Automated Creation of Analytic Catalog Records for Born-Digital Journal Articles

Abstract

This article summarizes the approach to bibliographic metadata developed at the University of Michigan Library for journal articles published and archived in HathiTrust using the mPach toolset, which allows journal editors to create born-digital open-access journals and create their own metadata as a byproduct of the publishing process. Specifically, mPach allows a journal editor to convert edited manuscripts from common source formats such as Microsoft Word into JATS (Z39.96-2012) XML and embed structured metadata about the article and journal. Since HathiTrust currently uses MARC as its common-denominator metadata format, JATS metadata are automatically mapped to MARC fields, creating one analytic record per article but without normalizing to follow RDA rules for transcription from primary sources of information or creating entries according to name authorities. For each new journal, a serial record for the journal is created manually by a serials cataloger. This serial record and each analytic record for articles in that journal link to a "collection" for the journal built using the HathiTrust Collections feature.

Keywords

e-journals, analytic records, JATS, digital preservation

The library as journal publisher

The University of Michigan Library has long digitized journals that are in the public domain, making them available through the Web. The U-M Library's homegrown system for doing this, called DLXS,[1] has informed the design of the infrastructure for HathiTrust,[2] a multi-institutional shared preservation-quality digital repository that the U-M Library now uses as the primary home for its reformatted content. But the DLXS infrastructure continues to be actively used by Michigan Publishing, the part of the U-M Library that serves as the main academic publisher of the University of Michigan, to publish born-digital journals because HathiTrust cannot yet accommodate this type of content.

The U-M Library is not alone in serving as a publisher of journals. The Library Publishing Coalition,[3] conducted a survey of library-based publishing efforts in 2013 and found that of more than 110 responding institutions, over 70% publish faculty-led journals and 54% publish student-led journals.[4] While libraries often offer publishing support as a service for their users, the housing of these operations in libraries also provides an opportunity to improve on the situation for the preservation of e-journal literature.

The deficit in journal preservation

Until quite recently, publishers produced documents on physical media, and libraries acquired and preserved copies of these documents. But in the era of the Internet, when publishers host content online, the library's role in acquiring and preserving the content is in jeopardy: without special licensing arrangements such as

those often provided by open-access journals, a library has no legal right to make a copy of the content for preservation.

Various business models have evolved to address this situation, especially for journals, which are increasingly available only online. For non-open-access journals, research libraries often negotiate the right to create a digital copy of any content acquired during the period of subscription[5] and make this content available only to their patrons,[6] though few are equipped to provide this kind of restricted access and archiving with integrated browse and search functions. To address the more pressing concern of publishers going out of business without *any* libraries holding a copy of the content, libraries and publishers have collaborated in initiatives like LOCKSS,[7] CLOCKSS,[8] and Portico[9] in order to guarantee that one or more copy of the content will become available if it is no longer available from the publisher. Similarly, the Koninklijke Bibliotheek and Elsevier reached an agreement in 2002 whereby the KB will preserve Elsevier journals under terms similar to those governing journals that use LOCKSS, CLOCKSS, and Portico.[10] Still, there are problems with these models. LOCKSS uses Web crawling,[11] which captures only the appearance of Web pages but not their underlying structure or search functionality. Portico and the KB, on the other hand, rely on publishers to deliver journal articles in valid file formats, and not just the version first published but also any corrected versions of these articles. CLOCKSS uses either Web crawling or publisher-provided content.[12]

One way to ensure that a library always has access to the latest content is for the library to operate the very system used to publish the journal. A survey in 2010 of a cross-section of North American academic libraries found that, of 144 responding institutions, 43 offered "operational publishing services" to their scholars at the institution.[13] Of these 43 institutions, most host publications using open-source

software such as Open Journal Systems (OJS)[14] or DSpace,[15] while about a quarter use Digital Commons,[16] a hosted platform provided by bepress. OJS and Digital Commons are also the dominant publishing platforms according to the Library Publishing Coalition's 2013 survey.[17]

Unfortunately, all of these platforms deliver to users only those files (primarily PDF files) created and uploaded by a journal editor. Since the library is not in a position to control the software and workflows used to create these files, the library can only provide bitwise preservation of the files, severely hampering future migration of the content.

A higher standard for preservation

Since libraries are increasingly involved in journal publishing, HathiTrust is a natural place to archive and provide access to journal literature to ensure its long-term preservation and discoverability. A TRAC-certified repository,[18] HathiTrust already archives and provides access to reformatted library holdings, but the U-M Library, a founding member of HathiTrust, sees an opportunity to use HathiTrust for publishing born-digital journals as well. To develop an infrastructure in support of low-cost university-based publishing that addresses the needs and values of both content creators and librarians, the U-M Library is funding the creation of mPach,[19] an open-source, end-to-end publishing system in which the act of publishing and the act of archiving are unified. In other words, archiving in HathiTrust happens as a byproduct of publication rather than being carried out after the fact. mPach leverages existing components of HathiTrust and available open-source software where appropriate. Note that by policy HathiTrust only closes access to content for legal reasons, not because a

rightsholder wants to restrict access. Therefore, mPach only supports the publishing of open-access journals.

Archiving is not as simple as saving a copy of a file produced by a journal editor, as OJS and institutional repositories generally do. Instead, the content needs to be stored in a format that allows digital preservation. PDF/A, a non-proprietary variant of the PDF family standardized as ISO 19005, is often suggested for such needs, but a PDF/A file is poorly suited for use with screen readers for the visually impaired and for any non-paginated display, and is suboptimal for searching and data mining.

Rather than preserving the paginated appearance of a document, the text of the article needs to be stored in a format that reflects its structure and semantics, with associated media in formats that can be preserved and rendered. mPach has developed a specification for journal articles that uses the Journal Article Tag Suite (JATS), an application of NISO Z39.96-2012,[20] for the text and stores this with high-quality versions of media objects and with a METS record containing structural and preservation metadata.

An overview of mPach

There are three major parts of mPach (see figure 1), each of which includes components in various stages of development at the time of writing:

- **the peer review and editorial system**: what authors and reviewers interact with
- **Prepper**: what prepares the article for ingest into HathiTrust for archiving and publication

- **modified HathiTrust components**: various modifications to existing components of the HathiTrust environment to support born-digital journal articles

[Figure 1: Major parts of mPach]

As a modular system, mPach can be used with any peer review and editorial system that is capable of interacting with Prepper; however, the developers have chosen to provide OJS as the default option. Despite having no support for digital preservation, OJS is already widely used for library-based journal publishing, and mPach's integration with this software will allow for a smooth transition of journals already published using OJS into the HathiTrust repository. Integration with mPach requires that manuscripts that reach the "layout" stage in OJS be sent to Prepper, which prepares the HathiTrust Submission Information Package (SIP).

Prepper provides a user interface for the editor of a journal: a dashboard for administering the journal and putting article manuscripts through a production process—akin to composition and typesetting—that prepares all content according to the preservation standard developed for mPach content in HathiTrust. Prepper invokes Norm, an application that converts manuscripts from Office Open XML ("DOCX") format[21] into XML that conforms to JATS. DOCX is the default option because, like OJS, it is widely used in the editorial process of journals published by libraries. The Prepper interface also guides the staff member through a review of validation errors detected by Norm's conversion, uploading high-resolution figures, supplying "alt text" for figures, previewing the article as rendered using the default stylesheet (based on the Preview XSLT stylesheets[22]), uploading supplementary material,[23] and submitting for ingest into

HathiTrust. Prepper keeps track of articles so that a revised version can be submitted for ingest. Currently the ingest process overwrites any previous version of an item with the same identifier, but eventually HathiTrust will archive past versions and allow users to navigate among them.

mPach requires a number of significant modifications to HathiTrust components and workflows originally designed to support reformatted print materials. The reading interface in HathiTrust, which previously supported only display of digitized page images, renders JATS XML in HTML and allows a user to download a dynamically generated PDF and EPUB, display metadata specific to articles, and link to a special "collection" for the journal in HathiTrust's Collections application[24] that allows for browsing volumes and issues of the journal. The HathiTrust Data API[25] allows for the content of each article to be retrieved for use outside of the native HathiTrust interface.

Journal- and article-level metadata in mPach

Discovery of known items in HathiTrust using metadata like title and author is currently provided by a catalog of MARC records, with one record per bibliographic entity in the repository. Digitized journals are handled with holdings data, like in an online library catalog. HathiTrust partner institutions contribute MARC records for the scanned volumes according to HathiTrust's Bibliographic Metadata Specifications,[26] an extension of MARC 21 minimal-level requirements.

Since the article, not an issue or bound volume, is the primary unit of interest for readers of a journal, mPach is designed with the article as the unit of archiving and publishing, supporting even the publication of journals without volumes or issues. Each article has its own analytic catalog record, tied to a monographic record for the journal as a whole. Monographic records for the serial as a whole are created manually by a

catalog using information provided by the editor of a journal that will be published using mPach, whereas analytic records are created automatically by Prepper using a mapping of JATS XML metadata elements to MARC fields.[27] In the case of a journal previously published in print and digitized by HathiTrust which then transitions to using mPach, or one that succeeds a previous title, the relevant existing monographic record for the serial will be modified to add the holdings available through mPach or to point to the successor title, respectively.

The automatic mapping of content of JATS XML elements to MARC leads to the creation of analytic records whose fields contain data not conforming to a cataloging code like AACR2 or RDA. In particular:

1. Records do not have titles of journal articles transcribed according to the cataloging code (capitalizing only the first word and proper nouns); instead, titles appear in the record as displayed in the article, possibly in headline style.
2. There is no statement of responsibility but simply a list of authors.
3. Names are handled as the mPach user spelled them and divided them into forenames and surnames.

Conclusion

As part of mPach, a platform for publishing born-digital journals in a preservation-quality repository, the University of Michigan Library developed a model for both journal-level and article-level metadata. Journal-level metadata is created manually when a new journal is established in mPach, but article-level metadata is created automatically as part of the process of publishing an article. Specifically, since the mPach workflow produces structured article-level bibliographic metadata in the

JATS XML format, the system can map JATS elements to MARC fields with only a few infidelities to a cataloging code, none of which impedes discovery in the HathiTrust catalog.

Notes

1. DLXS, http://www.dlxs.org/.

2. HathiTrust Digital Library, http://www.hathitrust.org/.

3. Library Publishing Coalition, http://www.librarypublishing.org/.

4. Sarah K. Lippincott, ed., *Library Publishing Directory 2014* (Library Publishing Coalition, 2013): x, accessed June 27, 2014, http://www.librarypublishing.org/sites/librarypublishing.org/files/documents/LPC_LPDirectory2014.pdf.

5. Sadie L. Honey, "Preservation of Electronic Scholarly Publishing: An Analysis of Three Approaches," *portal: Libraries and the Academy* 5, no. 1 (January 2005): 59-75, doi:10.1353/pla.2005.0005.

6. NISO SERU Standing Committee, "SERU: A Shared Electronic Resource Understanding: A Recommended Practice of the National Information Standards Organization" (Baltimore: National Information Standards Organization, 2012), http://www.niso.org/publications/rp/RP-7-2012_SERU.pdf.

7. Lots of Copies Keeps Stuff Safe, http://www.lockss.org/.

8. CLOCKSS, http://www.clockss.org/.

9. Portico, http://www.portico.org/.

10. "National Library of the Netherlands and Elsevier Science Make Digital Preservation History: Permanent Digital Archive Assures Perpetual Accessibility of Scientific Heritage," August 20, 2002, accessed August 29, 2013,

http://www.kb.nl/en/news/news-archive-2002/national-library-of-the-netherlands-and-elsevier-science-make-digital-preservation-history (site discontinued).

11. “How LOCKSS Works,” LOCKSS, accessed June 27, 2014, http://www.lockss.org/about/how-it-works/.

12. “How CLOCKSS Works,” CLOCKSS, accessed June 27, 2014, http://www.clockss.org/clockss/How_CLOCKSS_Works.

13. James L. Mullins, Catherine Murray-Rust, Joyce L. Ogburn, Raym Crow, October Ivens, Allyson Mower, Daureen Nesdill, Mark Newton, Julie Speer, and Charles Watkinson, *Library Publishing Services: Strategies for Success: Final Research Report* (March 2012), accessed June 27, 2014, http://wp.sparc.arl.org/lps/.

14. Open Journal Systems, http://pkp.sfu.ca/ojs/.

15. DSpace, http://www.dspace.org/.

16. Digital Commons, http://digitalcommons.bepress.com/.

17. Lippincott, *Library Publishing Directory*.

18. “Trusted Repository Audit Checklist (TRAC),” Center for Research Libraries, http://www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying-0.

19. mPach, http://www.lib.umich.edu/mpach.

20. Journal Article Tag Suite, http://jats.nlm.nih.gov/.

21. “Office Open XML,” Wikipedia, http://en.wikipedia.org/wiki/Office_Open_XML.

22. NISO Journal Article Tag Set (JATS) Version 1.0: Preview XSLT Stylesheets, https://github.com/NCBITools/JATSPreviewStylesheets.

23. Recommended Practices for Online Supplemental Journal Article Materials: A Recommended Practice of the National Information Standards Organization and the

National Federation of Advanced Information Services (January 2013),

http://www.niso.org/publications/rp/rp-15-2013.

24. “Collections,” HathiTrust Digital Library, http://babel.hathitrust.org/cgi/mb.

25. “HathiTrust Data API,” HathiTrust Digital Library,

http://www.hathitrust.org/data_api.

26. “Bibliographic Metadata Specifications,” HathiTrust Digital Library,

http://www.hathitrust.org/bib_specifications.

27. “JATS-to-MARC Mapping,” JATS Wiki,

http://webservices.itcs.umich.edu/mediawiki/jats/index.php/JATS-to-MARC_mapping.

Contributor Notes

Kevin S. Hawkins

Director of Library Publishing

University of North Texas