OR2019 Presentation Submission
14th International Open Repositories Conference, June 10th-13th, Hamburg, Germany

Uncomplicating the business of repositories

Emily G. Morton-Owens, University of Pennsylvania Libraries, egmowens@upenn.edu

Katherine Lynch, University of Pennsylvania Libraries, katherly@upenn.edu

Session Type

- Presentation

Abstract

In this presentation, we discuss how our library runs our repository in production to meet the needs of our "business" as efficiently as possible. We have an interest in limiting the number of digital platforms we manage, for the purposes of sustainability and efficiency, but we must also consider how well a general platform can meet specific user needs.

A governance group of administrators, in conference with stakeholders and developers, seeks to find the best way to accommodate each collection or functional need, with an eye to minimizing technical complexity, offering stakeholders self-serve options when possible, and maintaining a single canonical copy of each object. We will present some case studies of how material has been handled in our developing digital ecosystem, where preservation and access sometimes present conflicting priorities. We are exploring how our repository can best evolve to support our aims of making data and documents freely available.

Conference Themes

- Repositories - evolution or revolution?
- Supporting open scholarship and cultural heritage
- Open and sustainable

Keywords

Repository scope, repository planning, IT governance

Audience

Administrators, repository managers, and librarians who are also working with growing collections, multiple platforms, and platforms under development

Background

Starting in 2016, the Penn Libraries have been working on a highly generalized digital repository based on Fedora and Samvera, with significant customizations to achieve our aims of avoiding software lock-in and offering cost-effective storage at a large scale. At OR 2018, we presented about these technical aims and how we achieved them.[1] Now that our repository, Colenda, is in production, we have moved on to a new phase of the project that requires a greater focus on business requirements analysis to ensure that we are using our new tool to the greatest effect.

Presentation content

Following the launch of our Samvera-based repository, Colenda, the Penn Libraries team has been working with administrators, librarians, and other stakeholders to allocate our effort to support new initiatives and formats of material. These discussions raise questions about scope that must be answered by strategy because there are few technical limitations in play. While some institutions have multiple platforms that they use for repository-esque purposes, we have a single new platform that can theoretically support many uses, but limited developer effort that forces us to analyze and prioritize what comes first. We also have legacy platforms with pros and cons, one of the more interesting pros being that our users still rely on many of these platforms on a regular basis, with the unusual ramification that a faithful user base to legacy software and hardware means preserving functionality and access to content without the disruption that typical sunsetting processes entail. This proposal considers the factors of how we organizationally meet user needs.

Our initial work on the Colenda platform has supported the ingestion of approximately 30 TB (7,776 objects) as of January 2019, most of which consists of traditional, book/manuscript/single-sheet type images. Much of the developer effort went into crafting a highly generalized approach to metadata management and transformation as well as the creation of workflows for staff to process and ingest objects of any format or structure. With those workflows underway, the next steps raise questions about the scope of the tool, because it was designed in a general way to accept any type of data object as long as it was accompanied by specified minimal metadata, which we call PQC or Penn Qualified Core.

For example, the most recent development work on Colenda has been to add an audio/video player in support of material that we are digitizing under a CLIR grant to provide better access to our Marian Anderson collection. It was easy to decide, in this case, because the feature had obvious ongoing usefulness and made it possible for us to apply for a grant we were interest in pursuing.

We are currently also engaged in a project to re-evaluate our use of bepress's Digital Commons product after it was acquired by Elsevier. From a technical standpoint, there is no question that Colenda could support both the content and the workflows around what we have branded as "Scholarly Commons." (We do not promote self-ingest for Scholarly Commons; library staff ingest the material for our faculty. This obviates what would be one of the biggest obstacles to using Colenda for institutional repository purposes: creating an end-user-friendly workflow. Instead, we would train an additional cohort of staff to use the staff-oriented ingest process.) Still, the value of mixing this content into a discovery environment with a diverse collection of other items is unclear, and it may require its own discovery portal to be welcoming to users. In this case, it may be better to keep the material housed separately. If the content requires a discovery context completely apart from its presence in Colenda for its primary use and audience, it is likely better suited for a different platform.

Another example is related to our Materials Library. Housed in the Fine Arts Library, the Materials Library provides samples of traditional and emergent art and construction materials. Library staff have created photo and video surrogates of the materials for purposes of both preservation and access, because, as always, we want to make our collections available to non-local users wherever copyright permits. In this case, there remain important questions of user needs and behavior that we must take on to understand the value of this effort compared to others—how might users search and use their results? What value does a non-tactile version of this data offer, and how can users access that value?

In each case, Colenda is capable of providing data storage, preservation, and access at a minimal level that achieves some of the library's aims in terms of data management, but to support customized, contextual access and use that is specific to the unique content requires additional work (business requirements analysis and coding). Further, each of these collections has previously been, or in the future could be, supported to some extent by another platform.

The presentation will briefly discuss questions of scope and governance that have gone into the project once we moved beyond the basic features of our first phase of development. We have a governance group of administrators that attempts to limit the proliferation of digital platforms while ensuring that creative and strategic collections have a home. This group works with the developers and librarians about what is needed and what is possible.

In evaluating the services and content that we support, we ask ourselves the question, when is it acceptable to duplicate derivative content in two separate platforms, in the interest of simplifying and appropriately compartmentalizing our services? For instance, because Colenda is the canonical home of most digital objects featured in our online exhibits, our natural first inclination was to find a way to stream content from Colenda directly into another interface that would present as a completely different application. However, we found this would place an undue burden on the repository, as it would necessitate us refactoring code around a core component (the Solr index) of our Samvera instance in order to support functionality that has nothing to do with long-term preservation of these same assets. As long as there is one canonical representation of an object (and everyone stewarding the content understands which copy this is), derivatives used for other contexts can be copies. It reduces the complexity of our services and makes user interaction with the data simpler without taking up too much storage. In this case, we ultimately decided to decouple Colenda and exhibits, instead using Omeka to create online exhibits. While this meant introducing another platform into our infrastructure, Omeka was vetted as sufficiently straightforward to maintain that the complexity it introduced was minor enough to outweigh the abstract complexity of the alternative. Rather than imposing our own ideas of elegance on users, we ultimately opted to offer them a solution that would be largely self-serve and allow them to create representations of collections or exhibits without relying on IT staff.

This choice also met local user needs in a better way, simplifying their work for creating links between exhibits and the repository. Curators can simply look up an object in Colenda, download a high-resolution derivative image from the discovery layer, and upload that to their Omeka site. Omeka also allows curators to link back to alternative representations of the content, so users discovering the content in Omeka can find their way back to the full, long-term object in Colenda. Finally, Omeka as part of a LAMP stack represents a clean, robust exit strategy for migrating exhibit content into another platform down the road, with the considerations of one and not two software platforms to consider when the time comes.

By comparison, once we had a repository as the canonical source of truth for our objects, this presented an opportunity to integrate it with an existing system: OPenn (a website providing predictable and programmatic access by front-end users to open-access images and metadata). All content in OPenn was regarded as needing the long-term preservation benefits of Colenda, so all content in OPenn was intended to be ingested into Colenda. Fully duplicating OPenn's content in Colenda would mean creating duplicate, full-resolution representations of these objects in two isolated platforms, which would generate ambiguity around which version was the canonical one as well as make interoperability between the two platforms very complex if even achievable. As such, it was ultimately decided that OPenn should be reimplemented to integrate directly with Colenda, serving its master images and metadata from Colenda's storage pool, with all updates to objects taking place in Colenda and trickling down to OPenn. This represents some added complexity to implement; however it reduces the ongoing complexity of maintaining two siloed applications that draw from the same content and, from the user perspective, serve similar purposes.

This proposal is partly inspired by another institution's statement, at OR 2018, that they had around twenty different repository instances. In order to keep our infrastructure maintainable by our relatively small staff, we are actively attempting to limit the number of platforms or instances, while at the same time using the right tools for each task. In our work to uncomplicate the necessarily complex landscape of supporting varying content where one size does not fit all, we are always mindful of what factors count when making these decisions: service purposes, introducing new platforms, long-term exit strategies, scale, and

sustainability. This reduces pressure on our development staff by simplifying our ecosystem but increases the need to be organized and aligned at a policy and governance level.

Repository System

- Digital Commons
- Fedora
- Samvera

Conclusion

In trying to build a repository that avoided software lock-in or constraints on the types of objects we could support, we pursued the promise of a constantly evolving collection that could include any assets that enhance research, teaching, and learning at Penn. Now that we have tools to meet general user needs for sustainable preservation, we are looking ahead to new object formats and software integration use cases. We have a created a forum where we develop policies that ensure that the use of our repository matches the Libraries' strategic vision to make data open and support innovation in the digital humanities.

References

[1] Lynch, Katherine. Morton-Owens, E. G. Defensive Design: developing a system-agnostic repository for sustainable long-term preservation. Open Repositories 2018: Bozeman, MT. June 2018.