Case Study: Conversion from Olive ActivePaper to Veridian Software

Illinois Digital Newspaper Collections

The University of Illinois at Urbana-Champaign University Library recently launched a new version of their digital newspaper archive, Illinois Digital Newspaper Collections (IDNC). The IDNC is a free online archive of digitized historic newspapers and trade journals organized in 4 different collections. The IDNC includes interactive features allowing users to tag articles, correct OCR text, and share on social media.

BACKGROUND

For the past 10 years the Illinois Digital Newspaper Collections was made available online through ActivePaper Archive (APA) by Olive Software. The site has many historic newspaper titles, with article-level segmentation, in Olive's PR XML format. The collection includes historic Illinois newspapers, farm weeklies, entertainment publications, and college newspapers that together paint a colorful picture of eastern and mid-western America in the 19th and early 20th centuries. The rich and historically significant content within the collection drew the interest of university researchers, faculty, staff and the Illinois community in general. Users who wanted to know what campus and community life was like during this time period could access the collection online, free of charge, and conduct searches based on people, places, events, and dates of interest.

In addition to this the university was awarded three National Digital Newspaper Program (NDNP) grants, in 2009, 2011, and 2013, to digitize culturally significant Illinois newspapers. With this funding the university, as part of the Illinois Newspaper Project, was able to digitize another 13 newspaper titles for inclusion in Chronicling America. These titles totalled approximately 200,000 newspaper pages in the METS/ALTO format, but with only page-level display.

SITUATION

In September, 2012, University of Illinois at Urbana-Champaign Library formed a Newspaper Delivery and Preservation Working Group to discuss the sustainability of the Library's repository architecture for managing the preservation of and access to digital newspaper collections. The working group evaluated the APA software it was using for the Illinois Digital Newspaper Collections and decided it was not meeting the needs of the library and the users. Furthermore, as the technology was aging, it became increasingly important to migrate off the outdated and unsupported Windows 2003 servers for which the software was built. The working group evaluated multiple other software solutions, including upgrading to Olive's APA version 5. After consideration they recommended the university chose the Veridian Software, developed by DL Consulting.

SOLUTION

The Veridian Software provided a compelling alternative to Olive for a variety of reasons:

- Veridian is highly scalable and presented the only reasonable conversion option for this large, 900,000 page collection of article-level newspapers.
- DL Consulting has the capability to convert Olive PR XML data to article-level METS/ALTO, the industry standard used by most large newspaper digitization projects including the Chronicling America project at the Library of Congress.
- The University of Illinois was able to include their NDNP data in the new collection, alongside their existing Olive-produced data, because Veridian accepts data in the METS/ALTO format.
- For several years Veridian has been the software of choice for newspaper digitization projects at many large U.S. university libraries, including those at Princeton, Columbia, and Cornell.
- There are a number of very large Veridian collections at national libraries around the world, including those in New Zealand, Singapore, Viet-

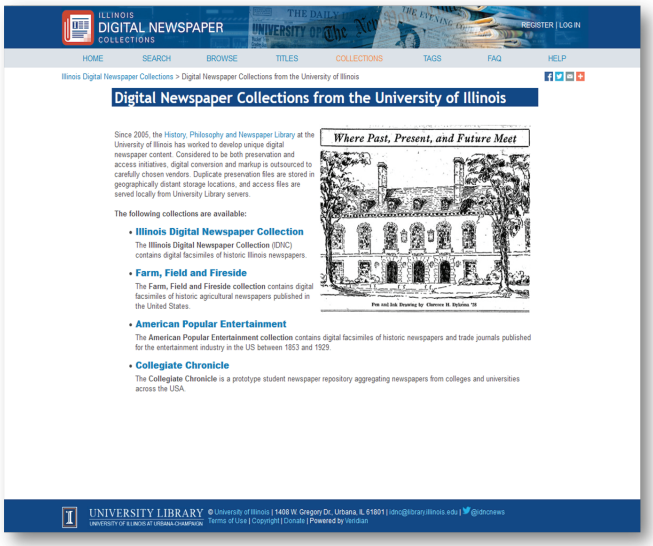
nam, and Estonia, making Veridian one of the few options proven to work well with multi-million page newspaper collections.

- The University of Illinois was able to continue hosting their collection by licensing the Veridian Software. DL Consulting also provides hosting and maintenance services depending on the needs of the library.
- Veridian provides open access to the underlying newspaper data through an XML Interface API.
- Veridian runs on the Apache Solr search platform, a fast, scalable, industry-standard search engine.
- Veridian works on tablets and other touch screen devices out of the box.
- User Text Correction, a feature unique to Veridian which facilitates the correction of errors in the digitized text caused during the OCR process, is not only popular with users it also encourages registration and helps to build a loyal online community around the collection.
- Registered users can add tags and comments to pages, articles, advertisements, or photographs, enriching and improving the collection over time.
- Veridian is highly configurable and can be modified to suit any customization requirements.

EXPERIENCE

Converting 900,000 pages of PR XML data to METS/ALTO is no small task. As with any project of this size there were a few issues that needed to be addressed throughout the project. For example, a fraction of the digitized text experienced a high occurrence of OCR errors, some of the content in the collection was missing, and a number of images were only available in low resolution. The team at DL Consulting worked with the University of Illinois to address each issue. Ultimately they produced an organized and easily accessible archive of 45 historic newspapers in 4 collections: the Illinois Digital Newspaper Collection, Farm Field and Fireside collection of historic farm newspapers and magazines, American Popular Entertainment Vaudeville newspapers, and the Collegiate Chronicle college newspapers.

In the first few weeks since the new Illinois Digital Newspaper Collections has been live 60 users have registered on the site. These 60 people have already corrected close to 22,000 lines of text with the user text correction feature in Veridian. The crowdsourcing features including user text correction, tags, and comments are unique to Veridian Software and the collections that take advantage of crowdsourcing have seen a large increase in the average length of time users spend interacting with the content. Crowdsourcing, combined with Veridian's built-in Search Engine Optimization, is expected to dramatically improve the discoverability of this collection and encourage a high level of user engagement with the content.



idnc.library.illinois.edu

Illinois Digital Newspaper Collections

* Collection of newspapers & trade journals
* 45 titles
* 89,821 issues
* 1,115,613 pages
* 6,648,636 articles
* 60 registers users (to date)

Veridian
digital collection management software

PO Box 12669, Chartwell, Hamilton, 3210 New Zealand
ph: +64 (7) 857-0830 fx: +64 (7) 857-0831 contact@veridiansoftware.com