Providing Access to Digitized Newspapers: A Case Study from the University of Illinois at Urbana-Champaign Library

A report by Kyle Rimkus and Kirk Hess (2014)

Abstract

In 2013, the University of Illinois at Urbana-Champaign Library assessed its preservation and access infrastructure for locally digitized historical newspaper collections. At the time, the library was locally serving 900,000 pages of web-accessible historical newspapers using an internally managed system, and 200,000 pages via the Library of Congress'(LC) *Chronicling America* project. When the library reviewed its repository architecture for locally digitized newspapers, they also conducted a user survey, performed an environmental scan of digital newspaper management systems at peer institutions, and established user requirements. Following this analysis, the library implemented the Veridian digital newspaper platform in 2014 and transferred its digitized newspaper collections into it, and all of its digitized newspapers are now available from a single access point. This paper provides a detailed overview of the library's assessment process, and a summary of the current status its digital newspaper repository services.

Background

In late 2012, the University of Illinois at Urbana-Champaign (UIUC) Library identified an internal need to assess its systems infrastructure for digitized historical newspaper collections. At the time, the Library was relying on two services to provide access to locally held historical newspapers digitized from paper or microfilm: the Library of Congress' (LC) *Chronicling America* portal (http://chroniclingamerica.loc.gov) and the UIUC Library's Illinois Digital Newspaper Collection (IDNC) site (http://idnc.library.illinois.edu).

These access systems received digitized newspaper content from two distinct project teams within the library. *Chronicling America*, for example, provides a home to more than 200,000 newspaper pages

digitized from microfilm under the banner of the Illinois Digital Newspaper Project (IDNP, http://www.library.illinois.edu/inp/), itself funded by the National Digital Newspaper Program of the NEH and the Library of Congress. *Chronicling America* relies on a robust storage and delivery infrastructure managed by the Library of Congress, with a public access system that allows users to conduct full-text searches across its vast corpus of newspapers from throughout to the United States, to zoom into individual page images, and to download PDF versions of those page images.

In contrast, the UIUC History, Philosophy, and Newspaper Library's IDNC project digitized and provided access to newspapers, many from Illinois or with a strong Illinois connection, that did not fall under the purview of federally-funded IDNP collection-building. In addition to Illinois-specific collections, the project focused on historically significant US farm weeklies published in the late nineteenth and early twentieth centuries, historic newspapers and trade journals published for the entertainment industry in the US between 1853 and 1929, and collegiate student newspapers from across the US. From 2005 to 2014, the library licensed a locally- hosted software platform called Olive ActivePaper Archive to provide its users with open access to over 900,000 pages of its IDNC newspaper content.

Originally, there was no overlap between the newspapers UIUC stored in its Olive ActivePaper Archive instance and those it contributed to *Chronicling America*. Thematically, most of the newspapers the library contributed to both systems were originally published in Illinois, and were intended to appeal to the same audience of students and scholars interested in state history. Splitting this content between two access systems did not directly benefit the library's patrons, as it produced two access points for UIUC newspaper collections. In 2013, the library began to reconsider this strategy.

Initial Concerns

While the solutions outlined above served the library well in its first decade of newspaper digitization, internal staff began to express concerns in 2012 related to the sustainability of the library's digital newspaper repository infrastructure, in particular if it should continue to rely on Olive ActivePaper as a software solution. Technical staff pointed out that improvements to the software were slow to materialize over the years, that it was running in-house in a deprecated Windows 2003 erver environment, and that other viable options had begun to appear on the market. To explore these concerns, the library charged the Newspaper Delivery and Preservation Working Group with assessing UIUC's repository architecture for digital newspaper collections. The working group consisted of the former Assistant Archivist for Music and Fine Arts, the Digital Humanities Specialist), the Preservation Librarian and group chair), the Manager of Infrastructure Management and Support, and the Visiting Metadata Librarian for Web-Scale Discovery. The working group conducted a comprehensive analysis, summarized below, before making recommendations.

Definition of Needs

As a first step in evaluating software platforms for digital newspaper access, the working group held discussions with internal library stakeholders to identify end user needs and desired management features. This permitted the group to frame a balanced analysis of software options for providing online access to the library's digitized newspapers. These criteria are listed below.

Features for the end user

- *Article-level text-search (article segmentation): Newspapers often contain articles that begin on one segment of a page and continue on another, frequently out of sequence (e.g. "article continued on page x"). The practice of mapping the image file coordinates for where articles begin and end, as well as associated character transcript metadata, is known as article segmentation. While extracting structured data from newspaper page images in an automated fashion is challenging*

and often produces inaccurate results UIUC provided article segmentation in its IDNC project, and preferred a solution that would continue to enable it.

- *Ability to download content as open access (OA) when applicable: Scholars do not always want to work with digitized OA content in a software delivery platform within a web browser, because many prefer to view or manipulate digital newspapers using third-party software.. UIUC sought to enable patrons to download portions of newspapers or entire issues of newspapers, when in the public domain, in a readily- accessible file format such as PDF.*
- *Ability to search across a repository of newspapers: UIUC wanted its patrons to be able to perform full-text searches across all of its newspaper collections.*
- *Ability to access the images as well as the text of page images and advertisements): Given scholarly interest in newspaper imagery such as historical advertisements, UIUC wanted to ensure that individual newspaper pages were presented in an effective user interface.*
- *Crowd sourced Optical Character Recognition (OCR) correction, transcription: OCR transcripts, especially those generated from digital files derived from murky or splotchy microfilm, are frequently inaccurate. Given strong community interest in historical newspapers, several libraries have experimented with cultivating "crowd sourced" OCR correction within their local communities of users. UIUC was interested in having this capacity.*
- *Ability to tag articles: Similar to the OCR correction feature, UIUC sought to enable patrons to enhance descriptive information about newspaper articles by allowing them to create and apply their own descriptive metadata tags.*

Features for the local administrator of content access/management system

- *Ability to divide the repository into collections of materials: UIUC sought a software application that could serve not only individual newspapers, but divide them into collections with common themes for ease of use.*
- *Use METS/ALTO as the native text encoding standard: The Metadata Encoding and*

Transmission Standard (METS) and Analyzed Layout and Text Object (ALTO) schemas, both maintained by LC's Network Development and MARC Standards Office, have emerged as the standard metadata and page-data representation schema for digitized newspapers. METS provides the structural and administrative metadata for a digital object while the ALTO extension contains OCR text mapping of the physical location of page elements such as articles and advertisements. In keeping with this best practice, UIUC sought a METS/ALTO compliant system.

- *Sustainable, low-cost production and management workflow. UIUC sought a system that would not bring with it substantial rising recurring costs to sustain ongoing maintenance and digital production of newly digitized newspaper content.*
- *Trusted access to files for researchers. UIUC sought a system that would provide reliable access to Illinois newspaper content that could be cited as a scholarly source over time.*
- *Portability of data.UIUC wished to ensure its own ability to move content out of any system under consideration, without undue difficulty or expense.*
- *Ability to quickly and easily update new content. UIUC sought a system that would allow its collection managers to bring collections of newly digitized newspaper content online with a minimum of difficulty and waiting time.*

In addition to this internal discussion, the UIUC Library conducted an informal survey of users of the Illinois Digital Newspaper Collections. Theysolicited respondents by placing a request displayed prominently at the top of its digitized newspapers website with a link to a brief survey querying their most highly valued features in a new UIUC newspaper site.

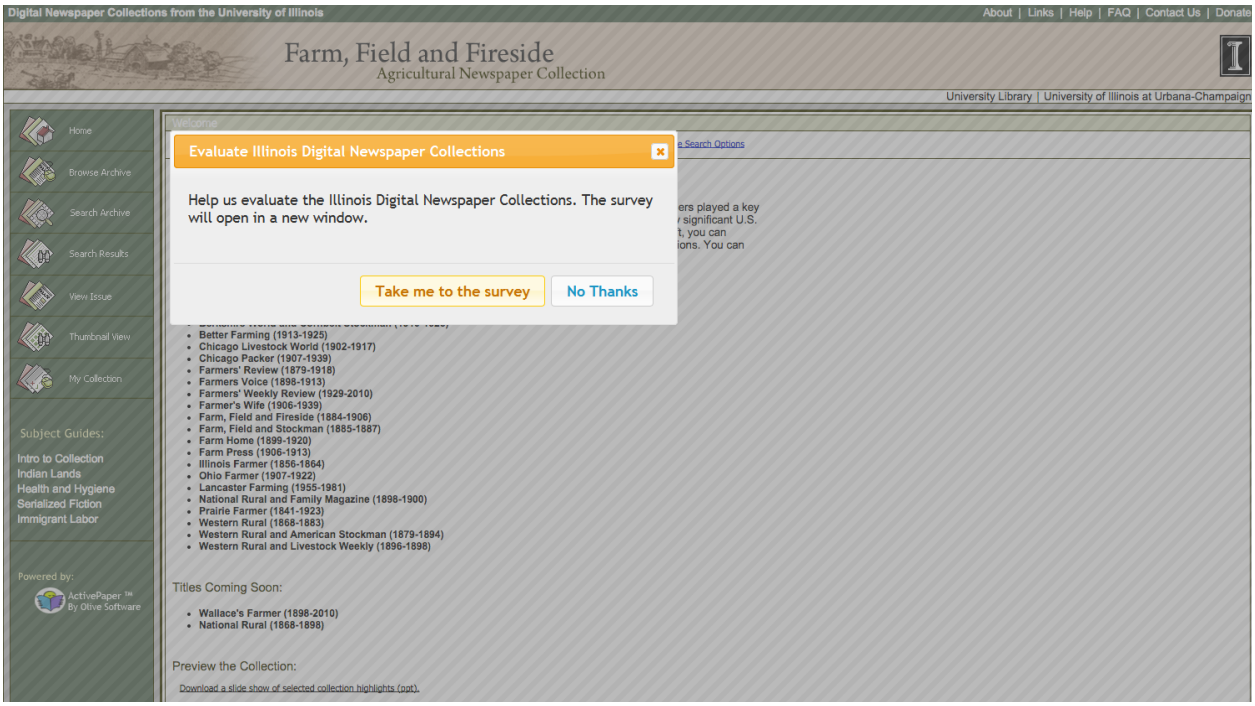


Figure 1. Surveying end users.

The survey, whose full results are available in Appendix I, attracted eighty-eight respondents, whose preferences largely aligned with the library's internal priorities. For example, when asked about features most important to them, the top response, 72 percent, focused on the application's search feature. Regarding new features, the two most strongly desired by users were OCR text correction and adding the ability to tag and comment on articles.

6. We are currently migrating the collections to a new management software system. The migration will add some new features to the collections. Which new feature would be of interest to you?	*Percent*	*Count*
Correcting text errors within the collection	**47%**	**41**
Sharing links and articles on social networking sites, such as Twitter	**10%**	**9**
Using social networking sites to follow news about the collections	**13%**	**11**
Adding tags and comments to articles within the collection	**48%**	**42**
If not listed above, what feature would you recommend?	**18%**	**16**

Figure 2. Survey results.

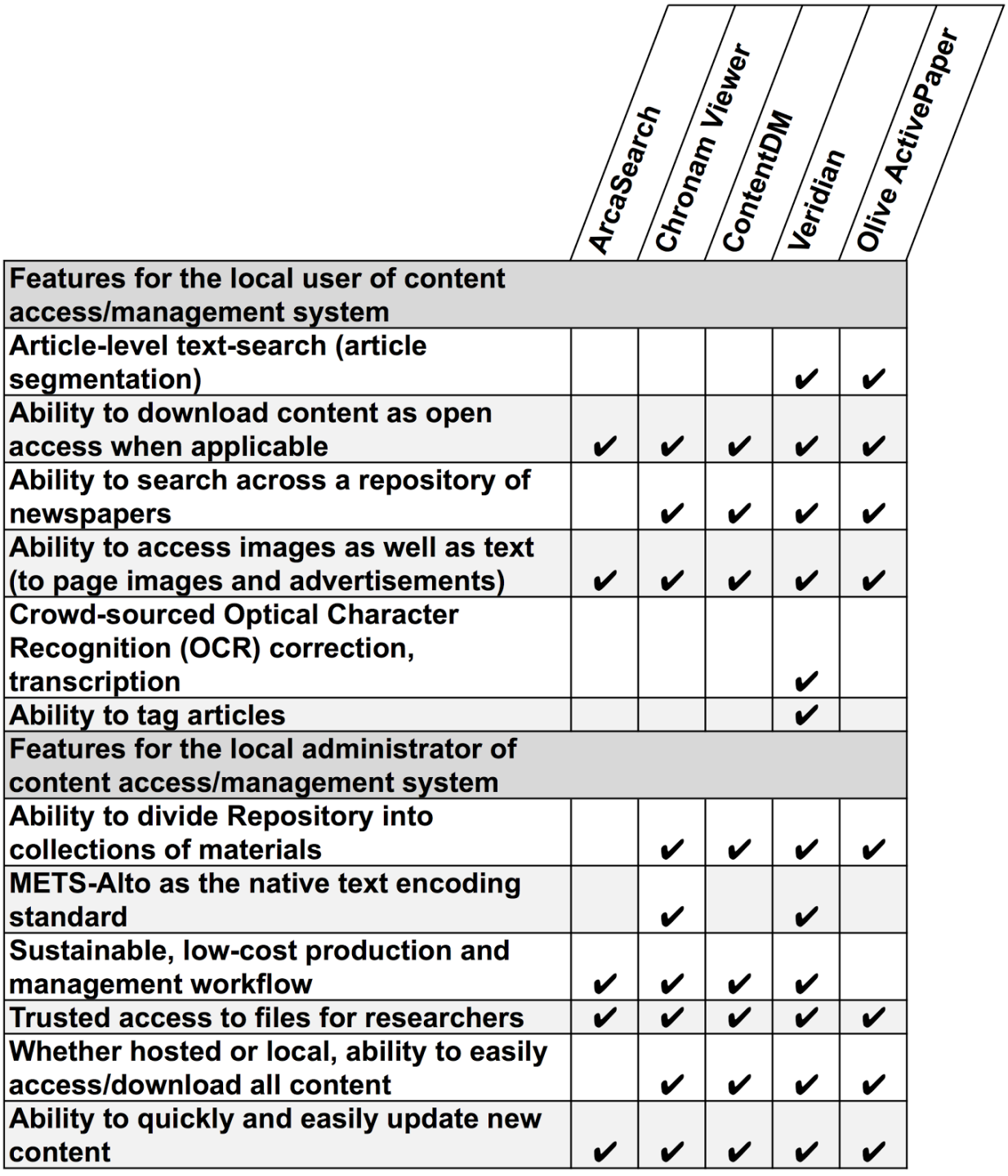
While informal, this user survey helped decision-makers at UIUC ensure that their internalpriorities for digitized newspaper delivery were in line with those of their end user community.

Access Systems Available for Academic Libraries

UIUC identified several access systems academic libraries commonly deploy to provide access to collections of digitized newspapers. They are listed below, with some comments on their ability to meet UIUC's needs (see figure 3 for a rubric view of this analysis):

- ArcaSearch (www.arcasearch.com/main/highered.html) is a PDF-based storage and access solution, and does not offer article segmentation or OCR correction. The user interface supports many document types and is not specifically tailored to scholars of historical newspapers.
- Chronam Viewer from *Chronicling America* (https://github.com/LibraryOfCongress/chronam) LC's *Chronicling America* project makes its viewer available as open source software. It has been adopted by other academic institutions such as the University of Oregon. The tool offers page-level but not article-level segmentation, and is METS/ALTO based. It does not offer crowd sourced OCR Correction.

- CONTENTdm (www.contentdm.org/) is a robust digital library platform used primarily to provide access to digitized visual resources and to create descriptive metadata for them. While some institutions use CONTENTdm for their newspaper collections, CONTENTdm was not intended specifically to host newspaper collections, and for this reason its search and discovery features are awkward for newspaper scholars.
- Veridian from Digital Library Consulting (www.dlconsulting.com/veridian/). Used most notably by the California Digital Library (http://cdnc.ucr.edu/cdnc), Veridian supports article-level segmentation and OCR text correction by users. In addition, it is METS/ALTO compliant.
- Local solutions. Some universities such as the University of Florida have built their own solutions for hosting newspapers. However, none of these appeared particularly viable for UIUC's needs.
- Olive ActivePaper Archive (www.olivesoftware.com/products/activepaperarchive.asp). Deployed at UIUC at the time of this research, Olive is also used by Committee on Institutional Cooperation (CIC) members Ohio State and Penn State for their newspaper archives. While it supports article level segmentation, it uses PrXML as its metadata standard rather than METS/ALTO and does not offer crowd sourced OCR correction.



	ArcaSearch	Chronam Viewer	ContentDM	Veridian	Olive ActivePaper
Features for the local user of content access/management system					
Article-level text-search (article segmentation)				✔	✔
Ability to download content as open access when applicable	✔	✔	✔	✔	✔
Ability to search across a repository of newspapers		✔	✔	✔	✔
Ability to access images as well as text (to page images and advertisements)	✔	✔	✔	✔	✔
Crowd-sourced Optical Character Recognition (OCR) correction, transcription				✔	
Ability to tag articles				✔	
Features for the local administrator of content access/management system					
Ability to divide Repository into collections of materials		✔	✔	✔	✔
METS-Alto as the native text encoding standard		✔		✔	
Sustainable, low-cost production and management workflow	✔	✔	✔	✔	
Trusted access to files for researchers	✔	✔	✔	✔	✔
Whether hosted or local, ability to easily access/download all content		✔	✔	✔	✔
Ability to quickly and easily update new content	✔	✔	✔	✔	✔

Figure 3. Requirements.

Digital Newspapers at Peer Institutions

UIUC scanned the online digital newspaper programs of its CIC peers and initiated conference calls with several, including Olive users Penn State and Ohio State, plus Indiana University, Purdue University, the University of Michigan, and the University of Nebraska. Beyond the CIC, UIUC spoke with newspaper staff from the University of Oregon and the California Digital Library to learn from their own robust digital newspaper repository programs.

At the time of these conversations, the UIUC newspaper program found itself in a situation similar to that of Penn State and Ohio State, all users of Olive ActivePaper. These users of Olive ActivePaper were generally satisfied with how the software operated and displayed its content, and it seemed to meet most libraries' basic needs; however, it has not evolved very rapidly in recent years, and certain users expressed interest in adding new user-centered features such as OCR correction and custom tagging of articles. In the CIC, however, there did not appear to be another digital newspaper program of similar magnitude to the Illinois program.

Like UIUC, the University of Oregon newspaper program is also an NDNP partner. Unlike UIUC, Oregon has chosen to use the open source NDNP viewer as its access tool. Oregon has customized the software and is committed to maintaining their installation of it. While this was considered as a viable solution by UIUC, it does not at present meet certain very important user needs identified by the working group, such as article-level segmentation.

In contrast, the California Digital Library is one of the United States' most significant users of the Digital Library Consulting (DLC) Veridian product. Their view of the software was very positive, based on end user feedback and their own experiences working with DLC staff. They adopted Veridian in 2008, and believe that DLC has a stable business model. This recommendation, in addition to Veridian's impressive feature set as outlined above, led to UIUC's decision to adopt it as its digital newspaper repository

platform.

A Platform Change

In reviewing the environment of digital newspaper repository solutions, the Newspaper Delivery and Preservation Working Group recommended shifting UIUC's platform for locally digitized content from Olive ActivePaper to Veridian. It found that Olive lacks academic libraries as the primary focus of its business model and has been slow to add new features in the six years of its use at UIUC. Olive met UIUC's core end-user needs when it was first implemented, but the PrXML standard it uses to represent article-level metadata has not been broadly adopted by library practitioners.Additionally, Olive has not indicated that they intend to use the preferred METS/ALTO standard natively. Lastly, Olive does not seem poised to offer features such as crowd sourced OCR correction or the ability to tag articles. These were the primary motivating factors that motivated UIUC to recommend adopting Veridian.

UIUC saw Veridian as an attractive option because its parent company Digital Library Consulting caters specifically to academic libraries, with Veridian sold specifically to those who manage large collections of digital newspapers. It provided all of the features the working group defined as critical needs. While retaining features from Olive such as article-level text-search (article segmentation) and the ability to divide a repository of materials into individual collections, it also offered forward-looking features such as crowd sourced OCR correction, the ability to tag articles, and METS/ALTO as its native text encoding standard. On the information technology management side, Veridian fit in with UIUC's preference for investing in repository solutions that could be used out-of-the-box with a minimum of customization.

Implementation

UIUC began its Veridian implementation October 2013. It hired a part-time graduate assistant coordinator to manage the Olive-to-Veridian conversion activities in collaboration with the Digital Newspaper Project Assistant. Digital Library Consulting wrote a PrXML to METS/ALTO crosswalk script, and converted

approximately 900,000 of UIUC's PrXML files into METS/ALTO by mid-December 2013. Subsequently, UIUC ingested the converted metadata and page images into the Veridian platform and reviewed the newly loaded content.They discovered a number of quality problems such as missing issues, unsegmented mastheads, zooming problems due to low-resolution images, and incorrect segmentation. Project staff worked with Veridian staff to fix these problems.

After the metadata conversion and content ingest, UIUC staff spent January and February 2014 fixing bugs and customizing the site's end-user interface with a new banner and color scheme. Digital Library Consulting helped the library integrate its five Olive collections into a single portal, and to create a landing page for each collection along with accompanying pages for documentation and acknowledgements. In addition, UIUC ingested all 200,000 pages of its Illinois Newspaper Project collections into Veridian. To accomplish this, the library took files digitized according to the LC's *Chronicling America* specifications, derived page and issue-level PDFs from them, and ingested content, metadata, PDF files, and JPEG-2000 image files from into Veridian. The resulting site was soft-launched at the beginning of February 2014.

By the end of February, UIUC began marketing by distributing informal announcements to local and national email lists notifying scholars and colleague institutions of its new newspaper portal. Veridian's built-in Search Engine Optimization (SEO) feature, which was not available in Olive ActivePaper, allowed Google to fully index the repository and make newspaper content available to web searches. By April 2014, the library had set up social media feeds to promote its newspaper repository on Twitter, Facebook , and Pinterest In addition, the library marketed the site through an official press release distributed to state and local genealogical organizations and historical societies. Local media were contacted via phone, which resulted in a newspaper article and a staff radio interview on local public radio in July 2014.

Impact

The library's SEO and marketing efforts succeeded in attracting new users to its newspaper collections, as shown by tracking data for visits on both the Olive ActivePaper and the Veridian sites. From the beginning of February through the end of June 30, 2014, the Veridian IDNC site had over 70,000 sessions and 49,000 users. By comparison, the previous Olive iteration of the site, tracked over the same time period in 2013, had 14,000 sessions and 7,000 users, a five- and seven-fold increase, respectively (see figure 4 below). In addition, the crowdsourcing OCR correction feature attracted over 120 registered users who have corrected over 42,000 lines of text. A column on genealogy research published in a local newspaper underscored the popularity of this feature.

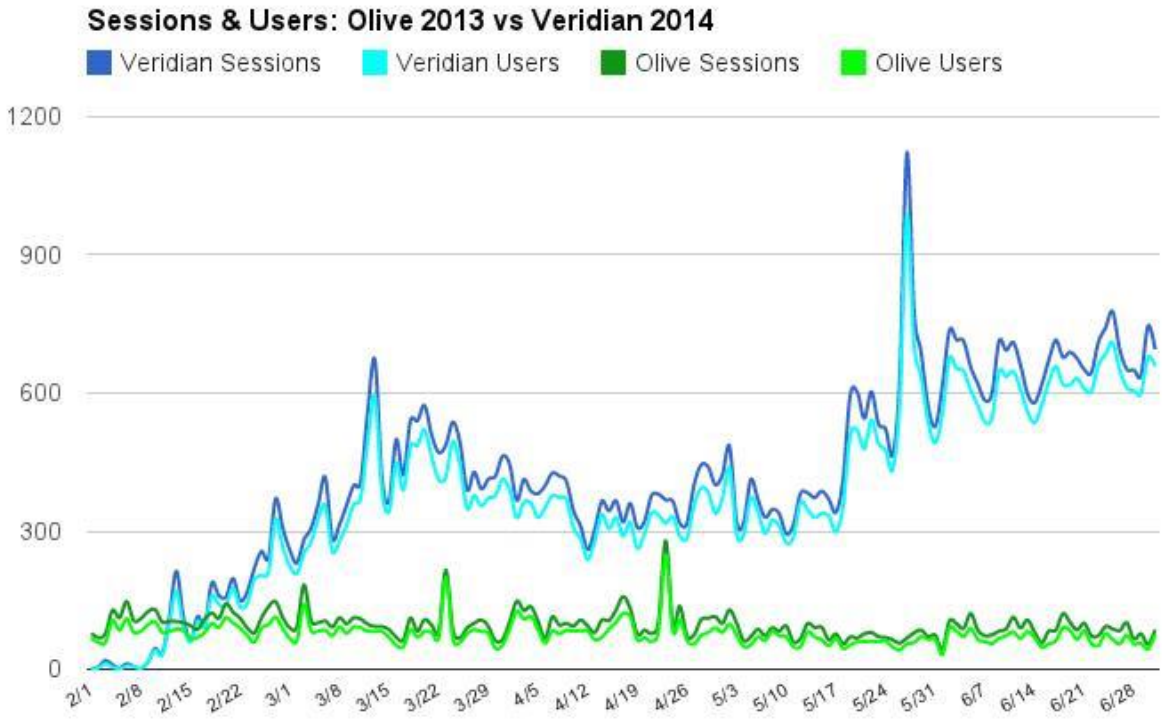


Figure 4. User statistics.

Next Steps

In the next year, the library hopes to allow more scholars to mine its unique newspaper content for use in research, by using Veridian's underlying Apache Solr search engine and its native text-mining application

programming interfaces. The library has been approached by the Viral Texts project, "which seeks to develop theoretical models that will help scholars better understand what qualities—both textual and thematic—helped particular news stories, short fiction, and poetry 'go viral' in nineteenth-century newspapers and magazines," and is investigating regenerating OCR for its digital files with the hope of producing improved output, a strategy used to some success by the Early Modern OCR Project (eMOP). In addition, the library plans to branch out from digitized newspaper content by ingesting born-digital editions of the local student newspaper the *Daily Illini* (http://www.dailyillini.com).

The UIUC Library will continue to emphasize marketing, user outreach, and an improved experience for its users of digital newspaper collections. In mid-2014, the library's IDNC team received a small internal marketing grant that it intends to use as prize money for a text correction contest for Illinois faculty and staff. Digital Library Consulting has improved its PrXML to METS/ALTO conversion script, and has worked with UIUC to improve Veridian's interface, especially where incorrect zoom levels have made certain pages difficult to view in Illinois's repository. Segmentation quality issues are a continuing problem, but these cannot be fixed very easily without a tool for redrawing boundary boxes and outputting updated ALTO elements; UIUC is hopeful that this will become a future crowdsourcing feature in Veridian.

Conclusion

The UIUC Library's 2013 investigation of its digital newspaper repository platform consisted of a full assessment of user and end-user needs and a comparison of available digital newspaper software platforms. At its conclusion, the library recommended discontinuing use of the Olive ActivePaper platform in favor of Digital Library Consulting's Veridian solution. Veridian offered the library several benefits, among them its METS/ALTO metadata standard, and its strong features for end-user engagement such as article tagging and crowdsourced OCR correction. In early 2014, Veridian was

implemented as the library's home repository for digitized newspapers from its local collections. The library also ingested all UIUC newspaper content previously only available in LC's *Chronicling America* site. As a result of the library's migration to Veridian, all of its digitized newspapers became available in a single user portal, and it found a dramatic spike in usage. This rise in usage was credited to Veridian's SEO features, which aggressively index its newspapers in Google, and the Library's own outreach and marketing efforts.

Bibliography

Justin Littman, "A Technical Approach and Distributed Model for Validation of Digital Objects," *D-Lib Magazine* 12, no. 5 (2006), dx.doi.org/10.1045/may2006-littman.

Justin Littman, "Actualized Preservation Threats: Practical Lessons from Chronicling America," *D-Lib Magazine* 13, no. 7-8 (2007), dx.doi.org/10.1045/july2007-littman.

Thomas McMurdo and Birdie MacLennan, "The Vermont Digital Newspaper Project and the National Digital Newspaper Program," *Library Resources & Technical Services* 57, no. 3 (2013): 148–63, dx.doi.org/10.5860/lrts.57n3.148.

Robert B. Allen, Weizhong Zhu, and Robert Sieczkiewicz, "What to Do With a Million Pages of Digitized Historical Newspapers?" (Illinois: iConference 2010 Papers, February 3, 2010), accessed August 13, 2014, http://hdl.handle.net/2142/14932.

Caroline Danielset al., "Community as Resource: Crowdsourcing Transcription of an Historic Newspaper," *Journal of Electronic Resources Librarianship* 26, no. 1 (2014): 36–48, dx.doi.org/10.1080/1941126X.2014.877332.

Kenning Arlitsch, L. Yapp, and Karen Edge, "The Utah Digital Newspapers Project," *D-Lib Magazine* 9, no. 3 (2003), dx.doi.org/10.1045/march2003-arlitsch.

Joan Griffis, "Illinois Ancestors: State Newspapers Have New Website," *News Gazette*, accessed August 13, 2014, www.news-gazette.com/living/2014-06-25/illinois-ancestors-state-newspapers-have-new-website.html.