rather than accumulating assets of its own. This approach builds knowledge and resources in the extended community of beneficiaries whom the Institute assists.

As the project's lead institution, Educopia will ensure that the project and its deliverables focus on open source and community-oriented frameworks that can be collaboratively implemented by cultural memory organizations. Philosophically, this approach empowers libraries, research centers, and museums and facilitates their active leadership and participation (rather than outsourcing) in the realm of digital preservation in ways that are consistent with their curatorial responsibilities in the print/physical realm.

Project DDP groups: MetaArchive, Chronopolis, UNT-CODA

MetaArchive Cooperative

The MetaArchive Cooperative provides trustworthy[27] low-cost, high-impact preservation services to help ensure the long-term accessibility of the digital assets of cultural memory organizations. The Cooperative is an independent membership association with 17 member institutions that functions as a community-owned, community-led initiative (hosted by the Educopia Institute). Its collaborative networks are comprised of libraries, archives, and other cultural memory organizations that seek to cooperatively preserve their digital materials, not by outsourcing to other organizations, but by actively participating in the preservation of their own content. Members identify collections that they want to preserve and prepare them for preservation according to leading standards. Using a technical framework that is based on the LOCKSS software and enhanced through modular data curation tools created and deployed by the MetaArchive Cooperative, these collections are ingested into a geographically distributed network where they are stored on secure file servers that are housed by the member institutions. These servers

[27] MetaArchive Cooperative. "MetaArchive Cooperative TRAC Audit Checklist." April 5, 2010. Available at: http://metaarchive.org/resources. The Cooperative conducted an external audit using TRAC in 2009, demonstrating MetaArchive's conformance in each of TRAC's 84 categories.

dynamically monitor and repair content, minimizing the risk that information be lost due to human error, technology failure, or natural disaster. The Cooperative actively engages in strategic alliances in order to foster interoperability and adoption of open source and community-based approaches. For example, the Cooperative is currently working with UNT and Chronopolis to promote interoperability between the LOCKSS software and the iRODS client. The Cooperative also consulted with and provided a model for many of the current DDP networks in operation, including PeDALS[28], ADPNet[29], and Data-PASS.

Chronopolis

Chronopolis is a digital preservation data grid framework developed by the San Diego Supercomputer Center (SDSC) at UC San Diego, the UC San Diego Libraries (UCSDL), and their partners at the National Center for Atmospheric Research (NCAR) in Colorado and the University of Maryland's Institute for Advanced Computer Studies (UMIACS). A key goal of the Chronopolis project is to provide cross-domain collection sharing for long-term preservation. Using existing high-speed educational and research networks and mass-scale storage infrastructure investments, the partnership is designed to leverage the data storage capabilities at SDSC, NCAR, and UMIACS to provide a preservation data grid (based on the iRODS open source framework) that emphasizes highly redundant data storage systems. Chronopolis has spent a number of years working through all aspects of digital preservation, from bit-level storage to high-level metadata management. They have been supported by multiple funding streams totaling several million dollars, from the Library of Congress' NDIIP Program, the California Digital Library (CDL), and related local organizations. They have also worked toward establishing strong collaborations with other national efforts, including the MetaArchive Cooperative (see above) and the Interuniversity Consortium for Political and Social Research (ICPSR).

UNT-CODA

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 CDL microservices 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 is implementing its integrated off-site replication in 2010. The CDL-based microservices that support the current instance of CODA are being experimented with for optimizing workflows across both instances of the repository.

Project Content Contributors

As documented in Appendix A: Content, each of our partners currently curates a number of digital newspaper collections. The partners of this project have been selected because they have diverse holdings that are, we believe, representative of the field. These are not institutions, in other words, that will bring "ideal" collections to the table. Instead, their collections will present myriad problems that we anticipate will help us to better understand and address the problems inherent in the larger field. Represented among our seven content contributors are NDNP leaders (University of Utah, Penn State, UNT), institutions that are just beginning their digital newspaper acquisition and digitization initiatives (Clemson University, Boston College, GA Tech), institutions with normalized collections (UNT), and institutions with diverse and un-normalized legacy digitized collections (VA Tech, Penn State). As documented in Appendix B: Case Studies, they also represent a wide range of file types, encoding practices, metadata implementation, and repository systems. This diversity, coupled with the Advisory Board's additional knowledge, will be a great asset to the project as we explore the challenges of creating preservation ready newspaper collections and ingesting them into three DDP frameworks.

[28] Persistent Digital Archives and Library System (PeDALS): http://pedalspreservation.org/.
[29] The Alabama Digital Preservation Network: http://www.adpn.org/.

HISTORY, SCOPE, AND DURATION

History

The "Chronicles in Preservation" project has its roots in a series of conversations between Penn State, the UKY, the UNT, the Library of Congress, and the MetaArchive Cooperative. This group—which includes three NDNP sites and the Library of Congress—recognized that the *Chronicling America* project is necessarily limited to the preservation needs for a specific sub-genre of digital newspaper content: newspapers digitized according to the NDNP's standards and housed within its centralized preservation repository at Library of Congress. These NDNP institutions reported having significant legacy collections that were digitized and encoded to evolving standards across the last two decades. They did not know what work they needed to undertake in order to normalize these collections for preservation purposes. They also did not know where they would preserve them, as the central preservation repository they were each using for their NDNP collections is necessarily restricted to NDNP collections.

The conversation continued within the MetaArchive Cooperative's membership. Member institutions reported having significant collections that fall outside of the NDNP scope, including digitized campus, local, and state-based newspapers and born-digital newspaper content. Members expressed their dismay at the lack of guidelines for selection, appraisal, data structuring, and other preservation-readiness activities for these newspaper collections. To date, only one member has preserved newspaper collections in the MetaArchive network, although most of our members curate such collections.

Throughout the planning process, both the project partners and Advisory Board have emphasized the importance of these three open-source DDP frameworks, as they provide a foundation for community-owned and community-controlled preservation networks. All of our project partners agree that U.S.-based cultural memory organizations today face critical decisions about their work that could have an enormous impact on the future of the field. Will cultural memory organizations choose to demonstrate leadership and accept responsibility for digital artifacts in a manner that is consistent with their work with physical artifacts, or will they outsource these tasks to other, non-library/archive/museum entities? In order to achieve the former, and thus maintain their role as content stewards in the digital age, cultural memory organizations need mechanisms that allow them to collaborate in efficient and sustainable ways. Open source distributed digital preservation frameworks such as the three studied here provide one of the most promising means to accomplish this goal, and as such, were of great interest to our project team.

Preliminary Research

The Educopia Institute has conducted preliminary research that revealed that institutions are creating and storing their digital files in wildly diverse ways.[30] Most institutions are not yet pursuing preservation activities but report that they anticipate undertaking preservation activities in the next three-to-five years. They report needing advice with regards to the appraisal, selection, and prioritization of materials for preservation. They also report needing assistance in moving content from the access-oriented repository systems in which it is housed (including vendor-based systems) into preservation systems. Above all, institutions report a desire to conduct their own digital preservation activities rather than outsourcing this core mission of their institutions. Fully 73% of respondents from both surveys report interest in "participating in community-based preservation networks," while only 30% report interest in "reasonably priced vendor-based solutions."[31]

The project team has also conducted extensive research into existing standards, specifications, and guidelines that provide foundations for the work we propose herein, as documented in "Methodology."

[30] Survey of Preservation Readiness for Museum and Archive Curators (MetaArchive, 2009); Survey of Preservation Readiness for Newspaper Curators (Educopia, 2010).
[31] Skinner and McMillan.

Scope and Duration

The project team will conduct a two-year research project (May 2011 to April 2013) to better understand and meet the needs of cultural memory organizations with regards to the preservation of their newspaper content. This research will result in *guidelines* for preparing digital newspaper collections for preservation, *interoperability tools* to facilitate their exchange between repositories, and a *comparative analysis* of three DDP frameworks. In so doing, it will enhance the long-term sustainability of this essential content genre for tomorrow's humanities scholars and researchers.

METHODOLOGY AND STANDARDS

The following three sections, which correspond to the three research questions of this project and their associated outcomes, document more fully the activities that we will undertake and the methodologies that underlie each component.

1. How can curators effectively and efficiently prepare their existing digitized and born-digital newspaper collections for preservation?

As described above, libraries and other cultural memory organizations curate a substantial body of digital newspaper content. The genesis of these collections is often a series of iterative and cumulative digitization and born-digital acquisition efforts with idiosyncratic and ad-hoc data storage structures that vary radically in their file types, structures, and metadata.[32] As our surveys of our project partners, Advisory Board members, and the broader digital library community have demonstrated, institutions have limited resources to expend on the normalization or restructuring of their legacy digital content.[33] With limited staffing and time, how can institutions prepare such collections for preservation?

We seek to encourage excellent preservation practices, but we also must ensure that the perfect does not become the enemy of the good.[34] If institutions believe that they are incapable of readying their content for preservation according to emerging standards and guidelines, they may not take any action at all—much to the detriment of our nation's digital heritage. If they instead may engage in an incremental process that allows them to begin preserving content now, while slowly and steadily building toward an optimal level of preservation readiness, they will be more likely to participate in preservation activities now. Engaging at all is an important first step that the vast majority of our nation's cultural memory organizations have not yet taken. Once institutions begin preserving content, they will begin building the requisite expertise and knowledge in this area to prepare new collections and normalize legacy collections according to optimal standards.

In this project, we will study and document how best to meet the preservation needs of these collections. We will produce a set of guidelines that explicitly differentiate between the *essential* and the *optimal* in preservation readiness activities and that document the incremental steps that institutions may take to move from the *essential* to the *optimal* level of preservation readiness in their local environments.

As briefly documented in Appendix B: Case Studies, a range of issues will be studied by the project team documented in the *Guidelines* white paper, including the following:

- **Acquisition:** As institutions acquire newspaper collections from news agencies (e.g., VA Tech's set of local and international digital newspapers) or from other cultural memory curators (e.g., UNT's Texas Digital Newspaper Program), should they demand that acquired content adhere to particular standards?
- **Appraisal and prioritization:** Although attention has been given to the selection, appraisal, and prioritization of newspaper content for *digitization*, none has yet focused on this process for

[32] By "data storage structures" we mean the entire range of methods by which data is stored, including directories, administrative metadata, and other data management techniques.
[33] 2010 Survey of Newspaper Curators (Educopia, 2010).
[34] "Le mieux est l'ennemi du bien" from Voltaire's *Dictionnaire Philosophique* (1764).

preservation. How should a curator weigh such variables as file formats, storage media, and encoding levels when making prioritization decisions? How can they assess which collections are at most risk?

- **Collection metadata:** Metadata plays different roles in access and preservation realms. What role should metadata schemas and standards (*e.g.*, PREMIS, METS) play in preserving newspaper collections? If content stewards do not ensure the use of current "best practice" standards, what are the ramifications? What metadata fields/components are *essential* for preservation readiness and which are *optimal*? How may institutions begin with the essential and build to the optimal over time?
- **File types and migration:** How may an institution assess its newspaper file types, know which of these file types are at risk or no longer supported, and know which file types will require migration at iterative stages of their preservation work? How can institutions migrate their newspaper files, and which of the resulting files (original and migrated) is most important to preserve? What documentation (metadata) does the institution need to provide for migrated files?
- **Data structures:** How might the file structures that a newspaper collection uses either enhance or compromise its preservation readiness? Are there principles that can be applied to make newspaper collections easier to ingest into preservation networks? For example, what is the difference between a collection that is stored such that all files are 1) named with standard and meaningful conventions and 2) organized in folders by title with issue-level subfolders (e.g., Clemson's proposed structure) vs. one that is stored such that files are 1) named with unique identifiers that have no specific relationship to the object's contents, and 2) are stored in one large folder that contains other, non-newspaper content (e.g., a DSpace repository structure such as GA Tech's)?
- **Documentation:** How might documentation aid in a collection's preservation readiness? Could institutions use descriptive documentation as a bridge between the *essential* and the *optimal*? For example, if an institution cannot yet produce item-level preservation-oriented metadata, might they store documentation regarding how their conventions and data structures currently work so that, when removed from the context of their local repository and preserved at the bit-level, such collections could be reassembled for access purposes from the preservation copy if necessary?
- **Intellectual property considerations**: What rights issues arise when preserving newspaper content in a replicated and distributed manner? How may content curators address IP issues effectively, both as they acquire and digitize collections *and* as they preserve existing collections? What standard language may be used in contracts or MOUs with news agencies and other content owners to ensure curators have necessary rights to preserve that content?
- **Costs of remediation:** What does it cost to ready content for preservation according to essential and optimal practices? What may an institution gain by preparing its content at the optimal level?

These and other relevant issues will be analyzed in detail by the project staff, project team, and Advisory Board during the project. There are two aims for this work. First, it will inform the work of each project partner as they ready their own collections for preservation, and second, it will provide a set of guidelines that may be used to assist a broad range of other institutions as they engage in preservation readiness work for their digital newspaper collections in their local environments.

The project team will begin studying preservation readiness issues in the Start-Up project phase by conducting an in-depth follow-up survey via videoconference with each project partner about the status and condition of each of the collections they are preserving during this project. The PI, Project Manager, and Chronicles Committee will review existing standards and best practices, including OAIS, PREMIS, METS, and the NDNP guidelines, and will perform a gap analysis to provide a comparison between these standards and our partners' current realities.

The Advisory Board will review the case studies and gap analysis, and will meet with the project team to determine the applicability of existing standards and to identify what additional considerations we need to address for this genre of content. The Advisory Board and project team will draft an outline of the *Guidelines* in July 2011 that documents the *essential* standards that must be met in order for newspaper

content to be considered preservation ready and the *optimal* standards that institutions should seek to meet in order to ensure ideal viability and usability for these collections in the future.

The project team will use the first full draft of the *Guidelines* (completed March 2012) to produce a "preservation plan" for each partner institution by April 2012. The project partners will use these plans to ready their collections for ingest, producing preservation-ready SIPs by August 2012. As they engage in this process, their individual findings will inform the evolving draft of the *Guidelines*.

The project team will produce the *Guidelines* in an iteratively reviewed process with the Advisory Board, and will post a semi-final version for a two-month public review in November 2012-January 2013. The resulting white paper will be published as a freely downloadable PDF as a project deliverable and will be publicized through presentations, and listserv/blog announcements. The Educopia Institute will also produce an open wiki based on the *Guidelines* that will be open to other groups that wish to edit, amend, or append information based on their own experiences in preservation readiness activities.

2. How can curators ingest preservation-ready newspaper content into existing DDP solutions?
The project team will explore a set of common repository exchange scenarios faced by institutions as they preserve their digital newspaper collections. As described in Appendix B: Case Studies, some of our partners store master and access copies together; others store them in two systems. Some bind metadata to digital objects (through METS wrappers); most have metadata stored separately from objects, and one has no metadata at all. Some have page-level metadata; others have issue-level or collection-level. Some institutions use in-house systems; others have outsourced encoding and access to proprietary systems. Each of these factors impacts the way that content is prepared for ingest and submission into any single repository and how it is exchanged with preservation systems, including those evaluated in this project.

Ambitious efforts have been undertaken toward achieving standardized and reliable exchanges of content between various centralized repositories using standards (PREMIS, METS) and specifications (BagIt) for stabilizing content exchange. Both PREMIS[35] and METS[36] schemas are accepted approaches for recording relationships surrounding digital objects and activities taken upon them over time. BagIt is an efficient, simple packaging and transfer format that incorporates a human-readable manifest file that lists digital objects and their checksums and serves as an authoritative inventory for content exchange.

A promising effort toward repository exchange was piloted by UIUC's HandS (Hub and Spoke) Project (2005-10).[37] The Project has created a preservation exchange workflow that models digital objects in standardized ways (meeting OAIS requirements for SIPS, AIPS and DIPs)[38] for transfer between various centralized and access-oriented repository infrastructures (DSpace, Greenstone, FEDORA, and Eprints). Once exchanged, these digital objects can be accommodated to internal repository specifications, and reorganized according to the previous profile specifications when needed for dissemination and exchange.

More recently, TIPR: Towards Interoperable Preservation Repositories (Cornell, NYU, FCLA)[39] has sought to exchange AIPs between centralized preservation infrastructures to ensure survivability and succession of content. TIPR has developed the Repository Exchange Package (RXP) specification (modeled on BagIt), a hierarchical packaging format consisting of PREMIS and METS-derived files and the corresponding content files. They are now piloting exchanges between their three repositories.

[35] PReservation Metadata: Implementation Strategies (PREMIS): http://www.loc.gov/standards/premis/.
[36] Metadata and Encoding Transmission Standard (METS): http://www.loc.gov/standards/mets/.
[37] Hub and Spoke Project (HandS): http://dli.grainger.uiuc.edu/echodep/hands/.
[38] Consultative Committee on Space Data Systems, *Reference Model for an Open Archival Information System, Pink Book,* CCSDS, 2009, available at: http://public.ccsds.org/sites/cwe/rids/Lists/CCSDS%206500P11/CCSDSAgency.aspx.
[39] Towards Interoperable Preservation Repositories (TIPR): http://wiki.fcla.edu:8000/TIPR/.

Similarly, MetaArchive, Chronopolis, and UNT have undertaken interoperability work to establish AIP transfers between LOCKSS and iRODS-based systems. Using open source scripts developed by UNT, MetaArchive and Chronopolis are using BagIt to retrieve, validate, and bundle content for exchange. The project partners also plan to explore interfacing a LOCKSS plugin with a BagIt structure to enable the exchange of AIPs from UNT's CODA repository into the MetaArchive's preservation network. This work will conclude by December 2010, and our findings will be directly relevant to this project.

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 ingested content from DSpace, CONTENTdm, Fedora, 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[40] and Fedora with iRODS.[41] 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.

In this project, we will study the complexities involved in streamlining such access-to-preservation repository exchanges. We will examine a range of issues, exemplified here by our preliminary research (see Appendix B: Case Studies). During these early investigations a number of questions arose regarding compatibilities between partner institutions' collections and access-oriented systems and the preservation systems studied here. 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 robust microservices-based approach for preparing SIPs to become AIPs extensible to 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 microservices interact with exchange packages that already include this information (e.g., Penn State's NDNP collections)?

To study these issues, the project's technical team will analyze the applicability of efforts such as HandS and TIPR to moving content between systems for meeting our project goals. In conjunction with our Chronicles Committee and Advisory Board, the project team will also study barriers to implementing PREMIS and METS and BagIt for our partners' collections and for these preservation environments. The project team will consider the needs of the participating sites' newspaper content, study the implications of other digital newspaper structures, and improve interoperability practices for this content genre's exchange with preservation repositories. We will build streamlined solutions with broad applicability that, where possible, expand on existing standards and tools, as described above, while handling the specific repository system challenges that are presented by common newspaper repository and file structures.

The Project Software Engineer will build interoperability tools to handle the validated exchange of content from the access-oriented repositories into the preservation frameworks represented in this project. Development work will take place in short cycles, with user feedback from each partner institution built into each iteration. This usability testing will be an ongoing and integral component of the Software Engineer's work. The exchange mechanisms will be made available through open source licensing/release and disseminated through GoogleCode, the Educopia website, and at least one conference paper. NDNP project participants will also share project results with the NDNP community during its annual meetings.

[40] See the DSpace project wiki for DSpace-SRB Integration: https://wiki.duraspace.org/display/DSPACE/DspaceSrbIntegration; See also iRODS project wiki for DSpace: https://www.irods.org/index.php/DSpace.
[41] See the iRODS project wiki for Fedora: https://www.irods.org/index.php/Fedora.

This project does not set out to establish a single unified workflow or exchange mechanism for preparing any given newspaper collection for ingest across all three preservation systems explored in this project. It does aim to reduce barriers to preservation by establishing systematic approaches for exchanging this content between commonly used access-oriented repositories and a set of mature preservation solutions.

3. What are the strengths and challenges faced when using three leading DDP solutions for preserving digital newspaper content?
This project will provide an evaluation of three leading technical approaches in the U.S. context (iRODS, LOCKSS, and CDL microservices) for institutions that want to preserve their diverse newspaper holdings in DDP frameworks. Each of these approaches has unique features and qualities that may be well suited to particular institutions' needs. This comparative analysis will also assist groups that host systems based on each of these three frameworks in their future development aims, as it will clearly document ways that each might improve or broaden its own preservation services.

Beginning in the Research Phase of the project the project technical team, the Metadata Advisor, and the Data Wrangler will closely study the selected preservation repository systems (Chronopolis, MetaArchive, UNT-CODA) and the export and import options currently available for each. During the Development Phase, this team will continue to study the issues, barriers, and successes that will arise in the data exchange process, and will use these collective findings to draft a *Comparative Analysis of Distributed Digital Preservation Frameworks*, using the partner institutions' collections as case studies. This draft will be shared with the Chronicles Committee for review and comment by December 2012.

The Comparative Analysis will document some of the main features of each system, explicitly analyzing both the underlying technologies (iRODS, LOCKSS, and CDL-microservices) and the specific production environments (Chronopolis, MetaArchive, UNT-CODA). Relevant features include the following:

- **Ingest:** As previously described, each of these three systems employs a different approach to ingest. For example, Chronopolis uses the iRODS client and has mainly transferred content using hard drives. LOCKSS uses web crawling, which requires that content be available (temporarily or permanently) via the Internet and be structured such that a web crawler may successfully ingest the content (e.g., using GET requests). CODA typically transfers content using hard drives and moves it through a specific data model. The project team will document the methodology of each ingest approach and what content structures best match each environment. It will also document the degree to which the project's interoperability tools efficiently facilitate repository-to-repository transfers into each environment.
- **Subsequent ingests:** How does each system handle "updates" to content, either due to a collection's growth or due to an intentional change made to a collection? Must the content be re-ingested as a new collection, or can the framework iteratively add to the collection? What are the preservation implications of each approach?
- **Data Modeling:** What processes are enacted on the data after it is ingested? For example, CODA runs all ingested content through a common set of processes, performing curatorial functions such as extracting information about each file or object and using it to populate METS records and CONSER-derived bibliographic metadata. Chronopolis and MetaArchive currently do not perform any automated functions that change the contributed collections, and instead expect the content curator to perform curatorial functions as part of the preservation readiness process for each collection. Can an institution that has created preservation-ready SIPs engage with the CODA system without incurring data changes in that environment, and is it possible for institutions that have *not* created such SIPs to responsibly preserve their content in the Chronopolis or MetaArchive environments? How does each system gauge the preservation readiness (or lack thereof) of collections it ingests?
- **Storage Environment:** How might the storage environment of each system bear on the preservation activities it needs to perform for newspaper collections? What are key differences between tape back-ups and spinning disk storage (e.g., for running automated checksums) or of data grid environments vs. servers, and how might those impact particular collection types in this genre?

- **Monitoring:** How does each system conduct its monitoring activities between its distributed copies? What implications might different methodologies (e.g., automated vs. human-based monitoring, or particular types of hashing or checksums) have for content curators and the collections they preserve?
- **Security:** What mechanisms do these systems use to ensure secure network-based environments for ingest, monitoring, and recovery? Do these systems employ data encryption, SSL, or other security technologies? Are there strengths and/or challenges in the use of these mechanisms for newspaper content? Are particular systems more or less secure, and might that have an impact on content that is considered sensitive by its content curator (e.g., material for which the institution has limited preservation rights and no access rights)? Do some systems include access components, and if so, what impact might that have for the preservation of content that is not cleared of IP restrictions?
- **Recovery:** What does content look like when it is recovered from any of these three systems? Are some recovery scenarios better matched to certain repository systems? What steps must a content contributor go through in order to recover their collections? Do the systems provide access components to assist users in reaching the preserved content? Do the systems only provide recovery content to the content contributor, who then must use that content to repopulate their local systems? What implication might this have for newspaper curators and the needs they have for their collections?
- **Scalability:** How does each system grow (both in terms of content and additional replication/monitoring partners), and what impact might this have as each system increases its content base over time? How does each system ensure its organizational stability, particularly where members/partners actively host components of the infrastructure? What happens if a host institution drops out unexpectedly?
- **Cost:** What are the real costs of operating each system and of preserving content in each system and how do those costs scale?

In our preliminary research, we studied the existing state of digital newspaper collections at our partner sites using a select set of example collections (see Appendix B: Case Studies). We analyzed the set of challenges this example content will present for ingest, archival storage workflows, monitoring practices, recovery strategies, and security in the three preservation environments. The findings from these analyses have helped us to scope some of the main topics that we will cover in the *Comparative Analysis*.

The analysis will depend upon the research undertaken in all areas of this project, including the preservation readiness surveys, planning, and implementation; the interoperability tools research and development, and the content ingest activities. Like the *Guidelines*, the *Comparative Analysis* will be published as a freely downloadable PDF on the Educopia website and will be publicized through presentations and through announcements on appropriate listservs and blogs. The Educopia Institute will also produce and host an open wiki based on the *Comparative Analysis* that will be open to the broader community to edit, amend, or append information based on their own experiences.

Evaluation

The Chronicles in Preservation project has three major research and development outcomes that require timely and thorough evaluation: the *Guidelines*, the interoperability tools, and the *Comparative Analysis*.

Guidelines to Preservation Readiness for Digital Newspapers

The *Guidelines* will be evaluated at iterative project phases. The project partners will use a draft of the *Guidelines* to help them scope and enact preservation readiness activities for each collection they submit for ingest. As these collections are submitted, the project team will evaluate the preservation readiness of each collection and will work with the project partners to determine what portions of the *Guidelines* may need revision or elaboration according to shortcomings in the readiness of their content or any confusion they experienced while preparing these collections. The Advisory Board will also evaluate this document through their comprehensive review in October 2012. The broader community will further evaluate the *Guidelines* through a public comment period that will take place from November 2012-December 2012.

Finally, the open wiki published by the Educopia Institute as a partner resource with the *Guidelines* in March 2013 will enable ongoing review and editing by the broader digital curation community.

The most effective marker of the *Guidelines'* long-term success will be its adoption by the broader digital library community. The Educopia Institute will monitor downloads and citations of this resource in an ongoing manner to continue measuring its impact in the future.

Interoperability tools

We have created evaluative measures for the interoperability tools throughout their study, development, project-based use, and release. The Technical Advisor (Mark Phillips, UNT) will oversee the Research phase of the interoperability study, and will evaluate the appropriateness of the Software Engineer's proposed design in conjunction with the Chronicles Committee and the lead developers for Chronopolis, MetaArchive, and UNT-CODA. The Software Engineer's development activities will be undertaken using agile development techniques, and will incorporate user feedback during each development iteration. These users will include the project partners who contribute content and the preservation repositories that ingest that content. The partners and preservation system administrators will further evaluate the tools as they are implemented and used to conduct repository exchange activities.

As with the *Guidelines*, the best measure of the tools effectiveness will come after their release via GoogleCode and dissemination by our extended project team. These tools should be helpful to any institution using any of the featured access-oriented repository systems that seeks to preserve content in one of the collaborative systems represented here. They should also assist other groups that use the underlying technologies (LOCKSS, iRODS, CDL microservices). The Educopia Institute will continue evaluating the success of these tools through their uptake and use within the broader community, as measured by downloads from GoogleCode and citations regarding their use.

Comparative Analysis of DDP Frameworks

The *Comparative Analysis* will be evaluated during its draft stages by the Chronicles Committee and by key representatives of each preservation framework. Success measures will be taken by the Educopia Institute via monitoring downloads and citations of the *Comparative Analysis* and monitoring the open wiki that is provided to the digital library community for comment and elaboration.

WORK PLAN

Deliverables

- White paper on *Guidelines for Digital Newspaper Collection Preservation Readiness*
- Repository interoperability tools
- Open source licensing and release of the tools for use by the extended PLN community
- *Comparative Analysis of Distributed Digital Preservation Frameworks*
- Presentations regarding this work at major conferences (CNI, others)

Activities

May 2011 – July 2011: Start-Up Phase

During this **start-up phase** of the project, we will engage in planning activities, set up our conference calls and meeting schedule, advertise positions, and hire staff.

May 2011-July 2011: The PI will work with the Technical Advisor, Content Advisor, Metadata Advisor, and Sustainability Advisor to establish an extensive survey and evaluation tool based on our pilot appraisal of partner collections (see Appendix B) to facilitate further information gathering regarding our partners' collections. This survey will be distributed to each partner in June 2011. The PI and Project Manager will meet individually with each partner to assist with collection assessments in June/July 2011.

June 2011-July 2011: The PI, Content Advisor, and Project Manager will revisit existing standards and,

using initial results from partner surveys, document their applicability for legacy and born-digital collections, including a gap analysis. We will share findings with the Advisory Board in July.

July 2011: Chronicles Committee meets
We will convene the Chronicles Committee (comprised of a lead from each partner institution) via videoconference to 1) review the project goals and deliverables, the project timeline, and the roles and responsibilities of each partner; 2) establish an Outreach Plan for presentations and dissemination of project results; and 3) review the partner survey results and their implications for establishing preservation-ready collections at each institution and exchanging those collections between their local repository infrastructures and the three preservation repositories (MetaArchive, Chronopolis, UNT-CODA). **Meeting Outcomes:** 1) shared understanding of the project activities; 2) Project Outreach Plan; 3) Case studies for all partner collections that document their current file types, encoding levels, metadata, file structures, and repository systems and the preservation readiness activities that each partner will undertake to normalize their content and ensure its long-term viability.

August 2011: Advisory Board meets
We will convene the Advisory Board to review their roles in the project and to help guide our initial drafting of the preservation readiness guidelines. To this end, they will 1) review the partner case studies (survey results); 2) provide feedback on our documentation of existing standards' potential applicability for legacy and born-digital collections; and 3) provide insights about reasonable guidelines that will meet the *essential* collection readiness needs and ambitious guidelines for *optimal* readiness preparation that the project team will factor into its work on outlining the guidelines during the project's research phase.

August 2011 – November 2012: Research Phase
During the **research phase**, we will conduct technical and organizational studies that will guide our documentation and technical development activities in the Development Phase of the project.
August 2011-September 2011: The Principal Investigator, Project Manager, Content Advisor, and Chronicles Committee will continue to study preservation readiness issues for digital newspaper content, building upon the NDNP guidelines and other identified standards. This group will use the partner institutions' collections as a base for understanding the range of challenges institutions may face in preparing their newspaper collections for preservation. The Advisory Board will help the project team to identify additional challenges that are not exemplified by these case studies. By September, the project team will deliver to the Advisory Board an outline of the white paper for review and comment.
August 2011-September 2011: The Project Software Engineer, in coordination with the Technical Advisor, the Project Manager, the Systems Administrator, and the Systems Programmer, will conduct a study of existing interoperability tools and specifications (including TIPR-RXP and BagIt).
September 2011-November 2011: The Project Software Engineer will work with the Project Manager, the Technical Advisor, the Metadata Advisor, and the Data Wrangler to study the repository systems (including Olive, CONTENTdm, DSpace, DigiTool) in which partner collections are currently stored, the structure of these collections, the preservation repository systems we will work with in this study (Chronopolis, MetaArchive, UNT-CODA), and the export/import options currently available for each.

December 2011 – April 2012: Transition Phase
The **transition phase** will focus on the transition from research to documentation and development work.
December 2011-March 2012: The Project Manager will work with the PI and Chronicles Committee to draft a white paper documenting the project's initial findings regarding preservation readiness for digital newspaper content. This white paper will provide a range of guidelines from the *essential* to the *optimal* that will address the needs of institutions of variable sizes and capacities for rectifying and normalizing their collections. This draft will be shared with the Advisory Board for review in March 2012.
December 2011-March 2012: The Project Software Engineer will continue studying the repository systems (local and preservation) and will begin experimenting with existing tools and specification to

identify barriers to repository exchange.
March 2012-April 2012: The Advisory Board will review and provide feedback on the *Guidelines* draft.
March 2012-April 2012: The Project Software Engineer, working with the Systems Programmer and Systems Administrator, will draft a development plan for the interoperability tools for review and approval by the PI, the Technical Advisor, the Chronicles Committee, and the Advisory Board. Once approved (March 2012), the Project Software Engineer will begin coding activities.
March 2012-April 2012: The Project Manager, the Data Wrangler, the Sustainability Advisor, and the Metadata Advisor will work with the Chronicles Committee to create a preservation readiness plan for each partner based on the initial white paper (*Guidelines*) research findings and will begin helping partner institutions to prepare their collections for preservation.

March 2012: Chronicles Committee and Advisory Board Meet
We will host a joint half-day meeting of the Chronicles Committee and Advisory Board to 1) evaluate the interoperability tools development plan; 2) elicit feedback from the Advisory Board regarding the draft of the *Guidelines*; and 3) evaluate the preservation readiness plans prepared by the Project Manager, Data Wrangler, and Metadata Advisor for each partner.
Meeting Outcomes: 1) Approved interoperability tools development plan; 2) approved draft for the *Guidelines* white paper; and 3) approved preservation readiness plans for each partner.

April 2012 – January 2013: Development Phase
The **development phase** focuses on the project's key technical and organizational development activities.
April 2012-August 2012: The Project Manager and Data Wrangler will work with local staff at each partner site to finalize readiness work for all collections, resulting in preservation-ready SIPs.
April 2012-August 2012: The Project Software Engineer will continue coding the interoperability mechanisms and will present work to the PI, Project Manager, Technical Advisor, and Chronicles Committee for iterative feedback and development cycles. As components of the mechanisms are complete, they will be used to complete test data exchanges between each repository system and the three identified preservation repositories (Chronopolis, MetaArchive, and UNT-CODA).
April 2012-October 2012: The PI, Project Manager, and Chronicles Committee will continue editing the *Guidelines* and will share a draft with the Advisory Board for review and comment in October 2012.
September 2012-December 2012: The Project Software Engineer, with assistance from the Systems Administrator, will use the interoperability tools to transfer the SIPs prepared by each member institution into each of the three identified preservation repositories (Chronopolis, MetaArchive, and UNT-CODA).
September 2012-December 2012: The PI, Project Manager, Technical Advisor, Project Software Engineer, Systems Programmer, and Systems Administrator will continue to study the issues, barriers, and successes that arise in the data exchange process, and will use their findings to draft the *Comparative Analysis of DDP Frameworks*, using the partner institutions' collections as case studies for challenges and strengths in each approach. This draft will be shared with the Chronicles Committee and with key representatives from each framework for review and comment by December 2012.
October 2012-December 2012: The Advisory Board will edit the *Guidelines* in October 2012. The Project Manager will incorporate the Advisory Board's feedback in November. The project team will then post the *Guidelines* for public review for a two-month period, inviting such review through major digital library/archive/museum listservs and blogs that will reach digital newspaper curator audiences.

January 2013 – April 2013: Wrap-up Phase
January 2013-March 2013: The PI, Project Manager, and Chronicles Committee will integrate the comments from the public review phase and will finalize the *Guidelines*.
January 2013-March 2013: The Project Software Engineer will finalize documentation, package code, and release the interoperability tools under an open source license through GoogleCode.
January 2013-March 2013: The Project Manager will work with the PI, Technical Advisor, Project Software Engineer, Systems Programmer, and Systems Administrator to finalize the *Comparative*

Analysis of DDP Frameworks, integrating the feedback from December 2012.
April 2013: The PI and project staff will write up and submit the project's final report.

STAFF

The project's PI, Chronicles Committee members, Staff, and Advisory Board are uniquely qualified to conduct this research and possess the requisite experience to manage and perform all proposed work.

Principal Investigator (25% cost match): Dr. Katherine Skinner, Executive Director of the Educopia Institute and program manager of the MetaArchive Cooperative, will serve as the Principal Investigator of the project and will supervise the project and its staff. Skinner has worked with the Educopia Institute since its inception in 2007. She has served as a Principal Investigator on three highly successful federal and private grant-funded projects in the past four years, two on digital preservation (NDIIPP, NHPRC), and one on access services (Mellon). She has extensive experience in digital preservation planning, policy creation, and implementation efforts, as well as in more general digital humanities work.

Project Manager (75%): The project manager will be responsible for daily project oversight and coordination. The PM will serve as a communications manager between project partners and will jointly conduct and document the preservation readiness surveys, assist with the *Guidelines*, help to establish preservation readiness plans, coordinate with each partner to ensure that content is prepared on schedule and to coordinate ingest mechanism user testing and implementation. A highly qualified candidate with experience in digital preservation has been identified and is available at the project start date.

Software Engineer (75%, 100%): A qualified software engineer will be hired to undertake research and programming work associated with project outcomes. These include the study of existing interoperability tools and specifications and repository system export and ingest mechanisms; development of the repository interoperability tools plan; implementation of the interoperability tools; exchange of content between repositories; and helping to write the *Comparative Analysis*. An identified candidate for this position with experience in the three DDP frameworks is available at the project start date.

Systems Administrator (15% cost match): Bill Robbins, Systems Administrator of the MetaArchive Cooperative, will work with the Project Software Engineer to evaluate existing tools, scope development, and facilitate the exchanges between repository systems. Robbins is a leading expert in Private LOCKSS Network creation and maintenance and has worked extensively with the Chronopolis team and the UNT team on fostering interoperability between their preservation infrastructures using the BagIt specification.

Systems Programmer (5% cost match): Monika Mevenkamp, Lead Programmer of the MetaArchive Cooperative, will work with the Software Engineer to undertake the research and programming work of the project, including the study of exchange mechanisms and the development of interoperability tools.

Data Wrangler (25%): A student will work to identify preservation readiness challenges for each partner institution, including file naming conventions, file types, metadata issues, encoding levels, and file structure issues. VA Tech has employed undergraduate student data wranglers since 2004 as part the MetaArchive's contract work and is uniquely situated to undertake this project work.

Content Advisor (3%): Dr. Martin Halbert (UNT) will serve as the Content Advisor. Halbert will work with the PI and the Project Manager to help prepare and implement the preservation readiness survey for partner institutions. The Content Advisor will also provide feedback and guidance throughout the development of the *Guidelines* documentation.

Metadata Advisor (2%): Hannah Tanner (UNT), an expert in metadata creation and management for

digital collections, will work with the PI and the Project Manager to identify and address metadata issues throughout the project work, especially focusing on the Preservation Readiness activities, including the partner survey development and implementation, the preparation of content at each partner site for preservation purposes, and the exchange of data between repositories.

Technical Advisor (2%): Mark Phillips (UNT), an expert in technical engineering and repository development, will oversee the work of the technical staff and will provide input throughout the interoperability tools study, development, and implementation. He will also help to conduct and write up the *Comparative Analysis* of DDP Frameworks.

Sustainability Advisor (5%): Gail McMillan (VA Tech), an expert in data curation for genre-based collections, will help to develop and implement the partner preservation readiness survey and the individualized preservation readiness plans for each partner institution.

Chronicles Committee Members (5% each): Each project partner has committed to assigning one or more of their staff members as a Chronicles Committee for the project. Chronicles Committee members and related staff from each partner institution will hold bi-weekly conference calls to address technical and organizational issues. All Chronicles Committee members will assign additional staff to the project to assist with collection preservation readiness preparation. Each of our Chronicles Committee members has included a letter of commitment in which they specify that they are committing 5% of their time on this project for the two-year period. The Committee is comprised of the following members, and CVs are included for each:

Mike Furlough (Penn State)
Cathy Hartman (UNT)
Gail McMillan (VA Tech)
Kenning Arlitsch (University of Utah)
Tyler Walters (GA Tech)
Emily Gore (Clemson University)
Bill Donovan (Boston College)
David Minor (SDSC)

Advisory Board (10 hrs each): The project includes four Advisors who are recognized experts in the field of newspaper digitization. These Advisors will help the Chronicles Committee outline and draft the *Guidelines to Preservation Readiness for Digital Newspaper Collections*. They will also evaluate the interoperability tools development plan and the preservation readiness plans for each partner and will review the *Guidelines* prior to its public release. The Advisory Board is comprised of the following members, and CVs are included for each:

Mary Molinaro (University of Kentucky)
Sue Kellerman (Penn State)
Bob Horton (Minnesota Historical Society)
Liz Bishoff (The Bishoff Group)

DISSEMINATION

Dissemination activities will focus on three primary areas: sharing the results of our preservation readiness studies, sharing interoperability mechanisms with other cultural memory organizations, and advancing knowledge of the distributed digital preservation approach to newspaper preservation.

To these ends, we will submit proposals to library and archive focused conferences (including CNI and iPRES), and will deliver at least four presentations during the project period about our findings. We will disseminate the *Guidelines* and the *Comparative Analysis* through relevant listservs, websites and presentations. We will promote use of the interoperability tools by making the source code and documentation openly available through GoogleCode and by publicizing this code via the Educopia website and relevant listservs (including iRODS, LOCKSS, and CDL user lists) and presentations.