Chronicles in Preservation Project

By: Martin Halbert and Katherine Skinner

Abstract:

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, have received $300,000 from the National Endowment for the Humanities to study, document, and model the use of data preparation and distributed digital preservation frameworks to collaboratively preserve digitized and born-digital newspaper collections.

Keywords: Management | Documentation | Security | Verification | Preservation of Digital Newspapers

Article:

1. Introduction

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 humanities 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 distributed preservation repositories?

2. Project aims

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. [1] The NDNP has developed preservation-oriented standards for current newspaper digitization practices.

This project will explore how these standards may be applied to 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.

3. Research questions

These issues are being investigated through the following research questions:

A. How can curators effectively and efficiently prepare their existing 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.

B. 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, DigiTool, and DSpace) and their ingest into DDP frameworks (iRODS, LOCKSS, CDL microservices).

C. 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 microservicesbased 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.

4. Acknowledgments

5. References

[1] Library of Congress. 2011. NDNP: Technical Guidelines for 2012 Awards. URL: http://www.loc.gov/ndnp/guidelines/NDNP_201214TechNot es.pdf (last accessed 09/26/2011).