Thoughts about a repository migration

thoughts produced, questions answered, and questions avoided

by Joe Cera and Michael Lindsey

https://n2t.net/ark:/85779/j41m23

Berkeley Law Library

Why migrate?

Quick background: We were using the repository for journal sites for the majority of the 13 student-run Berkeley Law journals and to display faculty publication lists attached to faculty profiles.

We decided to look into migration when we had a number of faculty members tell us that they didn't like the way that faculty publications were displayed on profiles.

We had an interest and a bag of skills that made us believe we could make the views more suited to our law school's look and feel

Most importantly, we had a lot of meetings and everyone involved in the decision making was open to discussing the possibility of migration.

The tipping point that made us commit to a migration was the ability to pull data in over the API and an understanding of templating:

- create one template for all faculty members
- create one basic template for all journals.

Berkeley Law Library

What we did during/after the migration

In preparation for the migration we worked with catalogers to help with mapping the bepress data to the TIND platform.

- Mapping non-MARC items to MARC items was somewhat straightforward helped greatly by the others in the room having recently worked through an ILS migration
- It made us think a little deeper about how repositories generally approach metadata and the lack of any kind of standard will result in challenges no matter the platform

We preserved the bepress snapshot and saved it to Google Drive

- A challenge for humans but not as much for computers? Every file associated with a record was simply named 'fulltext.pdf' - For the few items that didn't match, it made tracking down the mismatch a challenge.
- The export was more of a platform-specific data dump, rather than a semantically rich markup

We put the bepress identifiers in the TIND records

- helped us redirect requests for scholarship.law.berkeley.edu urls that were in the bepress form
- we use(d) mod-rewrite or something similar
- this connected searches for previous repository items with the new location of those items

The biggest surprise was that it felt as though we were migrating from two different repositories.

For the purposes of migration, Digital Commons and SelectedWorks were closely related but different systems requiring extra work.

Berkeley Law Library

APIs are the best

In our initial planning, we could see how the use of an API would let us do the primary thing we were interested in: displaying faculty publications in a way that matched the look and feel of the website and was flexible enough to be changed as we needed it.

This turned out to work exactly as we planned but does so much more:

- Having a write API on the target system was useful for batch operations during data cleanup (and there was a lot of data cleanup)
- The API lets us focus on the design or redesign of pages and the data on those pages with the confidence of knowing that, if the data exists, we can pull the data in
- Instead of putting items into the repository and hoping that people find them, we are putting items in the repository and thinking about where we can pull that data out and how we want it structured.
- The API has been used for faculty publications lists, journal sites, projects with the law school's Communications team, DOI scripting, and various other projects.

A proper API lets us express our data however we want, in current projects and in any new ones in the future.

Berkeley Law Library

What we found important

Carrying over existing systems to keep what works is important.

- TIND has built in DOI support but we had a working system through EZID and all of those systems continued without interruption
- Preserving keys from the legacy system is sound practice

We are not missing the loss of the ability to 'collect' from other structures.

- While this was often convenient, the limitations of collections and the lack of understanding about how those items were connected created a number of issues during use and during migration

Portability is calming

- Portability is an important point and hopefully will become part of everyone's organizational and personal data hygiene. having to scrutinize some vendor's idea of how to organize an export is no fun
- Building our own views through the API means we have a fuller understanding of the data models, we can control the presentation of repository data, and all of the views are independent of the repository platform (a functional API is very important)

Discussions focused on statistics and portability.

We had the luxury to choose portability and we believe that the increase in capabilities and reuse of the repository data has proven the value in that choice.

Berkeley Law Library

Issues that still exist

Statistics remain the biggest issue.

- We collected the statistics before migration but the two platforms do not collect statistics in the same way
- The best we can do is, for each item, connect the statistics from one platform to that of another but it really highlights the challenges of accurately counting things like downloads.

This experience has made us realize how important it is to know how any connections between repositories actually works.

- Any collected item that relies on the item existing within a platform will break if an institution migrates to another platform (or at least that is our understanding)
- This highlights the importance of platform independent identifiers like DOIs.

Few of us think outside of our own repository ecosystem. This makes it difficult to think about interoperability and the evolution of repository services.

Berkeley Law Library

Questions to ask

Ask about the details of a migration

- What does the data look like?
- What won't be captured in a migration?
- Ask others who have migrated to get a sense of what to expect

Consider the portability of your situation

- How dependent are you on the platform?
- How dependent are others on your data and the permanence of that data?
- Are you able to actually use your data in the way that you want to?

Think about the support from a new platform

- What is the direct support from the vendor?
- What does development look like and how is it communicated?
- Is the community a helpful resource?

Something to shop for with any data product:
Quality of experience getting data in and getting data out.

Berkeley Law Library