What's Past Is . . . Still Messing With Our Workflows

Jacqueline Whyte Appleby

Scholars Portal, Ontario Council of University Libraries, jacqueline.whyteappleby@utoronto.ca

Author ORCID Identifier: <https://orcid.org/0000-0002-4214-9917>

Follow this and additional works at: <https://docs.lib.purdue.edu/charleston>

Part of the [Library and Information Science Commons](#)

An indexed, print copy of the Proceedings is also available for purchase at:

<http://www.thepress.purdue.edu/series/charleston>.

You may also be interested in the new series, Charleston Insights in Library, Archival, and Information Sciences. Find out more at: <http://www.thepress.purdue.edu/series/charleston-insights-library-archival-and-information-sciences>.

Jacqueline Whyte Appleby, "What's Past Is . . . Still Messing With Our Workflows" (2017). *Proceedings of the Charleston Library Conference*.

<http://dx.doi.org/10.5703/1288284316701>

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.

What's Past Is . . . Still Messing With Our Workflows

Jacqueline Whyte Appleby, Scholars Portal, Ontario Council of University Libraries

Abstract

Scholars Portal has been aggregating locally loaded e-books for Ontario universities on an ebrary-backed platform since 2009—eons ago in the world of library technology! Over the last year, Scholars Portal Books has received a rewrite from the ground up, and this time the focus is on building a platform that anticipates the future of e-book access and usage. No surprise that there have been many challenges along the way, and most of them are unique to consortia: How do we handle corrections to old PDFs? What do we do with six ISBNs? More broadly, how do we support local scholarship at scale, and how can we make space for the open educational resources increasingly being integrated into higher education? This session looked at the complex e-book landscape through a consortial lens—from licensing and entitlements management, to wrangling a dozen XML schemas and implementing ever-changing DRM restrictions, toward the ultimate goal of preserving Ontario universities' books content for the long term.

Introduction

This paper considers the future of locally loaded e-books from a consortial perspective. As we redevelop our home-grown e-books platform from the ground up, we've looked back at the history of the project, examining closely the assumptions that informed our early work. We're also looking around, aiming to understand our users on a much broader scale, including not only readers, but staff, publishers, and the general public, in design and functionality considerations. Finally, we're looking ahead, acknowledging that the boundaries of what makes a "book" are ever-changing, and that the models for making and acquiring book-like content are unlikely to stay the same over the long or even medium term.

Context

Ontario's 21 university libraries have been collaborating since the 1960s and jointly funding digital projects since the early 2000s via Scholars Portal, the service arm of the Ontario Council of University Libraries (OCUL). Scholars Portal Journals was certified by the Center for Research Libraries as a Trustworthy Digital Repository in 2013, and Scholars Portal Books will be the next service audited.

Scholars Portal Books

Scholars Portal Books launched in 2010 with the intention of aggregating consortially purchased e-books and presenting them on a single interface, enhancing the user's discovery and reading experience. The platform today hosts over 250,000

commercial titles from most major academic publishers, as well as almost 500,000 public domain titles from the Internet Archive and a growing collection of open access titles. In general, we receive deliveries of files and associated metadata from a publisher or vendor via an FTP, run it through a publisher-specific loader to get it on the platform, and then distribute MARC records to schools, as well as sending database updates to Serials Solutions. In 2015, OCUL granted Scholars Portal two years of special funding to redevelop the platform. At the end of 2017, the project was almost complete.

Early Assumptions

While Scholars Portal Books continues to load newly acquired content weekly, the software it was initially built on is outdated. Just as pressingly, the e-book landscape has shifted quite dramatically in the last eight years, and many of our early assumptions about how e-books would be purchased, packaged, and used have proven to be untrue—or have had a much shorter shelf life than anticipated. As we work to redevelop the platform and to build software in anticipation of what is not yet true, examining the assumptions and reasoning underlying our early workflow, coding, and design decisions has been crucial.

Assumption: Formats and File Packaging

The first assumption we made was that standards for sending e-books would be somewhat standardized. Our experience with the Journals platform showed

a diversity of metadata standards gradually getting more harmonious as time went on. The format of one PDF per article is overwhelmingly the dominant form, and the near-constant presence of an ISSN, volume, and issue make for the easy concatenation of files.

In 2009, as our platform development began, e-books metadata, formats, and data structures were less well defined, but there was the belief that this would settle down over time; that a preferred format would win out, and that loading e-books would soon lend itself to extensive automation. This has not proven to be the case.

In terms of metadata, some publishers send only MARC records, but many send ONIX or rich full-text XML, but the XML is not in a standard format. Sometimes metadata will come at the book level, and other times there will be chapter-level metadata as well. One sends us beautifully specific TEI—but since they're the only ones doing it, it's just a whole new separate workflow. It is not uncommon to get multiple formats from the same publisher, especially if the metadata has been produced over a long period of time and by many different people. And even if the metadata is standard, the content may not be packaged in a singular format, making it challenging to write a loader. The inherent variety in the structure of books is an obvious hurdle; not all publishers send individually packaged chapters, but many do, and there may or may not be sections or parts to a book on top of that, as well as separately designated front matter, appendices, and supplements.

And how will these pieces come together? Will the chapters each have their own subfolder within the book's folder? Sometimes we'll get a single XML file containing all metadata for the entire delivery, other times there may be a separate file in each chapter-level subfolder. This variety of structures means that the time required for analyzing each new publisher can be extensive, and there is often the need to do this analysis for publishers we've worked with long term; because there is no standard, if a new person is hired to package the files, it is often the case that he or she will put them together in a slightly different configuration.

Assumption: For Every Book, a MARC

The second assumption we made was around MARC records, which were still the dominant book metadata record type in 2009. While we knew they had

a shelf life, we anticipated schools requiring them for the foreseeable future, and so tied ourselves to them—the software of the old platform required a MARC record to load a book.

While all of our schools do still rely on MARCs for at least some aspects of discovery, our ability to get those MARCs has proven to be frustrating, and often futile. The use of third parties in MARC creation has meant there are often delays and miscommunication. Because we are not a school that purchases content, we have consistently had trouble convincing vendors to allow us to download their records, and we have often resorted to asking someone at a purchasing school to download and share them with us, so we can serve the content back to them.

Without a MARC, we had no way to load and display the content, which meant that trouble with a missing record—or, just as often, hundreds of records—caused huge delays in the availability of purchased content.

Assumption: Everyone Buys Everything

Our final assumption was around the long-term sustainability of Big Deals. As the platform was being initially developed, purchasing huge packages of content across subject areas was still the norm. Our original entitlements management system does not account for nested or subpackages, for titles that have been sold in multiple configurations, or for backlists that are being sold in more granular fashion over time.

There are also now more third parties licensing content, which means that while some of our schools may purchase content directly through the publisher, other schools may have purchased content through a major vendor. In addition to the challenge of managing title lists when they have most but not full overlap, schools may have different license permissions for the same book—some will have purchased only single-user access, while others have no such restrictions.

Lastly, mergers and acquisitions between publishers and vendors sometimes means that while legacy backlists are still packaged and sold separately, the metadata associated with these items no longer distinguishes them as separate from their parent company, making it challenging for our loaders to understand which collection they should be associated with in the entitlements app.

In sum, a lack of standardization of e-book formats, a reliance on MARC, the ever-increasing presence of third parties in the creation, sales, and distribution of content and metadata, and an increasing diversity of licensing options and grants, meant that our legacy model for loading and providing access to e-books for Ontario universities was no longer sufficient.

Redevelopment for the User

As we redevelop the platform, we have the opportunity to examine past assumptions and to think carefully about the more open, flexible system we would like to build for our users. While a focus on an enhanced user experience has always been the goal of the platform, we're now thinking more broadly about who our users might be.

Standardization: Our Platform as a User

We have accepted that we may never get standardized books metadata and have opted to standardize on ingest—every book loaded onto our platform first has its metadata, whatever form it arrives in, transformed to BITS, the Books Interchange Tag Suite. BITS is the sister suite to JATS, which is the format that all of our journals metadata is transformed to. BITS was chosen as our singular format after extensive research and years of successful work with JATS. It is highly flexible and extendible, allowing for extensive about-the-object and about-the-narrative metadata at the book, part, and chapter level. For our purposes, the fact that it is in harmony with JATS is ideal for data sharing between books and journals repositories as the two formats become less discrete. Use of JATS has grown, and ideally, BITS will catch on, but we are prepared to continue to develop and tweak crosswalks for all incoming content.

The enhancement that this standardization provides cannot be overstated. With so many inconsistencies in how MARC subfields are applied, it was extremely challenging to differentiate different types of ISBNs, to properly recognize different contributor types, and to parse any chapter-level information. Our BITS records allow for much richer and more specific ways to sort, filter, and display the data.

E-Books for All: Users With Disabilities

Scholars Portal has run the Accessible Content ePortal (ACE) since 2013. It is a view of SP Books that is limited to students who have a token from their

local accessibility office. Books requested by disabled students are scanned centrally at the University of Toronto, then opened to ACE students at all schools that have a local copy of the book. Over time the repository has grown to hold over 5,000 scholarly books, with quite large collections in areas where graduate students have made extensive requests. But ACE has always existed separately from the regular Books platform content; with the redevelopment, we are bringing the entirety of their licensed books collection to ACE students. As much of the content is born-digital PDFs, it is at least as accessible, if not more accessible, than the scanned content they have previously used.

ACE students will also see an option to request an alternate format for any title, should the standard PDF be unworkable for their needs. Using the Internet Archive's derivatives tool, we can generate an alternate format and get it to them, usually within 48 hours. I wish to stress that the ability to transform content for a user with a disability is integral to OCU's equity policies, and schools do not sign licenses that do not allow for this transformation to take place.

"What's in the Deal?": Staff Users

Our users are also all of the staff at OCUL libraries who may or may not have been involved in the signing of particular deals, but who have questions about the content, including wanting to confirm what was ultimately included in a deal, how often titles are being accessed by their users, and whether the collection is properly represented in their discovery layer. These questions have generally been answered by an e-mail back-and-forth or via an assortment of confusing and difficult to maintain wiki pages. With the redevelopment, an admin tool that can query the Books database and return such answers will be open to OCUL staff, allowing them read-only access to the entire database underlying the system. They will be able to generate KBART files by collection or load date, packages of MARCs, along with a log noting all books in the collection missing MARCs, and COUNTER-compliant usage statistics, all in real time, in whatever configuration of titles and timelines makes the most sense for their analysis.

Likewise, publishers will have access to COUNTER-compliant statistics on their own collections, allowing them to pull usage data in aggregate or by school on a schedule that fits their operations.

Digital Preservation: The Users of the Future

The more we talk about an open library (and we do, though I haven't much here), the better we understand that our users are forever. Integral to the redevelopment planning has been the development of a workflow to preserve e-books content for the long term using Archivematica. Scholars Portal and a number of OCUL schools are already using Archivematica for the creation of Archival Information Packages in other contexts, and we feel confident that it is the appropriate tool for packaging book files, metadata, and versions through time.

In order to preserve content long term, Scholars Portal needs the right to locally load the content, the right to host the content in perpetuity, and the right to transform that content over time to formats of the future. Most OCUL schools are not members of Portico, and Scholars Portal is their central preservation strategy—they do not sign licenses that do not allow for these rights.

Unsolved Challenges

There remain a number of so far unsolved challenges for our loading and display workflows. The first is the handling of corrections to books. On the Journals platform, corrections are relatively straightforward—if an article is redacted, a new blank “copy” of the article is sent to us and the PDF file is overwritten on the platform, though the original file remains preserved in our file system. Corrections are generally published in future articles, and a link is made between them via DOI. Opening a journal correction file will quickly help us, or anyone else who stumbles across it, understand the nature of the correction.

Books corrections are more complicated. Sometimes a publisher will send us a new version of a book because they have corrected some typos and other small mistakes. While they may resend us the entire book, they may only send us affected chapters. There is never a manifest included with corrections, stating how serious the correction is (is the thesis of the work incorrect, or is the name of a participant misspelled on page 317?). Sometimes corrections are sent in a separate folder labeled “Corrections” or “Updates,” but they are also often interspersed with the general book delivery, leaving us unclear whether this is an updated version or simply an accidental resend. The latter case does happen fairly

often, and by default our loader discards an ISBN it has loaded before.

An additional challenge is improperly marked up PDFs. Where chapters are not appropriately tagged in the text, a table of contents may be difficult to navigate. For users with disabilities, poor tagging of PDFs can leave them unclear about where they are in a document and unable to access the specific material they are looking for. Publishers producing born-digital PDFs: It is unacceptable to not, at minimum, mark up your PDFs at the chapter level, with clear headings. Please find some online tutorials for doing this, and make sure it gets done. In the very near future, I suspect this requirement will be a clause in our licenses.

Looking Ahead

While many of these changes will help improve the user experience, all are based on assumptions about the type of content we will work with—generally PDFs, with a selection of XML-based documents. We hope to see more ePubs in our future, but they remain rare for now.

In practice, we already receive a broader variety of content types, and we anticipate this growing as digital scholarship continues to transform scholarly communications processes and outcomes. The bigger question, then, is not how we will support the traditional e-monograph long term, but how we will do this while also expanding the types of content and the variety of user experiences and interactions on the platform. In particular, we see the following three resource types as likely to have a larger presence on the platform in the year ahead:

Locally Published One-Offs and Open Access Content

As a preservation tool and an access point, Scholars Portal Books is already the platform of choice for some Ontario faculty who have self-published a book, or put together conference proceedings or a similar themed collections.

We also receive requests from schools to load either specific titles or large collections of open access content they believe their users may wish to access, or that they do not feel has a secure preservation plan behind it, and no easy route for discovery. While many of these items may come in PDF format, for the one-offs there is often no structured metadata

associated with the file, and there is certainly no workflow in place for accommodating these requests, especially when we're being asked to contact a third-party publisher to request files.

As the preservation piece of the platform is fully committed next spring, and with the debut of our new and more usable front-end, we anticipate an uptake in requests of this sort. Since we've historically dealt with purchased collections from major vendors, an additional new consideration will be for the quality of the material. Should all requests be loaded? Students see our platform as made up of content curated by their own libraries. What if we're loading, well, junk?

Additionally, because purchased content comes with a license, we have authoritative information about how entitlements for the collection should be managed. Where content is brought to us as "open," how do we confirm that we have the rights we're being told we do? Some processes for affirming the rights associated with a book, as well as a place to document those rights, will be needed.

Open Educational Resources

The development of a provincial Open Educational Resources (OER) strategy is a topic of great interest to OCUL schools, who are looking to help students save money, as well as make more of their school's teaching resources available for redevelopment and reuse. Again, we will want to balance the ease of adding content with the need to manage and display different possible reuse clauses. Other questions to consider include: In a situation in which a book has been remixed and reconfigured, can we keep and display both versions in a way that is useful? Can we integrate tools for remixing, commenting, copyediting, or commenting right into these resources? How do we make sure we properly acknowledge the lineage of content as it moves from use, to reuse, to new formats? These are metadata considerations, but also policy considerations, and we will need the support of many within OCUL as we move into this territory.

Web Archiving Derivatives

Web archiving is increasingly seen as an important tool for preserving online content that may

once have been collected in print, in particular for government publications and related gray literature, as well as Web content related to rapidly unfolding current events. Capturing content has proved fairly straightforward with the availability of tools such as Archive-It. The challenge is in creating useful metadata to help make these captures findable and understandable. In the case of harvested PDFs, they do not generally tend to be well marked-up, and so Web archiving tools are not capable of capturing any meaningful metadata about them. In this case, every document captured requires that a person manually create an associated set of metadata. In the case of websites, some tags may help point to key information, but most lack any semantic associations; how can we distinguish between the author of a news article and its criminal subject, at scale? More broadly, what should we be archiving? We know there's interest from many schools in building local collections of archived files or their derivatives. Is Scholars Portal Books an appropriate home for storing and accessing these files over the long term? We have recently become involved with the Canadian Web Archiving Coalition, and we hope to engage many national partners and stakeholders in these kinds of discussions.

Conclusions

What stands out for me when I look at the work we're currently focussed on, and the new directions we see ourselves taking in the coming years, is a focus on stewardship. The resources in our custody must be available for the long term, and they must be available to users in a format that meets their needs. We are not limiting this sense of stewardship to content with a license and price tag attached, and increasingly open, local, and government resources are seen as important collections for us to support.

The stewardship of e-books through future years will require more than just a fresh platform. My focus here has been largely on the technology used to enable access to an ever-broadening array of book-like content, but there are many other lenses through which to view this work, and it's not always easy to tease them apart. The future of e-book stewardship is a policy, budget, and licensing question, it is a staff development question, and in many cases it is a moral and ethical question. Thank you for giving me the space to ask some of these questions today.