1-14-2020

Cleaning Up Messy Records: Uncovering Match-Points in ILS and Repository Data

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

TSLL TechScans

A Blog to share the latest trends and technology tools for Technical Services Law Librarians

TUESDAY, JANUARY 14, 2020

Cleaning Up Messy Records: Solving Mysteries in Catalog and Repository Data

Many of us have our hands in multiple pots. Sometimes we are working with repository records, and other times records in our ILS. While embarking on what I mistakenly thought would be a simple series of tasks (linking from 856 fields to our freely accessible, digitized versions of the same items to our IR), I happened to uncover much messier data than expected. What a perfect opportunity to do some house-cleaning! In case you have undertaken similar work, and are considering comparing, cleaning, and (eventually) updating records in multiple locations, here are a few tips and resources I have found helpful on my own journey:

Have good list from the repository. My initial cleanup of Digital Commons items was fairly straight forward. I sat down with our International Law Librarian Anne to talk about the collection. We batch-downloaded the series as a spreadsheet, and updated the fields that did not match so that they did (example: type was not the same for each, some were articles, some dissertations - this one was easy, updated them all to be "dissertation" type). Then I saved it, and re-uploaded the batch sheet.

Have a good list from the ILS. The same set of items in our library catalog was more difficult to get a solid list of. We are in Innovative's Sierra, so I needed to use "create lists". This time sitting with Associate Director for Collection Services Wendy and I generated a list of items (not bibs). We made several lists, because we quickly discovered that what we thought our control field was (the 502 note) was inconsistent. Catalogers had changed over the many years these records were created, and at a certain point the 502 had changed from having only two periods in this item's abbreviation to having three (ex. LL.M. to L.L.M.). This minor flaw made things a bit more difficult. We ended up instead using the donor note field to get closer to the ideal number of items from my repository list. *In the future the location might also be a field to use for pulling this sort of list - but part of the ILS record cleanup was updating location codes, since items had recently been shifted from reserve to the basement.*

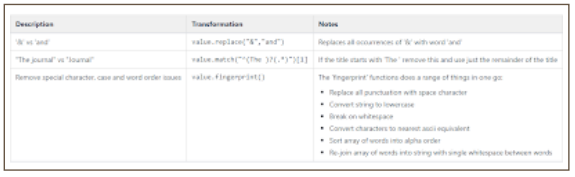
Fix your controls so they actually work. I ended up going with the list that had the highest item count (although now I had *more* than what was in my Digital Commons list) and updating these records first. Now they will have consistent 502's, and correct item locations. To verify the proper 502, Wendy and I consulted our office copy of the AACR (Anglo-American Cataloging Rules) -

- "Section 2.7 B13 Dissertations". Although LL.M. was not listed as a specific example, the rule states that you use: "*Thesis* followed by a brief statement of the degree for which the author was a candidate (e.g.. M.A. or Ph.D.), the name of the institution..., and the year in which the degree was granted."

To do this I used a combination of Global Update in Sierra and manual/individual edits. I added 856 fields to the records as well with subfield u (linking to the Digital Commons series landing page for this collection) and subfield 3 (for the text I wanted to appear as the hyperlink access point). You could do some of this in Rapid Update as well, if you're feeling extra confident! - *Note: In Sierra I had to start with a list of item records, but at a certain point I had to re-run my partially cleaned-up list to create a new list of bib records. I could only globally update fields in the bib record with a list of bibs (a list of items would not work!).*

Export a better list from the ILS. Now that I had a proper list from Sierra, I exported it to a text delimited file, then imported it into an Excel spreadsheet. I was ready to compare it to my repository spreadsheet and figure out what what missing in Digital Commons.To figure this out, I started by doing a "save as" of each spreadsheet, then narrowing both sets of data down to only the fields that I could use to compare between the two: Title, Author, and Publication Date (year was really what I was looking for). This presented further problems - not all titles in Digital Commons looked like what should be matching titles from the ILS records (ex. many repository title fields had been entered in all caps!). For Author, Digital Commons separated last and first name fields, but in the MARC records this was a single field. For Publication, the formatted date in Digital Commons records was very detailed and specific, while the only match-point in the ILS records was the 260 field (included publication date at the end as the subfield $c) - major thanks to our Collection Services Manager David for knowing this one off the top of his head! The 502 might again prove useful if the 260's were too difficult (since the "year in which the degree was granted" appeared at the end of this field for each item).

MarcEdit, OpenRefine, and beyond. At this point in my process of this particular project I am playing around with a combination of editing in MarcEdit, as well as a number of tricks I am reading up on for OpenRefine cleanup. If you are totally new to OpenRefine, there is a really handy wiki with screencast intros that I appreciated before diving in. So far, an extremely helpful resource I have come across has been *Comparing Two Sets of Data in OpenRefine How-To*. This entry shares step by step how to "Normalise titles to do comparison" using three main transformations. For my particular set of data, the value.fingerprint transform has given me good results, removing case from both sets of titles:



There is also an excellent page with more information specific to working with and cleaning up dates. I am still working with the cleanup of this set of items, but even though it is a work in progress this has been a wonderful learning experience. Each time I work on it I learn something new! I am excited about the things I have figured out in this process that can be applied to other sets of items in both our repository and library catalog records in the future. I'd like to thank several of my colleagues at UGA Law Library for providing various pieces of this project's puzzle. Without them I would not have made it this far with these particular data sets. *Thank you Anne, David, and Wendy for all your context, tips, tricks, and sharing your experiences with this collection.*

What types of cleanup are you doing with your library's data? What tips and resources have worked well for you? Please share with us in the comments!

Posted by Rachel Evans at 10:17 AM

Labels: cataloging, discoverability, metadata, repository

cataloging (321)
conferences (2)
digital images (2)
discoverability (2)
getting to know you (36)
government documents (25)
information technology (263)
library services platforms (1)
local systems (66)
management (77)
metadata (39)
new members (4)
preservation (74)
processing (17)
professional development (2)
quick question (4)
repository (6)
SEO (1)
serials (39)
web archiving (1)

Blog Archive

▼ 2020 (1)
▼ January (1)
Cleaning Up Messy Records: Solving Mysteries in Ca...
► 2019 (24)
► 2018 (31)
► 2017 (52)
► 2016 (38)
► 2015 (43)
► 2014 (42)
► 2013 (38)
► 2012 (48)
► 2011 (76)
► 2010 (80)
► 2009 (56)
► 2008 (150)
► 2007 (51)