MARAC SPRING 2023

S10: Challenges in Selecting a Digital Repository System

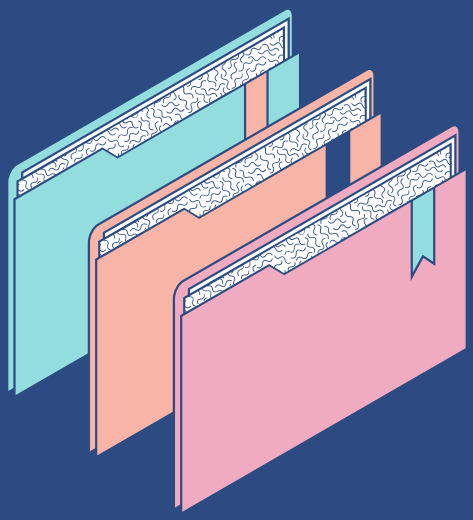
Thank you to our sponsor

MARAC wishes to thank the sponsor of this session, Lucidea. Lucidea offers knowledge management, integrated library systems, and collections management software solutions that help you and your organization redefine how your knowledge is shared.

We will be starting with Christine Anne George, who is the Associate Director for Research & Scholarship at the New York University Law Library. Prior to that, she was the Assistant Director for Faculty & Scholarly Services at the Cardozo Law Library. During her time there, she launched and oversaw Cardozo Law's IR, LARC, which contained both scholarly and archival collections. For the past six years, she has severed in leadership positions in the American Association of Law Libraries' Law Repositories Caucus and organized programming to keep law librarians up to date on the latest information concerning legal repositories.

My name is Alex Japha and I am the Digital Archives and Special Collections Librarian at Lehigh University in Bethlehem, Pennsylvania.

Lehigh is a medium sized research university. We are a merged organization between Libraries and Technology Services. The Special Collections department I work in is only three people, and I am responsible for digitization and our online presence. This includes adding content to our Digital Collections Repository, our Institutional repository, and migrating systems when necessary. So what will this presentation cover.



Overview

KEY TOPICS DISCUSSED IN THIS PRESENTATION

- Lehigh Repository History
- Decision to Migrate
- Selection Criteria
- Repository Options and Selection
- Migration and Launch
- Continuous Migration

First, I will give an overview of Lehigh's historical repository platforms, as that history has shaped and guided our current systems. Next, why we decided to migrate from those past platforms. How we decided on a future platform and chose our vendors. That will then bring us to the migration of content and the launch of our current system before finishing up with our future plans.

Digital repositories at Lehigh are over twenty years old, which is pretty early in the field. Given the limited options back in the early aughts, Lehigh implemented Greenstone as a kind of electronic bookshelf. This required some custom interfaces and websites to make the digital books available.

Pretty soon after that, Lehigh implemented an early version of CONTENTdm. Lehigh actually served as a pilot for several CDM features such as article segmented newspapers, EAD finding aids, and institutional repository. You can see Lehigh stuck with CDM for about 18 years because we were content with an unlimited license for local hosting, and had enough local storage space for our collections.

In 2012, Lehigh added a the Bepress digital commons institutional repository as a platform primarily for electronic theses and finding aids, which were being made available through Proquests platform. This worked well and we expanded the IR with journals, videos, and conference proceedings. However, the cloud hosted and subscription nature was problematic.

With Islandora in 2020, we were able to merge our digital collecitons and IR from CDM and bepress. We have two different front end interfaces but a singlw back end.

And finally, from 2023 forward we plan on using Islandora 2.0 for a single platform. Hopefully the merging of everything will help streamline our efforts.

Decision to Migrate

WHY LEAVE A FAMILIAR PLATFORM?

Bepress acquired by Elsevier in 2017

- Annual price increases
- Minimal additions beside ETDs
- UPENN Beprexit

Locally hosted CONTENTdm reaching end of life

- Server OS out-of-date
- No longer secure to IT
- Storage in new cloud hosted CDM cost thousands per TB

New University Librarian

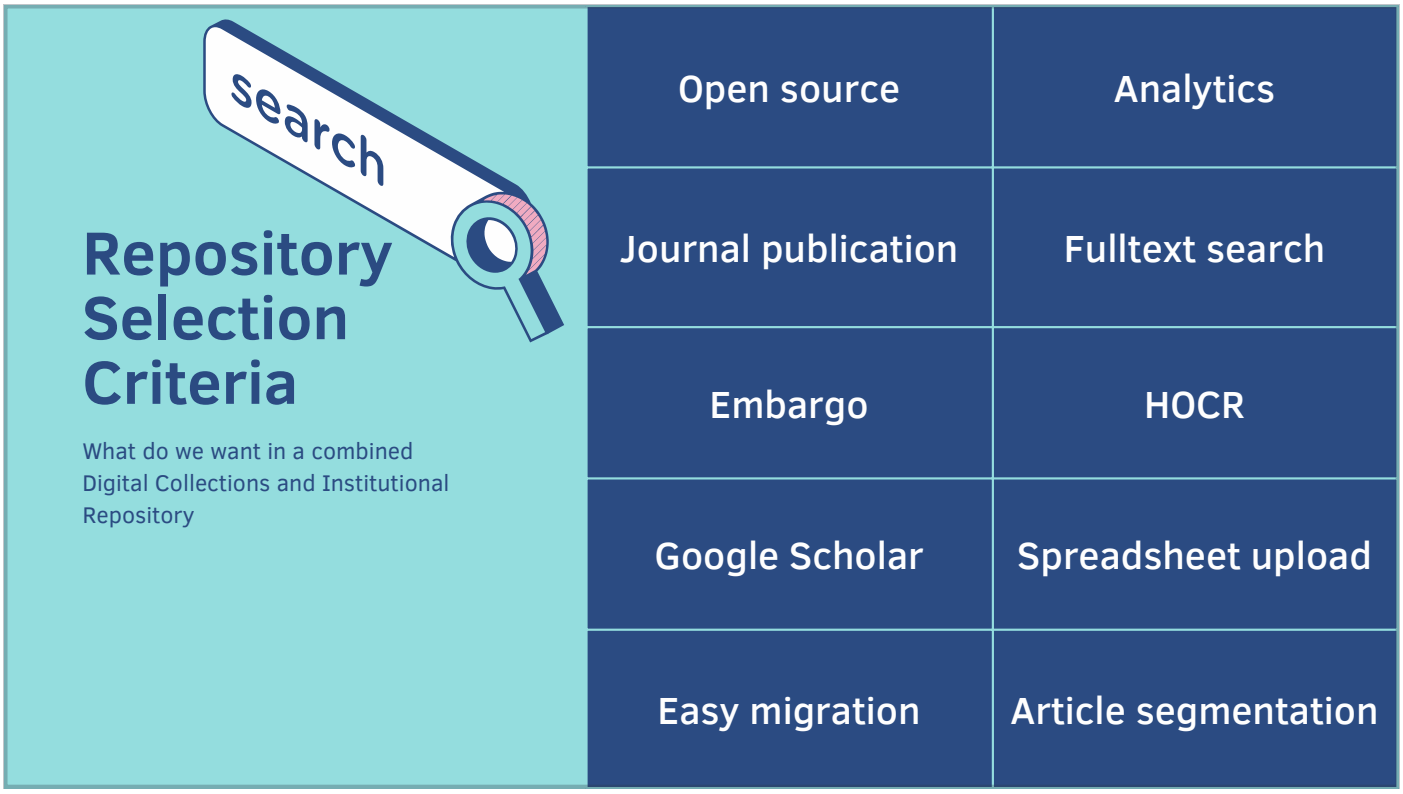
- Focus on open source solutions
- Dislike of annual payments

So why did we decide to change platforms if we were relatively happy with our existing technology.

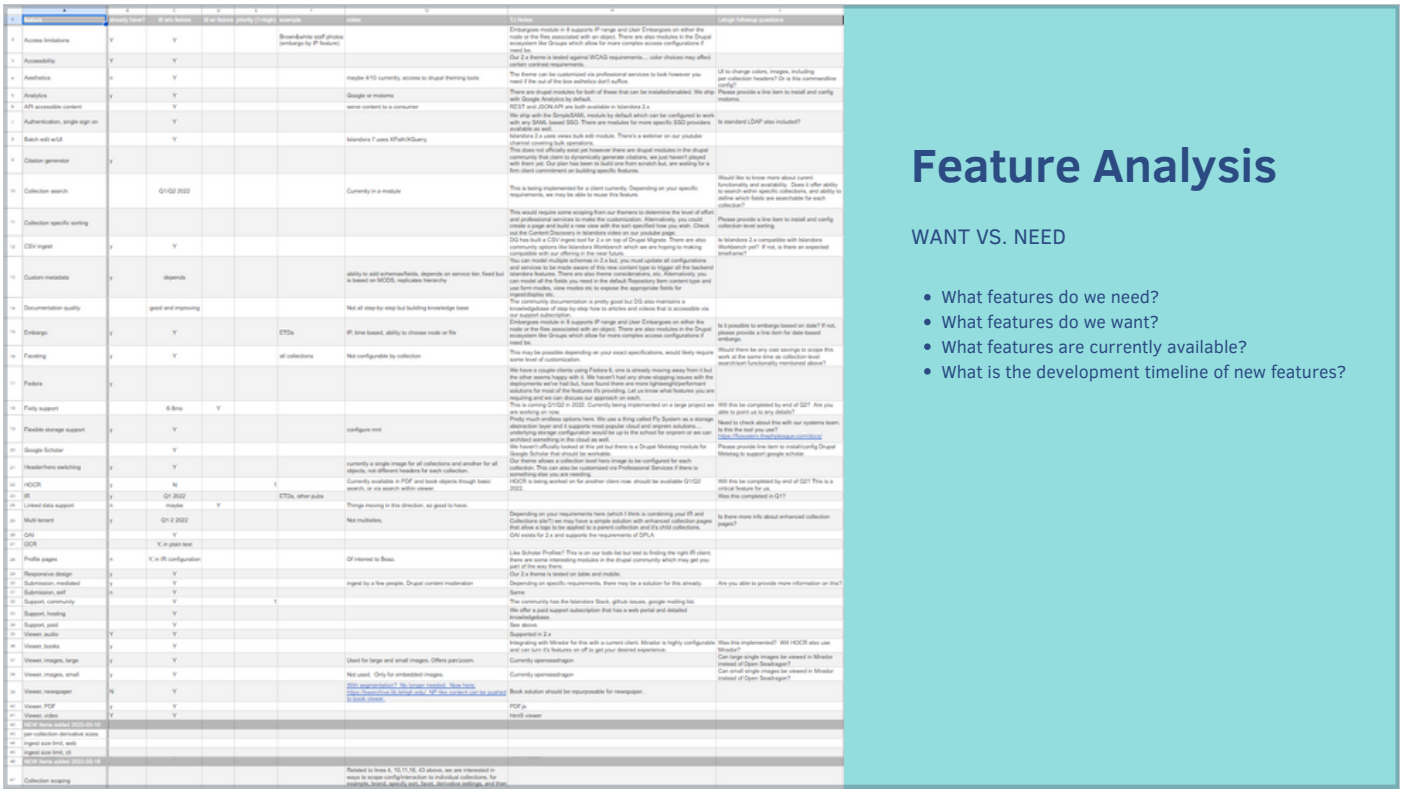
The first issue arose when Elsevier acquired Bepress in 2017. Our library has had less than stellar experiences with Elsevier in other areas like journal acquisitions, and the idea of a predatory publisher taking over an open access platform set off alarm bells. We also weren't adding a ton of material to the IR other than our ETDs. We had a journal being published at Lehigh, but the responsible faculty left. The annual price increases were becoming onerous for such minimal new content. It also felt like there was an organic groundswell around leaving Bepress after Elsevier purchased it. I'm not sure if other remember the University of Pennsylvania's Beprexit announcement, but that helped made a much smaller insitution like Lehigh realize we wouldn't be alone in switching platforms. However, Penn is still using Bepress to this day while we have since left.

CONTENTdm was an easier decision. OCLC had stopped supporting the locally hosted version of their software in favor of a new cloud hosted option. Due to the storage cost of the cloud storage, which I was told was $6,000 per terabyte, we ruled it out as an option. Without vendor support, our IT staff realized CDM would no longer be secure and we needed to migrate.

We also got a new University Librarian in 2019, who wanted to put a focus on open source software, like the FOLIO ILS that Lehigh helps develop. He also has a strong dislike of annual payments and subscriptions, which limited our options.



So how did we actually decide on our current repository system. Between the Special Collections and library technology departments along with the Institutional Repository committee, we came up with a ton of features and criteria match up with software. The ten criteria on the right aren't all of them but they are what we decided are the most important. Some of those criteria were met by other software options. Article segmented newspaper was a feature being offered by only one platform, Veridian, so we separated out our student newspaper there. Once we lost the active journal, we decided that Open Journal System would be a sufficient replacement for Bepress

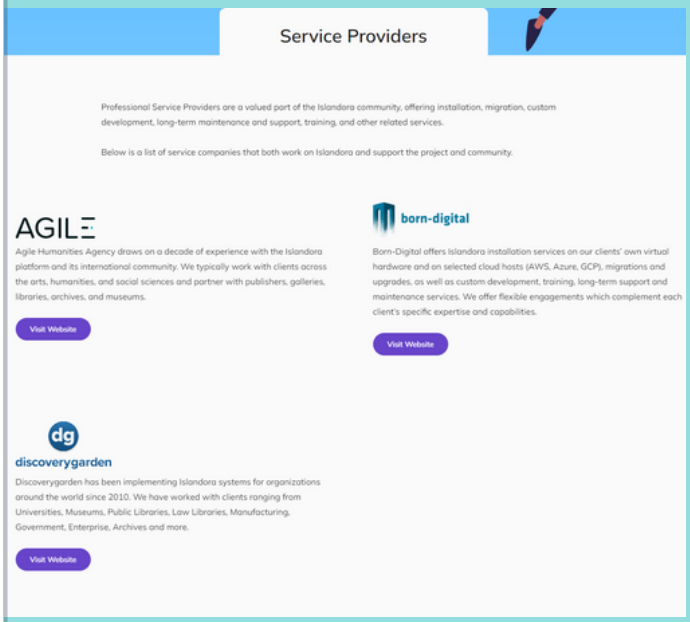


	feature	already have?	B w/o factors	B w/ factors	priority (1=high)	example	notes	1 / Notes	Lehigh followup questions
2	Access limitations	Y	Y			Brown&white staff photos (embargo by IP feature)		Embargoes module in 8 supports IP range and User Embargoes on either the node or the files associated with an object. There are also modules in the Drupal ecosystem like Groups which allow for more complex access configurations if need be.	
3	Accessibility	Y	Y					Our 2.x theme is tested against WCAG requirements... color choices may affect certain contrast requirements.	
4	Aesthetics	n	Y				maybe 4/10 currently, access to drupal theming tools	The theme can be customized via professional services to look however you need if the out of the box aesthetics don't suffice.	UI to change colors, images, including per-collection headers? Or is this commandline config?
5	Analytics	y	Y				Google or matomo	There are drupal modules for both of these that can be installed/enabled. We ship with Google Analytics by default.	Please provide a line item to install and config matomo.
6	API: accessible content		Y				serve content to a consumer	REST and JSON API are both available in Islandora 2.x	
7	Authentication, single sign on		Y					We ship with the Simple SAML module by default which can be configured to work with any SAML based SSO. There are modules for more specific SSO providers available as well.	Is standard LDAP also included?
8	Batch edit w/UI		Y				Islandora 7 uses XPath/XQuery,	Islandora 2.x uses views bulk edit module. There's a webinar on our youtube channel covering bulk operations.	
9	Citation generator	y						This does not officially exist yet however there are drupal modules in the drupal community that claim to dynamically generate citations, we just haven't played with them yet. Our plan has been to build one from scratch but, are waiting for a firm client commitment on building specific features.	
10	Collection search		Q1/Q2 2022				Currently in a module	This is being implemented for a client currently. Depending on your specific requirements, we may be able to reuse this feature.	Would like to know more about current functionality and availability. Does it offer ability to search within specific collections, and ability to define which fields are searchable for each collection?
11	Collection specific sorting							This would require some scoping from our themers to determine the level of effort and professional services to make the customization. Alternatively, you could create a page and build a new view with the sort specified how you wish. Check out the Content Discovery in Islandora video on our youtube page.	Please provide a line item to install and config collection level sorting.
12	CSV ingest	y	Y					DG has built a CSV ingest tool for 2.x on top of Drupal Migrate. There are also community options like Islandora Workbench which we are hoping to making compatible with our offering in the near future.	Is Islandora 2.x compatible with Islandora Workbench yet? If not, is there an expected timeframe?
13	Custom metadata	y	depends				ability to add schemas/fields, depends on service tier, fixed but is based on MODS, replicates hierarchy	You can model multiple schemas in 2.x but, you must update all configurations and services to be made aware of this new content type to trigger all the backend islandora features. There are also theme considerations, etc. Alternatively, you can model all the fields you need in the default Repository Item content type and use form modes, view modes etc to expose the appropriate fields for ingest/display etc.	
14	Documentation quality		good and improving				Not all step-by-step but building knowledge base	The community documentation is pretty good but DG also maintains a knowledgebase of step-by-step how to articles and videos that is accessible via our support subscription.	
15	Embargo	y	Y			ETDs	IP, time-based, ability to choose node or file	Embargoes module in 8 supports IP range and User Embargoes on either the node or the files associated with an object. There are also modules in the Drupal ecosystem like Groups which allow for more complex access configurations if need be.	Is it possible to embargo based on date? If not, please provide a line item for date-based embargo.
16	Faceting	y	Y			all collections	Not configurable by collection	This may be possible depending on your exact specifications, would likely require some level of customization.	Would there be any cost savings to scope this work at the same time as collection-level search/sort functionality mentioned above?
17	Fedora	y						We have a couple clients using Fedora 6, one is already moving away from it but the other seems happy with it. We haven't had any show-stopping issues with the deployments we've had but, have found there are more lightweight/performant solutions for most of the features it's providing. Let us know what features you are requiring and we can discuss our approach on each.	
18	Fedy support		6-9mo	Y				This is coming Q1/Q2 in 2022. Currently being implemented on a large project we are working on now.	Will this be completed by end of Q2? Are you able to point us to any details?
19	Flexible storage support	y	Y				configure mnt	Pretty much endless options here. We use a thing called Fly System as a storage abstraction layer and it supports most popular cloud and on-prem solutions... underlying storage configuration would be up to the school for on-prem or we can architect something in the cloud as well.	Need to check about this with our systems team. Is this the tool you use? https://flysystem.thephpleague.com/docs/
20	Google Scholar		Y					We haven't officially looked at this yet but there is a Drupal Metatag module for Google Scholar that should be workable.	Please provide line item to install/config Drupal Metatag to support google scholar.
21	Header/hero switching	y	Y				currently a single image for all collections and another for all objects, not different headers for each collection.	Our theme allows a collection level hero image to be configured for each collection. This can also be customized via Professional Services if there is something else you are needing.	
22	HOCR	y	N		1		Currently available in PDF and book objects though basic search, or via search within viewer.	HOCR is being worked on for another client now, should be available Q1/Q2 2022.	Will this be completed by end of Q2? This is a critical feature for us.
23	IR	y	Q1 2022			ETDs, other pubs			Was this completed in Q1?
24	Linked data support	n	maybe	Y			Things moving in this direction, so good to have.		
25	Multi-tenant	y	Q1-2 2022				Not multisites,	Depending on your requirements here (which I think is combining your IR and Collections site?) we may have a simple solution with enhanced collection pages that allow a topic to be applied to a parent collection and it's child collections.	Is there more info about enhanced collection pages?
26	OAI		Y					OAI exists for 2.x and supports the requirements of DPLA	
27	OCR		Y, in plain text						
28	Profile pages	n	Y, in IR configuration				Of interest to Bosu.	Like Scholar Profiles? This is on our todo list but tied to finding the right IR client, there are some interesting modules in the drupal community which may get you part of the way there.	
29	Responsive design	y	Y					Our 2.x theme is tested on table and mobile.	
30	Submission, mediated	y	Y				ingest by a few people. Drupal content moderation	Depending on specific requirements, there may be a solution for this already.	Are you able to provide more information on this?
31	Submission, self	n	Y					Same	
32	Support, community		Y		1			The community has the Islandora Stack, github issues, google mailing list.	
33	Support, hosting		Y					We offer a paid support subscription that has a web portal and detailed knowledgebase.	
34	Support, paid		Y					See above.	
35	Viewer, audio	Y	Y					Supported in 2.x	
36	Viewer, books	y	Y					Integrating with Mirador for this with a current client. Mirador is highly configurable and can turn it's features on off to get your desired experience.	Was this implemented? Will HOCR also use Mirador?
37	Viewer, images, large	y	Y				Used for large and small images. Offers pan/zoom.	Currently openseadragon	Can large single images be viewed in Mirador instead of Open Seadragon?
38	Viewer, images, small	y	Y				Not used. Only for embedded images.	Currently openseadragon	Can small single images be viewed in Mirador instead of Open Seadragon?
39	Viewer, newspaper	N	Y				With segmentation? No longer needed. Now here: https://newspapers.lib.lehigh.edu/. NP like content can be pushed to book viewer.	Book solution should be repurposable for newspaper.	
40	Viewer, PDF	y	Y					PDF.js	
41	Viewer, video	Y	Y					html5 viewer	
42	NEW items added 2022-05-10								
43	per-collection derivative sizes								
44	ingest size limit, web								
45	ingest size limit, cli								
46	NEW items added 2022-05-18								
47	Collection scoping						Related to lines 4, 10,11,16, 43 above, we are interested in ways to scope config/interaction to individual collections, for example, brand, specify sort, facet, derivative settings, and than

This 50+ row spreadsheet was our actual criteria matrix that we used to communicate with vendors. What was really difficult about this was identifying the bare minimum features we couldn't live without as opposed to the features that were just nice to have. It's also a moving target as new features are constantly being developed and becoming available.

I found that the available repoistory platform options can be neatly divided into two groups, open source and proprietary. The proprietary options have lots of benefits including more polished themes and designs, new and innovative features like Quartex's handwriting recognition, and integrations with more digital preservation tools. These were pretty much off the table for us due to the cost of storage and descriptions.

That left us with the open source options, of which there aren't as many. We quickly decided that DSpace was too IR centric. Samvera and Hyku wasn't as turn key or out of the box ready. We didn't have the technical staff to create our own samvera repository from scratch. With that, we were down to Islandora, which seemed to meet most of our requirements, particularly with the Islandora Scholar module.

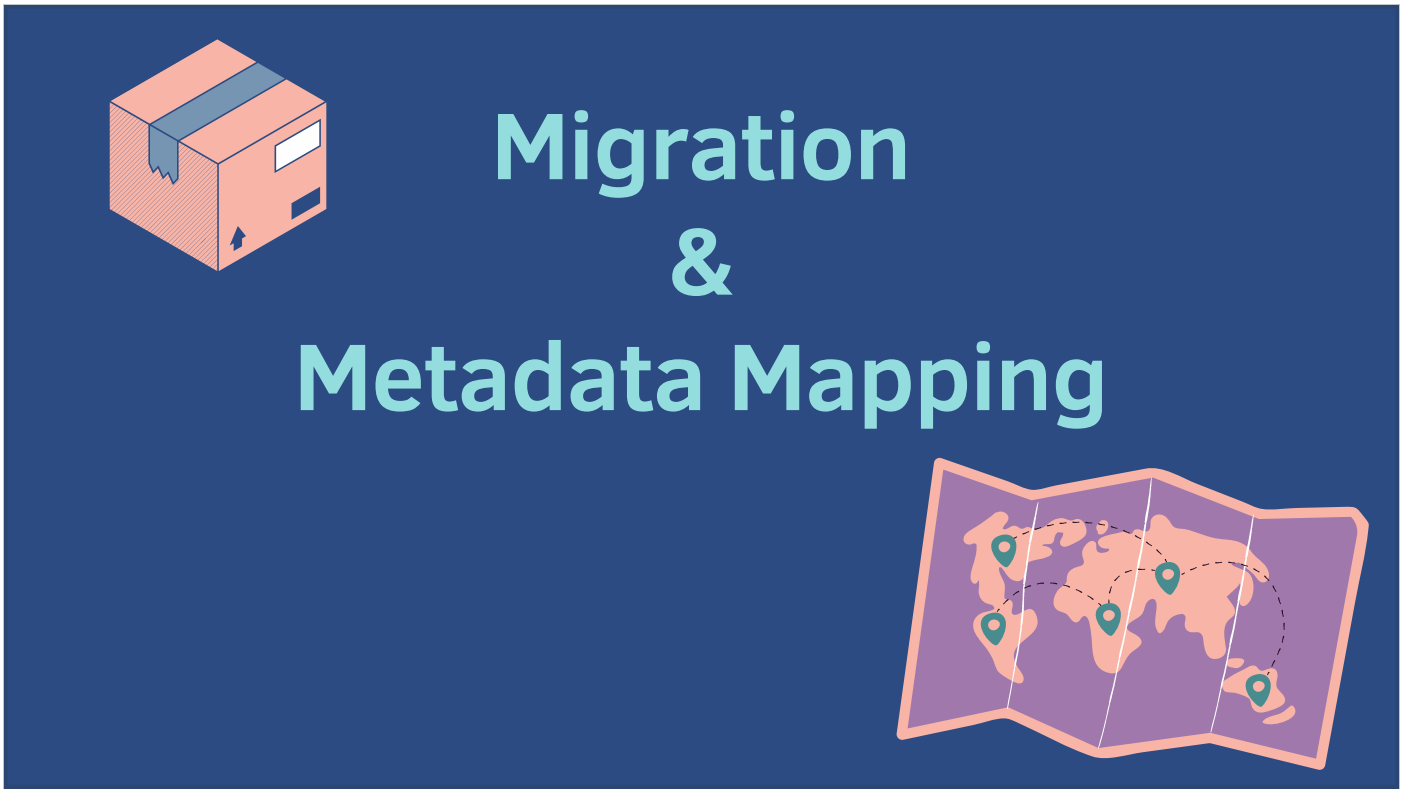


Vendor Selection

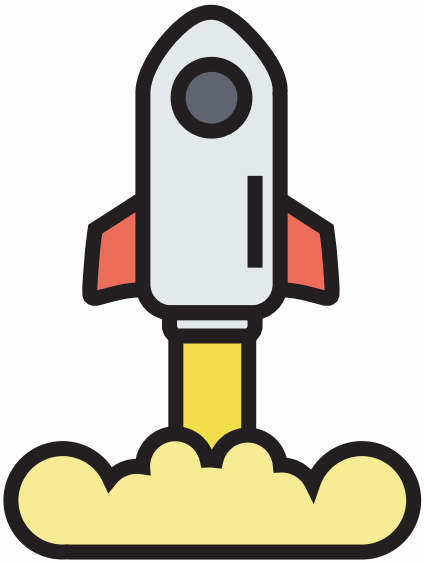
DO YOU NEED A VENDOR? WHICH ONE?

- **COST**
- Version of Islandora (Isle, community, etc.)
- Support & maintenance
- Theming
- Ease of migration
- In-house skills

Once we had the platform picked, we needed to decide on a vendor. The biggest factor here was cost. As contractors, these vendors CAN do pretty much whatever you want, but paying by the hour can get VERY expensive. For our current system, we decided to go with discoverygarden due to some of the customized features they had implemented, particularly around the migration and ingest tooling. We had decided to migrate on our own to save money, so having easy to use tools was important.



The metadata mapping and file migration took a year. Luckily, I had a lot of time to spend on it because it could be done at home during COVID. We had to create two different mappings templates for the digital collections and IR as different content had different needs. In CDM, each collection had different metadata, so I needed to create mappings for each collection trying to merge them to a single standard. This was a very iterative process that included endless failed ingests, missing metadata, or poorly formatted files. We eventually got all the content migrated over, but even 2.5-3 years later I'm still discovering odd issues and the result of what seemed like good decisions at the time but no longer match our standard procedures.



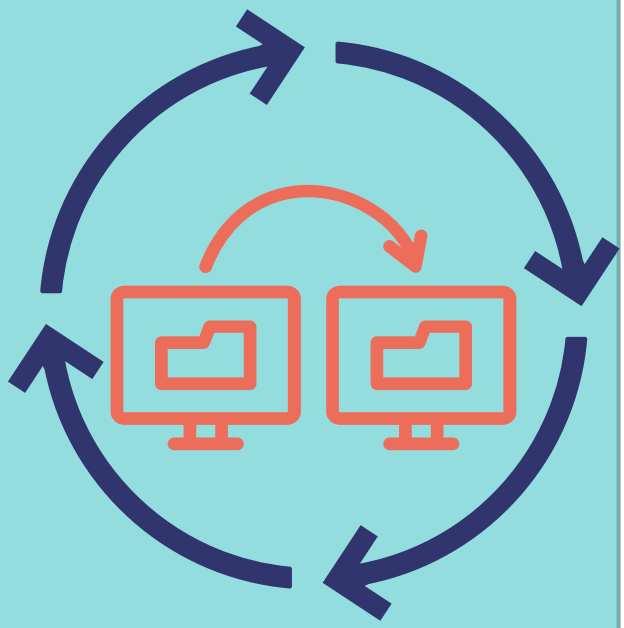
Launch & Review

- Launched in 2020, three years after initial decision
- Missing features
- Complex & challenging configuration
- Lackluster theming
- Vendor friction

I won't focus on the launch of our Islandora repositories as the them of this talk is selection. The main take aways are that it took about three years to go from decision to migrate to actually having a new platform. I don't think this is a crazy time span for implementing a new platform, but I'm curious to hear from everyone else.
Another takeaway is that just becuase a platform supports a feature doesn't mean that it can be easily implemented. Dealing with an outside vendor also adds complications. on the one hand, it took the technical load off of us, but the communication and negotiation still required a good deal of work.

Continuous Migration

- Islandora/Drupal 7 end of life
- Reassess selection criteria and options
- Islandora 2.X features in development



Now we are in a phase that my coworkers and I have started calling continuous migration. After being on CDM and Bepress of 8-18 years, we are now migrating again off of Islandora 7 after only seven years. This is mainly due to Islandora being based on Drupal 7, which is reaching end of life. Islandora 7 is a very mature platform with lots of modules and additional features, so we are not migrating to expand the functionality of the repository. Based on the 50+ row selection matrix I showed a few slides back, we drastically stripped down our selection criteria. Based on funding constraints, we were asked to deliver a pretty minimum viable product. However, many advanced features are currently in development and have anticipated release time frames in the near future. We hope that Islandora 2.0 will be more sustainable and long lasting due to the increased integration with Drupal. I don't believe that a Drupal 7 to 8 type architecture change is planned on the roadmap, so it should last for a bit. When the next migration does inevitably become necessary, hopefully the more mainline Drupal implementation will help smooth it out.

Selection 2.0

LESSONS LEARNED

Reassess in-house commitment

- Outsource migration to the vendor to save staff time.
- Understand configuration complexity from the start.
- Hire dedicated staff.

Merge collections and platforms

- Why have a separate institutional repository and digital collections?
- Moving from two Islandora repositories to one is easier to maintain.

Set realistic expectation

- Understand what features are available now vs. in the future.
- Limited customizability.
- Just because the feature is technically possible doesn't mean it can be easily implemented.

So based on the past six years of repository selection and migration experience, what lessons have we learned at Lehigh. The first is that you need to carefully and honestly asses the in-house commitments you are willing to make. While doing the migration ourselves was successful the first time around, we decided that it wasn't worth the staff labor, so we opted to outsource the migration work to the vendor. Configuring a repository to suit your institution's specific use cases is half the battle. Getting a clean install of a repository with everything set to default is not going to meet everyone or really anyone's needs. While hiring new staff isn't possible at many institutions, we are lucky to have an active search underway to add a repository centric developer to the library technology team. I'm sure the repository won't be their only responsibility, but it will enable us to be less reliant on a vendor

Merging our collections and platforms is a natural fit. While institutional repositories do have some specialized features, the content types can easily live alongside special collections digitized content. Having a single platform will hopefully be easier to maintain compared to two separate software platforms or two instances of the same software. This also encourages us to add new types of content, like research data, to the unified repository in the future, as well as potentially merging in some functionality that was previously moved out like newspapers.

Lastly, it is crucial to set realistic expectation to not mislead yourself, your stakeholders, or your administration about a repository. Some repositories offer flashy features, but won't necessarily have a robust user community or dedicated vendor for continued improvement or support. On the other hand, some repositories may be bare bones now, but have new features planned on the horizon. Having a development roadmap will help inform current decisions. Customizability of a

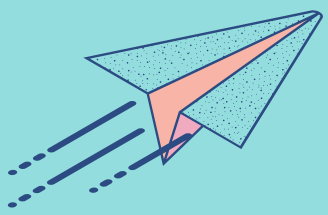
site system may not be easy or possible. We were promised the support of our internal web development team that never materialized, so our site design has more or less stayed the same since it launched. We were able to add on a few modules and plugins, but many of the features we thought would be present were never configured correctly, so they don't work. For example, Was Islandora 7 capable of Google Scholar integration? Yes. Does our site implement it well? No, because the configuration is complex and we haven't had dedicated staff to problem solve.

To wrap up, every institution's needs will be different and there is no one size fits all solution, but the selection process is going to be similar for many institutions, and can be improved through peer converstations.

Thank You!

More Questions?

Get in touch!



Christine George, she/her

Associate Director for Research & Scholarship
New York University Law Library
christine.george@nyu.edu

Alex Japha, he/him

Digital Archives and Special Collections Librarian
Lehigh University
asj316@lehigh.edu

Kevin Clair, he/him

Digital Collections Librarian
Penn State University
kmc35@psu.edu

Bethann Rea, she/her

Digital Collections Management Librarian
Penn State University
baz5008@psu.edu

Elizabeth James, she/her

Digital Archivist
West Virginia University
elizabeth.james1@mail.wvu.edu