1-1-2020

Messy Data Secrets Revealed: How We Digitized Theses and Improved Discovery

Anne Burnett
University of Georgia School of Law, aburnett@uga.edu

Rachel S. Evans
University of Georgia School of Law, rsevans@uga.edu

MESSY DATA SECRETS REVEALED:

Bound theses from the law library's collection

HOW WE **DIGITIZED THESES** AND IMPROVED DISCOVERY

By Anne Burnett and Rachel Evans

E mployees in libraries come and go. Collections change over time. The systems we use to share these collections—and the fields of data those systems have—are constantly evolving. The fluid nature of IT, discoverability, and access creates a host of avenues down which item details sometimes fall through the cracks. This article tells the story of one such collection, tracing its origins from accession and early digitization through to our continuing consolidation and cleanup of decades' worth of data. Along with the benefits of providing open access to the scholarship in the theses, their inclusion in the institutional repository (IR) has been advantageous to the LL.M. graduates themselves. They are now able to provide potential employers with a URL (or a professionally scanned PDF) when asked for a copy of their thesis.

Background and Collection Scope

The world map from our Digital Commons analytics dashboard plots the number of downloads the LL.M. Theses & Essays collection has accumulated since August 2006 from more than 150 countries.

Early Digitization

The digitization of this collection has occurred in a piecemeal manner. The first LL.M. theses added to our online repository were those that already existed in the university's Electronic Theses and Dissertations (ETD) Database. The graduate school began requiring all master's degree candidates to submit their theses in PDF, along with hard copies, in the late 1990s, allowing us to harvest many files from the ETD Database for inclusion in our Digital Commons repository. Unfortunately, not all LL.M. theses submitted after that requirement was implemented actually made their way to the database, as we discovered when we compared our catalog holdings with the contents of the database.

In addition, we had several decades of LL.M. theses predating the ETD Database—all of which required digitization. This meant we had around 170 additional hardbound

HOW TO THINK LIKE A GOOGLE BOT: SIX TIPS FOR IMPROVING DISCOVERY

Institutional repository (IR) platforms such as Digital Commons rely on Google and other search engines for discoverability. They depend on different sets of criteria than ILSs so that Google can effectively index our site, the collections, and the items there. Part of the motivation for cleaning up this and other series in our repository was to improve their discoverability. Here are six tips we try to keep in mind as we design our repository record metadata and when we are adding new items and cleaning up collections, in order to maximize discoverability on the open web.

SIX IR TIPS *when*
DESIGNING *for* **BOTS**

01 Title Fields Are Important
02 Human-Readable Is Better
03 Don't Use Too Many Keywords
04 Frequency, Consistency & Longevity
05 Bots Like Quick Load Times
06 Site Maps Are Critical, Especially for "Dead" Collections

1. Title fields are important. They are, perhaps, the most important field of any object or event metadata in your repository. No matter how many other fields have the data or keywords you think end users might search for, if the title does not contain it, then Google probably will not retrieve it (unless you have big bucks, of course, in which case you can use AdWords to pay your way to the top).

2. Human-readable is better. Your IR is not your library catalog. Your ILS is a (mostly) closed-off system. It was engineered by librarians with very strict cataloging rules, passed down over decades of meticulous fine-tuning with a field for every single possible bit of data. An IR is not an ILS. In the same way, Google is not your OPAC. Sure, you can use some of the same operators, and you may even form similar strings in each of the search bars. But the difference is that Google's algorithm is not a 100% known entity. Most of Google's users are performing natural language searches. Your IT or metadata librarians cannot access Google's back end and tell it what you want, what fields to provide searches for, what weight to give certain types of results, or how to display your results list. Google's algorithm not only likes, but craves, human-readable, not machine-readable. To recap: Google = human-readable, OPAC = machine-readable (aka MARC).

3. Don't use too many keywords. In addition to not getting overly wordy or technical in your IR fields, the field to especially watch out for is keywords. In Digital Commons, there is a nice keyword field. When we first started adding content to our repository, we went overboard with keywords. Too few, however, could hinder discoverability. The sweet spot to aim for is two to four keywords for Google's crawler. Experts in discoverability and SEO warn that if you use too many, search engines will actually ping or potentially ignore your content. Excessive metadata makes it assume this content isn't valid. Just be careful. This doesn't mean you should never use more than four keywords, but keep the majority of your content with three keywords or fewer to help search engines take your site more seriously.

4. Frequency, consistency, and longevity are key. Web developers are asked far too often, "When will Google crawl our site again?" This is a mystery to (almost) everyone, and while you can request a recrawl via Google's Webmaster Tools, there is no guarantee of the speed at which that will actually happen. One thing is certain: You will be recrawled more often the more frequently and consistently you update any site, including your repository. Long periods of no activity may result in flagging you as a dead site, so regular uploading, or even refreshing, content helps. Another related factor is longevity. This is simply the idea that the longer a site exists, the more time it has had to be crawled and the more it has appeared in search

theses (ranging from 85 to 200-plus pages) that required digitization. We were able to take advantage of a subsidy offered by LYRASIS at the time for digitization projects. The actual digitization was completed by Northern Micrographics, which was able to scan the items without damaging the bindings.

Because the theses were originally printed on one side of the page and scanned by high-volume book scanners,

results; this increases site traffic. Then this cycle returns to the beginning, because the more site visits your repository receives from organic Google searches, the more your site should rise in the results list. Your IR and its collections are becoming more closely associated with these searches over time. New sites will take time to get to that cycle point, but after many repeats (with the help of your frequent and consistent care and feeding), this will happen naturally.

5. Bots like quick load times. Since we don't really know when Google or other search engine bots will give the IR a visit, how can we make sure that when they do, they are finding you at your best? Load times are a big indicator. Things like high-res images of the latest guest lecturer, embedded Issu flipbooks of scanned books, and YouTube video of a 3-hour conference panel mean that multimedia may require a different approach. Some IRs have native streaming, which helps cut down on load times. If not, you may need to choose what is more important—the load time or the media—for keeping traffic coming to your repository. In the end, the faster your content loads, the more quickly it can be indexed. Bots are impatient. Make them wait too long, and they might just keep moving.

6. Site maps are critical, especially for dead collections. Even if your digitization project is not open-ended and you wind up with a static archive, Google will not forget about you as long as your site has a good skeleton in the form of a site map. Depending on the repository platform you are using, a site map may be generated for you as you create new content. It never hurts, however, to revisit your site structure and organization. Particularly for sites that have been around over a longer period of time, the site map may be pulling titles and other structural and organizational information that are either no longer accurate or are not as good as it should be. Structure expands as collections grow and new collections are added. Revisit your site map every so often as a regular maintenance task. It is essentially an outline of your site and all that it contains. As you make adjustments, monitor how your site (hopefully) rises in search results over time; increases in repository visits and downloads should follow suit.

an unforeseen issue was the inclusion of a blank page between every text page in each PDF. As we faced the grim prospect of student workers tediously deleting dozens or even hundreds of blank pages from each thesis, one of our student workers, Jennifer Cotton, mentioned that her coder father, Jay Cotton, had written a quick script that would easily remove the blank pages, adding only seconds per thesis to our workflow. The script that worked like a charm within the Adobe Acrobat console window is the following:

```
for (var p = 0; p < this.numPages; p++)
{
if (this.getPageNumWords(p) == 0) {
this.deletePages({nStart: p, nEnd: p});
}
}
```

Rise of the Repository

Since 2006, when the University of Georgia's Alexander Campbell King Law Library (UGA Law Library) began using bepress' Digital Commons platform as our official school repository, we have fairly consistently added theses to this collection. This series can be browsed and accessed at digitalcommons.law.uga.edu/stu_llm.

The metadata fields for repository records are different from our MARC records in the library catalog for the same items. Repository records contain fewer fields. Of course, much of the information in the fields is the same, but the field titles and formatting style are quite different.

Our LL.M. theses records in our repository include the following required fields:

- Title

- Authors

- Document type

- Publication date

- Keywords

- Disciplines

- Abstract

- Upload file

Using MARC Metadata to Generate Collection Lists

A major challenge in developing our comprehensive theses repository collection was simply creating a complete list of the items in the collection from the records we had. We turned first to MARC. Understanding the different purposes that the ILS and the IR serve helps in translating the data from one system to another, like we had to do when we found ourselves cleaning up data that has lived, disconnected, in two places. Although our MARC records for the LL.M. theses collection contain more fields than in our IR,

▶

Sierra Global Update

there are strict rules for how to format each field, making the decisions and guesswork less of a factor than for repository record cleanup. Thick books such as the AACR2, online MARC resources (loc.gov/marc/bibliographic), and RDA toolkits all provided clear guidance for double-checking fields and fixing data-formatting errors. However, a single missing "." or an additional space or "-" can throw your control fields out the window. With our LL.M. theses, some sleuthing support from multiple colleagues helped to locate one major inconsistency in the 502 Notes field—the very field we thought was our best control for creating complete lists of this collection in our ILS.

It should have been formatted this way:

|aThesis (LL.M.)--University of Georgia, 2007

Instead, it appeared in a variety of ways, including |aThesis (L.L.M.)--University of Georgia, 2007, i.e., with an extra period after the first "L"; |aThesis (L.L.M.) University of Georgia, 1997, with an extra period after the first "L" and missing hyphens; and |aThesis (LL.M)--University of Georgia, 1998, with no period after the "M."

Another inconsistent field that would normally be a good control to perform searches and create lists would have been the location field. However, the hard-copy theses had recently been moved from Reserves to our basement archives shelving, and locations had not yet been updated. About half of the item records had a location code of [res] and the other half [rbbas].

For our library and this particular collection, we finally located an unlikely but more consistently similar field, the donor note. As it turned out, most of our theses collection had a donor note stating "LLM Theses." This unexpected commonality made for a better control field than the others we had attempted to use for compiling an item list in our

ILS. This search returned 330 results—more than we had in our repository. To do this in Sierra, we used the Create Lists function to pull our largest set of items using that donor note. Next, we exported the list to text delimited files so we could better see and sort the data. The fields we exported included the following:

- 245 (title)

- 100 (author)

- Cataloged date

- 260 $c (publication date, subfield)

- 502 (note field)

- Bib location

- Item location

- Item status

- 856 $u (access URL, subfield)

Cleaning Up Data With Excel, Batch Revise, and OpenRefine

Now we had a fairly solid list from our ILS and a similar one from our IR. Every list in each platform already had some basic cleanup to get the lists as consistent as possible, and there was less than a 30-record difference between the two lists. Here are some of the approaches and tools we used to clean up the records in both our ILS and IR systems.

For Sierra ILS

Excel spreadsheets were the first tool we used after exporting MARC record data from Create Lists in Sierra. Excel helped with visually formatting column headers and doing basic manipulations, such as sorting alphabetically. With a few sorts on the note field, we quickly saw the scope of inconsistencies. This revealed problems with our locations, item status, and 856 fields. It also showed some sloppy cataloging, where someone must have been macro happy. There were random bits of information in some places, such as dimensions in the 260 field and the location of publication in an 856 field. Other fields were blank. Using some color highlights of cells and rows, we found where records in the ILS needed further cleanup. We also identified a list of titles to check if they were in the repository at all (those items lacking the 856 $u).

We did global updates and manual record changes for larger or smaller subsequent lists, making incremental corrections. This process involved revisiting print office copies of the AACR2 and double-checking the MARC 21 bibliographic standards.

For Digital Commons IR

Batch revise in our repository allowed us to download a CSV file containing all LL.M. theses in that series.

Records were exported to an Excel spreadsheet for checking.

Formatting and sorting this list in Excel (just as we had done with MARC records) allowed us to clean up repository records. Again, we could clearly see inconsistencies. For our IR metadata fields, the "document type" field was first, followed by "keyword" and "discipline" fields. One

THE CONTENT IS UNIQUE

TO OUR INSTITUTION AND CANNOT BE

OBTAINED THROUGH OTHER SITES,

DRIVING MORE TRAFFIC TO OUR REPOSITORY.

easy batch revision was the "document type" field. They all should have been "dissertation," but many had the default choice of "article" instead. After revising the Excel sheet, you can upload it directly back into Digital Commons, making this process a little faster than doing the same work with Sierra.

Transforming Records With OpenRefine

We also noticed from a quick visual comparison of title field data in our ILS and IR that many IR title fields had the titles in all caps. On the flip side, many MARC records in the ILS had the author name following a "/" in the title field.

OpenRefine was our next tool to deploy. For those new to OpenRefine, there is a learning curve, but it is not unlike Excel, Google Sheets, or any other spreadsheet application. It functions as your own little server though, so you may

need the assistance of IT to install it properly on your office computer. There are excellent Wikis and a robust GitHub space with plenty of resources, including screencasts at github.com/OpenRefine/OpenRefine/wiki/Screencasts.

And it was from there that we found the Online Library Environment and GOKb, a freely available data repository containing key publication information about electronic resources. Within this treasure chest of information was a how-to ("Comparing Two Sets of Data in OpenRefine"; openlibraryenvironment.atlassian.net/wiki/spaces/GOKB/pages/655657/Comparing+Two+Sets+of+Data+in+OpenRefine).

This was the missing piece of our cleanup puzzle, a few transformations that allowed us to match the title and date fields from the IR and ILS LL.M. theses records. We were able to compare the two OpenRefine projects, specifically comparing the titles to find matches, use expressions for iterative matching, and normalize the titles to make all of this happen. The transformations used were pretty conservative, but they were just what we needed, including the "fingerprint" functions that can do the following:

• Replace all punctuation with the space character

• Convert a string to lowercase

• Break on white space

• Convert characters to the nearest ASCII equivalent

• Sort an array of words into alphabetical order

• Rejoin an array of words into a string with a single white space between words
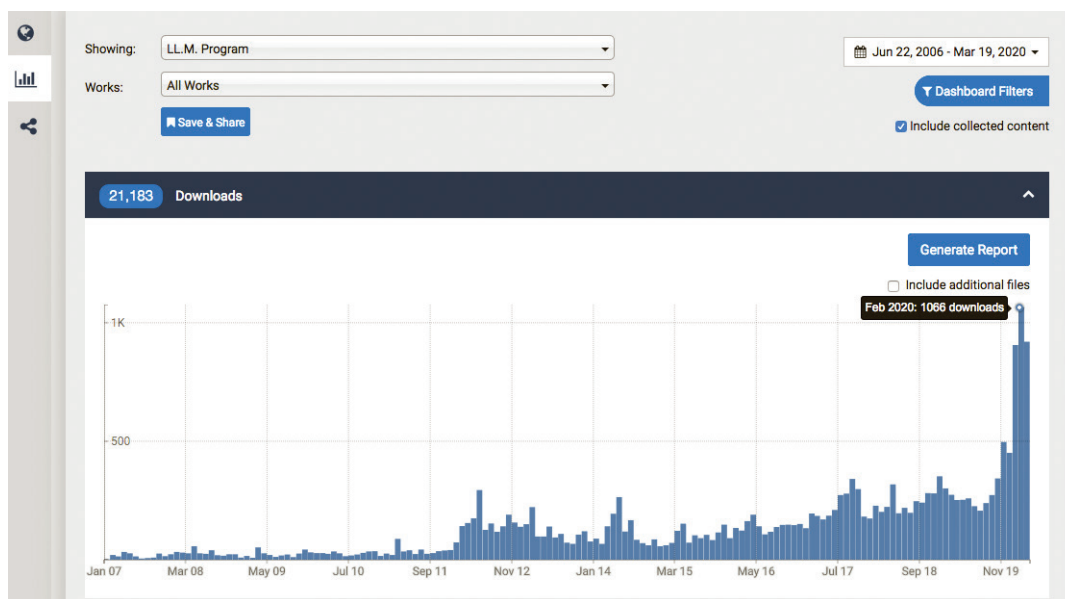
Discoverability has been dramatically improved.

Discoverability Awareness Reinforces Cleanup

During this process, we enjoyed taking advantage of our free discoverability report from bepress for our Digital Commons repository. A major takeaway from the report was the highlighting of our LL.M. theses collection. It turns out that this is one of our most popular series, and we were encouraged to continue with more collections like it. The content is unique to our institution and cannot be obtained through other sites, driving more traffic to our repository. It is also one of our most globally diverse collections as far as access and download. It has been accessed from more than 154 countries by 1,227 institutions, with more than 53,800 metadata hits and 385,000-plus downloads since August 2006. It contributes a great deal to the overall discoverability of our site.

Although our cleanup is still in progress—particularly the physical thesis inventory in our archives area—it has already proven to be worth the time and effort. Since cleanup began in October 2019, our referring URLs from the ILS (thanks to the updates to many MARC 856 fields) have gone up, as more users searching the catalog are able to access our IR items straight from the OPAC item record. Since November 2019, when we started updating the LL.M. series structure in our repository, downloads have skyrocketed from their cross-listed location in the LL.M. Program series. Also since November 2019, we have seen our batch revision work pay off, with increases in repository visits to this collection and increased downloads for the LL.M. theses and essays.

ABOUT THE AUTHORS

Anne Burnett, foreign and international law librarian, and **Rachel Evans**, metadata services and special collections librarian, work at the University of Georgia's Alexander Campbell King Law Library in Athens, Ga. There, they have collaborated together on repository projects such as the example in this article. For more information about this ongoing work, contact rsevans@uga.edu or aburnett@uga.edu.