IFLA International News Media Conference
27-28 April 2017
Landsbókasafn Íslands – Háskólabókasafn (The National and University Library of Iceland), Reykjavík, Iceland

Fitting a Square Peg into a Round Hole: Migrating Legacy News Content to Open ONI

Jeffrey A. Knapp
Penn State University Libraries
University Park, PA USA
knapp@psu.edu

Andrew Gearhart
Penn State University Libraries
University Park, PA USA
amg32@psu.edu

L. Suzanne Kellerman
Penn State University Libraries
University Park, PA USA
lsk3@psu.edu

Abstract:

Providing access to historical newspaper content has remained a strategic priority for the Penn State University Libraries for decades. As early participants of the United States Newspaper Program (USNP) in the mid-1980s, successful awardees of three National Digital Newspaper Program (NDNP) grants in the 2000s, and early adopters to host an in-house newspaper repository for its own student newspaper and U.S. Civil War era titles, the online collection has grown as substantially as its user base. As with all systems the time comes when adopting a new one is both necessary and prudent due to system obsolescence, infrequent vendor upgrades, and maintenance and user needs. Such was the case in 2016 when the University Libraries decided to investigate the use of the Library of Congress' open-source Newspaper Viewer Software to replace its aging Olive Software ActivePaper Archive™ system. Building the necessary expertise and system infrastructure to host an in-house instance of the Library of Congress' web-based software application was essential to the success of the project as was adopting the Scrum project management framework. Moreover, the actual process to migrate legacy news content from one system to another is not an easy task even though both systems maintain the same data and perform similar functions. This paper chronicles the decisions, planning and design phases of the project, the adoption of new organizational methodologies and roles, and the migration processes now underway to transfer 500,000 pages of historical legacy newspaper content to an Open ONI system hosted at the Penn State University Libraries.

Introduction

In the early 1990s the Penn State University Libraries launched an investigation to consider using digital technology solutions to preserve its deteriorating paper-based collections. Among the most vulnerable in the collection were its archival materials and historic newspaper holdings. With limited staff expertise and funding to build the new technology infrastructure and software needed, the Libraries sought commercial vendors to provide these services. Many of these first-generation service solutions were often proprietary-based, resulting in a total reliance on the vendor to resolve technical issues, to improve functionality, or to respond to user demands for better interface features. Moreover, the Libraries had to yield to the technical demands of the software by following precise specifications for uploading digitized content for discovery and access. Utilizing a vended service for newspaper digitization and online access was the norm for over a decade, starting in 2003. However, in 2016 new library leadership and the release of the Library of Congress' open-source software ("Chronam") changed the way in which the University Libraries could provide access to historical newspapers. Additionally, in 2016, the Libraries adopted the Scrum project management methodology, which emphasized an agile teamwork approach to adapt the product development of the LC's Chronam software for the Libraries.

This paper chronicles the Libraries' technological journey from relying on vendor-based proprietary software solutions to an open-source application. Included in the paper are the steps currently underway to migrate legacy newspaper content; in essence, fitting a square peg (the proprietary legacy data) into a round hole (an open-source Open ONI solution).

Testing digital image technology for decaying library collections

The Pennsylvania State University Libraries' first undertaking to convert analog paper content to digital content occurred in 1992 just as digital scanning technology was emerging. As a member of a consortium of eight academic institutions known as the LaGuardia Eight[1]—in honor of their meeting site at New York's LaGuardia Airport—the University Libraries was eager to participate in developing digital technology solutions for libraries as a means to combat the pervasive problem of decaying library collections nationwide (Council on Library and Information Resources, n.d.). From the informal meetings of these eight institutions grew a commitment to investigate the current digital technology landscape and to set a strategy to develop local digital collections while agreeing to share their findings. Each institution chose a demonstration project to test the technology with the goal to "advance the use and utility of digital technology for preserving and improving access to intellectual works of national and international importance through cooperative action and a supportive multi-institutional infrastructure" (Council on Library and Information Resources, 1992). With funding from the recently established Commission for Preservation and Access[2] the University Libraries proposed to test the viability of digitizing archival materials. The remaining seven consortium partners choose different material types to test. Two archival

collections from the Libraries' University Archives were selected for the test: 1) a collection of Pennsylvania agricultural county agents reports; and 2) a series of business records and personal correspondence from prominent United States labor leaders and educators.

By June 1992 the Libraries received the approval and funding for its demonstration project from the Commission. Based on the favorable testing and research results published by Cornell University and the Xerox Corporation, the Libraries secured an agreement with Xerox to supply the DocuTech Production Publisher hardware and software, a first generation "print-on-demand" networked system, for its project (Kenney & Personius, 1992).

The defined scope of the University Libraries' demonstration project was to:

1) Test the ability to create similar digital images from a considerable variation of source documents;
2) Reconfigure portions of a dispersed archival collection while maintaining the original file integrity;
3) Transmit digital files over data networks; and
4) Provide print-on-demand capabilities for archival materials.

Difficulties plagued the project from its inception as the library staff struggled with the new electronic digital image system. The large, weighty, and cumbersome 11" x 17" flatbed scanner was installed as a networked system in the University Libraries while the printing and publishing component of the system was located across campus at the University Business Services' print shop. Archival source documents consisting of business records and correspondence were scanned as proprietary 300 dots per inch bi-tonal images and saved to optical disks which could be retrieved, edited, and later printed on-demand at an output resolution of 600 x 600 dpi. At a 1993 conference Kellerman reported that among the problems were: "countless system error messages, the lack of documentation to resolve problems, numerous software glitches, network printing software glitches, and equipment failures…" Moreover, regular system upgrades to improve functionality and efficiency halted the project repeatedly throughout the demonstration's 18-month time frame. With each system upgrade, images previously scanned and stored on optical disks had to undergo conversion. At one point, over 30,000 images had to be converted which was the largest repository of digital images that the Xerox technicians had ever attempted to convert! Surprisingly, in the end, the digitized content was successfully converted. At the time, no digital standards existed. Beyond the initial on-site training by Xerox technicians, no documentation or guidelines on how to scan various document types (e.g., line drawings, photographs, etc.) existed. Regular conference calls with consortium project managers were greatly welcomed as a means to discuss progress and to share "tips,"—in essence, building best practices and guidelines on the fly.

While the Libraries achieved a measure of success in digitizing archival collections, and having overcome a multitude of technical challenges, confidence and enthusiasm to attempt more digitization projects was high. For ten years, from 1992 to 2002, the Libraries continued its digital libraries initiatives using the DocuTech system. However, as new digital technology solutions emerged featuring faster scanning devices, improved computer systems and discovery and presentation tools, the Libraries sought new off-the-shelf digital solutions, which essentially ended the era of the Xerox DocuTech Production Publisher at the University Libraries.

OCLC's NEW newspaper digitization solution

On May 28, 2002 the Online Computer Library Center (OCLC) and Preservation Resources announced that they had signed an agreement with Olive Software to provide libraries with "the tools to digitize historical newspaper archives." This online announcement further explained that OCLC would become the "exclusive worldwide distributor of Olive Software to the library market with the goal of the new program to help libraries preserve historical newspapers and open the content to online searchable access." Meg Bellinger, vice president of OCLC Digital and Preservation Resources, stated that through this partnership OCLC was taking a "leadership role in helping libraries preserve some of their most valuable resources" (Quint, 2002).

By 2003 the University Libraries signed on with OCLC and set its sights on a new digital initiative to scan and deliver the backfile of the Penn State University's student newspaper, the *Daily Collegian*. A meeting of librarians and staff from the *Daily Collegian* earlier in the year confirmed the urgency to digitize the paper and to make the online historical *Daily Collegian* archive a reality (Cheney, 2013).

Prior to digitization, access to the historical backfile of the *Daily Collegian* was limited and cumbersome at best. The University Libraries' News and Microforms Library and the Special Collections Library served as the reference center for the multitude of questions received about Penn State student life, student social behaviors, university happenings, and sporting events that could only be answered from the pages of articles, pictures, and editorials found in the *Collegian*. As early as the 1930s, the Special Collections Library had compiled a card file index of "important" events and people found among the *Collegian* pages. In 1984, the News and Microforms Library started its own *Collegian* index—indexing different types of events than the Special Collections Library was recording. Often, however, both indices fell woefully short, omitting articles and information that researchers were hoping to locate. *Collegian* student reporters and editorial staff frequently needed article access to the historical issues for writing assignments. When their own *Collegian* morgue file was found inadequate, they too turned to the Libraries for assistance (Kellerman, 2007). As for physical access to the *Collegian* backfile, the paper was limited to the microfilm copies produced consistently each year starting in the 1970s, or found in fragile bound paper volumes housed in the Special Collections Library. For the earliest holdings, the issues only existed in paper format.

With the promise of an integrated newspaper system, by December 2003 the Libraries acquired and installed a single library instance of ActivePaper Archive™ from OCLC and Olive Software as the software solution for the *Daily Collegian* project. This system allowed for newspapers, regardless of format (e.g., microfilm, paper or PDF files), to be transformed to an XML repository structure that included dates, articles, photographs, and advertisement publication information.

The first title targeted for the project was the *Free Lance*, which was published from 1887 to 1904. Following the *Free Lance*, the project continued moving forward to include each of the successive title changes of the paper through 1940. The number of years targeted for digitization was determined by the internal library funding available at the time. In some instances, paper issues had to be converted to microfilm first before digitization, as was the case with the backfile of the *Free Lance*. In other instances, existing print negative microfilm could be scanned directly. In other instances, poorly produced microfilm or badly scratched

positive service copy film had to be rejected and paper issues assembled for re-filming (Kellerman, 2006).

A publication history of the *Daily Collegian*:

- *Free Lance*, monthly publication, April 1887 – April 1904 (not previously microfilmed)
- *State Collegian*, weekly, September 28, 1904 – June 10, 1911 (on microfilm)
- *Penn State Collegian*, weekly; semi-weekly, September 28, 1911 – May 30, 1940 (had to be re-filmed due to poor quality)
- *Daily Collegian*, daily, September 5, 1940 – current (on microfilm)

The initial success of the *Daily Collegian* digitization project in 2003 prompted the Libraries to continue to digitize more newspaper holdings from microfilm and host the e-issues using the Olive Software application. Today, the Libraries' Olive Software repository boasts 500,000 newspaper pages. Titles and holdings currently online include: the *Daily Collegian*, 1887–2010; several Penn State campus newspapers; a substantial backfile of a Pennsylvania weekly farming newspaper, *Lancaster Farming*; and a collection of 116 rural Pennsylvania titles covering the United States Civil War era, from around the 1830s to the 1870s (Pennsylvania State University Libraries, n.d.).

Since 2003, multiple modifications and upgrades to the proprietary Olive platform were pushed out to resolve technical glitches. However, the user interface remained the same. As internet searching advanced (i.e., faster, reliable search results offered by Google), library practitioners and users requested similar search features, including crowd source editing tools, for the Olive newspaper repository. In 2007, Google approached the University Libraries asking permission to make the Olive newspaper repository accessible via their new News Archive Search product. After several months of crawling, testing the data, and fine tuning search algorithms, the pilot testing ended without visible results. By 2014, the viability of retaining Olive's ActivePaper Archive™ platform to deliver historical news content was being questioned by the Libraries' administrative leadership as new platforms were being promoted, namely OCLC's CONTENTdm[3], Veridian[4], the Library of Congress' newspaper viewer[5], the National Library of Australia's Trove portal[6], and the Internet Archive[7]. Promised upgrades to address requested new access and navigation solutions never materialized. Similarly, requests to Olive to meet the University's new web accessibility standards[8] failed.

Open-source Solution: Open ONI

What the Libraries needed was a responsive solution. Based upon the experience with Olive, the Libraries recognized the difficulty in obtaining and maintaining a solution from a third party amid evolving requirements. Likewise, developing their own custom solution suffered from the expensive and potentially disastrous need to maintain every aspect of the project. A road that forms a solution where the burden of development and maintenance is shouldered by more than one organization but is as flexible as in-house development seemed to be an unattainable fantasy. However, in many ways, this is what has been offered by successful open source projects. Critics of Open Source software, in an attempt to discredit it as a viable alternative to vended software, will argue that it is not truly free. Playing off of the open source adage that the software is "free as in speech" (where one is free to use and express

oneself with it) and "free as in beer" (it costs nothing), critics will cite that open source software is "free like a puppy." While the puppy analogy is intended to make a critical argument about open source software, it instead offers a telling metaphor. When one cares for a puppy, one can train it. In fact, there are classes for puppies to improve their behavior. If a puppy is not feeling well, one can do their best to make the puppy feel better, or it can be taken to a veterinarian. In fact, in virtually every dimension of the health and well being of the puppy, there is a community of practice that can help one make the most of their time with their dog. Having access to such a community and various helpful products make living with a puppy much more feasible. Similarly, by selecting open source software, one can benefit from the conventions of others. Experiences are interoperable with those of other open source software owners and, therefore, no user is alone.

It was this collaborative community aspect that allowed the bridge from data to system to occur. The Libraries was in search of a product that could adequately present the treasures of historical periodicals published in Pennsylvania. In examining options, the Library of Congress' Chronicling America (Chronam) project came to mind, as it was the foundation of the National Digital Newspaper Program (NDNP). Although it met virtually every need, the community around the software was sorely splintered. Each organization that worked with Chronam ran a highly customized version that made leveraging changes to the codebase difficult. This is because all changes would need to be rewritten into the system if the organization were to upgrade their instance with the changes available in the base Chronam code. This is where the Open Online Newspaper Initiative (Open ONI) was born.

The Open ONI project focuses on coalescing the feature improvements that forced organizations onto their own separate codebase islands and prevented them from contributing back to the project and moving forward. While the project has not incorporated all of the changes that organizations made in their forks of Chronam, significant progress has been made to make Open ONI into a platform that can be utilized on a very customized basis without breaking from the master branch of the codebase. For example, a complete system for easily running Open ONI was established using Docker. In less than 10 minutes, one can have Open ONI running in a development instance. This allows developers to quickly get the code, run it, and see where changes can be implemented. This rapid time to develop is critical to the project's future growth, as it should ease the process of growing the user base through rapid prototyping and easing developer onboarding. This is a function of utilizing a continuous delivery model, where single changes are quickly implemented to see if they work, rather than making large updates to software in the hope that they will all work well and be well received in a major version update. Continuous Integration also works hand-in-hand with Agile development frameworks such as Scrum. The inclusion of continuous integration is not without peril, however. The technologies are new and rapidly evolving, meaning that developer resources are hard pressed to learn and keep up with the changing landscape.

Migration

A Multi-source Conundrum

While the Open ONI project addressed the features, it did not address the existing data. Systems often have their own unique format from which they can import data. Open ONI, derived from Chronam, relied upon the open standard of NDNP for defining the data

structure. Initially, a team worked to evaluate the XML data structures of the Olive system in an attempt to determine how to transform the data into a format that could be restructured into NDNP format. Unfortunately, a crosswalk could not be determined because the Olive data structures were largely composed of unspecified data. Without knowing what all of the data elements were, where they needed to go could not be determined.

To solve the problem, the team decided to go back to the most recent source material for images, publication metadata, issue metadata, and page level optical character recognition (OCR) data. In the case of materials that had already been submitted as part of the National Digital Newspaper Program, this involved the process of harvesting, verifying, and loading the batches from the Library of Congress Chronicling America instance into an Open ONI instance. For items that were not a part of the NDNP, things became a bit more difficult. To ensure that all of the publications and issues could get loaded, a publication was selected for a test run that possessed the greatest number of potential hurdles, including non-inclusion in the NDNP project. This significantly increased the timeline to get to the first successful batch import, however, it ensured that the hurdles could be cleared with all of the subsequent batches.

Making a MARC

Part of the process of loading an NDNP batch into Open ONI involves retrieving a MAchine-Readable Cataloging (MARC) record that corresponds to each of the titles in the batch. Because some of the newer titles (i.e., those that were loaded into Olive, but were not part of the NDNP) were not considered by the Library of Congress to be general interest newspapers, their corresponding MARC records also were not available through Chronicling America. This required a local patch to Open ONI to allow loading of MARC records from an alternative store. Additionally, the image sets (developed from scans of microfilm) only contained bi-tonal data. This violates the NDNP standard for image quality, preventing the use of the batch verification tool created for the Library of Congress. To this end, the appropriate XML files had to be generated from local catalogs and edited to make the structure of the MARC.xml files work with Open ONI. In large part, difficulties surfaced due to the ongoing transition between MARC and MARC21 metadata formatting and cataloger practices.

Building NDNP Batches

Working with a third-party vendor to recreate the OCR metadata (thereby improving it) and generating the NDNP batches, the conversation was difficult to say the least. The vendor was apparently subcontracting to another vendor, which led to a communication breakdown. Initial versions of the batch did not have the correct data structures, did not have all of the XML files, and were not using the correct naming conventions for the files. After several extensive emails and batch rebuilds, the exact structure, naming conventions, and XML file contents could be hammered out.

Scrum Project Management Methodology

In order to take on a project like this, the Libraries' Administration decided they would need to implement a process that differed from those previously used. Managing digital projects require a degree of flexibility and constant monitoring that differs from traditional library

services, so a project management system utilized in the software industry made the most sense. The Libraries began using the "Scrum" method, a method that exists within the overall set of software development values, principles, and practices known as "Agile."

Agile is a set of values and principles—not processes or tools—intended as an alternative to earlier "documentation driven, heavyweight software development processes," allowing greater flexibility and responsiveness (Robinson, 2013). Scrum is a well-established framework for developing and sustaining complex products, "while productively and creatively delivering products of the highest possible value" (Schwaber & Sutherland, 2016, p. 3).

Scrum was chosen in part because one of its key roles is that of the Product Owner, and the Libraries' administration was interested in dispersing ownership for technology services throughout the libraries and making services more responsive to their stakeholders (Estlund, 2016). Simply put, users are more likely to hold accountable an individual charged with the responsibility for a service than they would a department, and they are probably more likely to report issues, too. The Product Owner decides on overall directions for the product, making sure that it moves in the direction of what the stakeholders (users, etc.) want to see.

The other two key roles are the Development Team and the Scrum Master. The Development Team is a self-organizing, cross-functional team of developers to do the work of completing the tasks assigned to them in order to create or improve the online product. The Scrum Master ensures that the methodology is followed and acts as a "servant-leader" for the Development Team by clearing obstacles, making sure sufficient resources are available, and clearing backlogged tasks (Schwaber & Sutherland, 2016).

This system has taken some time for those in the Libraries who have no experience with project management, particularly of the web development variety. Academic readers of the literature about Scrum may have some trouble seeing its relevance to scholarly pursuits, since Scrum is designed around meeting the needs of customers and increasing the likelihood of sales. However, even though the Libraries are offering this product as a free resource for use by anyone does not mean that it is not worthy of being optimized for user satisfaction like any profit-seeking product. With public-serving universities seeking ways of assessing the impact of their programs, taking an entrepreneurial mindset for web products is a natural step.

For some librarians, especially those in public service roles, this process, including the Product Owner role, can be difficult to work into one's work routine. After all, if one is a liaison librarian responsible for providing research assistance to students and faculty, teaching, and developing collections, those are priorities that tend to demand one's attention more than strategically planning a digitized collection that only a few people are asking about. But this work is very important—not only from the perspective of providing access to our historical news content, but also for the experience it provides to librarians. Librarians, who have been relegated to the role of customers, by licensing and leasing content from vendors, can now experience the process of building a content database from the ground up, including understanding things like file structure, standards, and the way users are actually using the product.

Librarians have been recasting their skills in response to change for a long time (O'Beirne, 2013). Just as they have embraced utilizing the internet, teaching of a growing list of literacies, doing outreach in their communities, engaging first-year students, and curating

data, the very adaptability of librarians makes them good candidates for the role of Product Owner. Due to their skill at finding knowledge resources, they tend to be efficient learners, and evidence suggests that digital project management is a skill being included more frequently in librarian job advertisements (Maatta, 2014). Further, Agile/Scrum and the role of Product Owner is a good cultural fit for librarians, since it places priorities on customer satisfaction and empathy for the product user (Burba, 2016)—skills librarians spend a great deal of their time honing.

Conclusion

Pegs are often square because it is the simplest shape to cut with a saw blade. Similarly, round holes are the easiest to cut using a round drill bit. Although the shape of an object might be determined by the available technology at the time of its creation, new tools, new technologies, and new roles will always be adopted to reshape them.

News was first printed on paper when that technology was invented. Microfilming and digitization adapted that technology to new media as time progressed. Libraries can adapt to the role of digitizing their own collections for financial sustainability, librarians can adapt to new roles as product owners working with web developers to develop and maintain these collections, and digital files developed for one platform can be adapted to work with a new one. The future success of libraries and librarians will hinge on their adaptability in light of the continued development of new technologies and new roles.

Notes

1. LaGuardia Eight members included: Yale University, University of Southern California, Cornell University, The University of Tennessee, The Pennsylvania State University, Harvard University, Princeton University, and Stanford University.
2. The Commission for Preservation and Access established in 1985. Patricia Battin, head librarian at Columbia University was named it first president.
3. See: OCLC. (2017). *Managing and presenting digital newspapers with CONTENTdm*. https://www.oclc.org/content/dam/training/CONTENTdm/pdf/Tutorials/Working%20with%20Content%20Types/Managing%20%26%20Presenting%20Digital%20Newspapers.pdf
4. Veridian, see: http://www.veridiansoftware.com/
5. Library of Congress' newspaper viewer, see: https://sourceforge.net/projects/loc-ndnp/
6. The National Library of Australia launched the Trove portal in August 2008. It was developed as the Library's online discovery service, including the Register of Australian Archives and Manuscripts, Picture Australia, Libraries Australia, Music Australia, Australia Dancing, Pandora web archive, ARROW Discovery Service, and the Australian Newspapers Beta service. See: http://trove.nla.gov.au/
7. Internet Archive pilot project to digitize newspapers, see: https://archive.org/details/newspapers
8. See Penn State's website accessibility standards at: http://accessibility.psu.edu/webpagetools/

References

Burba, D. (2016). Agile by Design. *PM Network*, *30*(10), 58–63.

Cheney, D. (2013). The Campus Newspaper—From Newspaper to Digital Archive, the Libraries' Role and Responsibilities. Presented at the IFLA 2013 Satellite Meeting on Newspapers & Genealogy and Local History Sections, Singapore. Retrieved from https://www.ifla.org/files/assets/newspapers/Singapore_2013_papers/day_1_04_2013_ifl a_satellite_cheneydebora_from_newspaper_to_digital_archive_the_libraries_role_and_re sponsibilities.pdf

Council on Library and Information Resources. (1992, December). Board Approves Contracts as Part of Digital Preservation Consortium Activities, (51). Retrieved from https://www.clir.org/pubs/archives/cpanews/cpanl51.html

Council on Library and Information Resources. (n.d.). History. Retrieved April 14, 2017, from https://www.clir.org/about/history

Estlund, K. (2016, February 29). Product owners and library IT services. Retrieved from http://sites.psu.edu/librarynews/2016/02/29/product-owners-and-library-it-services/

Kellerman, L. S. (2006, May). *If You Build It, They Will Come*. Webinar presented at the PALINET Webinar.

Kellerman, L. S. (2007, June). *Digitizing Campus Newspapers: Building the Historical Collegian Archive*. Presented at the Pennsylvania Library Association College and Research Division workshop, Lehigh University, Bethlehem, PA.

Kenney, A. R., & Personius, L. K. (1992). *Cornell/Xerox/Commission on Preservation and Access Joint study in digital preservation*. Retrieved from http://agris.fao.org/agris-search/search.do?recordID=US201300722814

Maatta, S. L. (2014). Renaissance Librarians. *Library Journal*, *139*(17), 26–26.

O'Beirne, R. (2013). We've been here before: Repeating patterns of re-invention of librarians' roles in academic libraries. *AIB Studi*, *53*(2), 83–94. https://doi.org/10.2426/aibstudi-9039

Pennsylvania State University Libraries. (n.d.). Digital Newspapers. Retrieved April 15, 2017, from https://libraries.psu.edu/about/collections/digital-newspapers

Quint, B. (2002, May 28). OCLC, Olive Software Ally to Digitize Library Newspaper Archives. Retrieved April 14, 2017, from http://newsbreaks.infotoday.com/NewsBreaks/OCLC-Olive-Software-Ally-to-Digitize-Library-Newspaper-Archives-17172.asp

Robinson, M. (2013, August 26). What is Agile? Retrieved from https://agileinsights.wordpress.com/tag/agile-development/

Schwaber, K., & Sutherland, J. (2016, July). The Scrum Guide: The definitive guide to Scrum. Retrieved April 2, 2017, from https://www.scrumalliance.org/why-scrum/scrum-guide