

What we learned from the Historical Newspaper pilot

In 2015, WPLC undertook a pilot project to establish a platform for access to digitized newspapers from libraries around the state. The pilot involved both pre-digitized content and content digitized from microfilm at the Wisconsin Historical Society (WHS). The Wisconsin Newspaper Association (WNA) platform utilized by BadgerLink was identified as the repository for the digitized content. For further details about the project, please see the project description on the WPLC website:

<http://www.wplc.info/node/286>

While the content is not yet live in the WNA platform, we have learned a lot from the pilot so far:

This project requires a collaborative approach.

In order to complete the pilot, a number of partners were involved:

- DPI: Because the content is loaded into the BadgerLink platform for Wisconsin newspapers, it was important for DPI to be at the table.
- Wisconsin Historical Society (WHS): As the keepers of the preservation microfilm for many Wisconsin newspapers and as the agency providing scanning for the pilot, WHS provided information about holdings and collections as well as technical expertise.
- Wisconsin Newspaper Association (WNA): The platform for the project is managed by WNA. They also have strong connections to newspaper publishers around the state and also provided expertise in acquiring rights.

The partners and WiLS (as representatives of WPLC) met multiple times to discuss the pilot and creating a model that could be sustainable longer term.

Metadata is simple....as long as files are named correctly.

The WNA platform requires that files be named consistently with issue and page information. If done correctly, there is no need for the library to create any additional metadata for the uploaded content.

Copyright is less simple, but there is help.

Libraries interested in loading newspaper content will have to get permission of the publisher if the newspaper is from 1923 or later. However, the WNA has strong connections with newspaper publishers and can work with them to obtain these permissions.

Pre-digitized content may require new OCR, and OCR for new and pre-digitized content takes significant time.

Some of the content that had been digitized prior to the pilot needed new OCR processing in order to create searchable transcripts, which took significant time and specialized software to complete. While the computer processing time takes the majority of the time, there is a significant amount of staff time needed to move files along through the process. When working with pre-digitized content, staff time is also needed for renaming and reformatting files, such as separating or combining pdfs.

Batching content will be most efficient and cost-effective.

WNA charges \$.13/page for the first 50,000 pages processed in a single load and \$.11/page for all pages after 50,000. It therefore benefits us to do fewer, larger loads rather than many smaller loads throughout the course of the year.

There is strong interest in continuing the project.

Seven libraries have already expressed interest in loading their content into the platform once the pilot is done. Many libraries that were part of the pilot are interested in loading additional content, too. The pilot has shown that this platform is a solution to the long-standing problem of how to provide centralized access to historical newspaper content in Wisconsin.

How to continue: an approach for discussion

Given the strong interest in continuing the project, and, assuming that the WNA platform proves to be an appropriate repository for the content, we would propose that WPLC facilitate two bulk uploads of historical newspaper content annually and continue to pay the annual hosting fee for the collection.

These two bulk uploads will contain:

- A maximum of 100,000 total pages of scanned microfilm from the Wisconsin Historical Society.
- A maximum of 40,000 total pages of pre-existing scanned content provided by libraries.
- A maximum of 10,000 pages from any one library.

The contributing libraries will be responsible for:

- Costs for scanning.
- Costs for uploading content to the WNA platform.
- For any pre-existing scanned content: all reformatting, renaming, and OCR creation necessary to meet the requirements of WNA.

WPLC would be responsible for:

- Annual costs for hosting (roughly \$2,000-\$2,500 after loading 280,000 additional pages)
- Project management time to coordinate with libraries, partners, and governing committee (see below). (approximately 40-50 hours per year)

We would further propose that the Newspaper Workgroup and the Collaborative Digitization Workgroup be discontinued and that a new Digital Historical Resources Committee is formed to advise and assist with the ongoing development of the newspaper project and other collaborative efforts related to the digitization, presentation and storage of historical content. If the board approves this recommendation, WiLS will work with these workgroups to draft a charge and determine a meeting schedule and workplan for the committee prior to the next WPLC Board meeting. In addition, once the content is loaded into the WNA platform, WiLS will work with the pilot libraries and either the Newspaper Workgroup or the Digital Historical Resources Committee on finalizing the report of the pilot.