David Provost

LIS 644

Metadata Edit & Migration

Reflection Paper

This assignment was very interesting for me. In a previous career, I did a significant amount of data cleaning, trying to standardize the data across research done at different times with different criteria. At the time, if tools like Open Refine existed, I (and my supervisors) were not aware they existed. I relied on Excel's tools (primarily Find and Replace) and manual editing to accomplish the task, but Open Refine would have been a huge help. For example, the tool we used to divide one cell into multiple cells based on a divider within the cell. As part of some previous professional development, I learned about and did some exercises using regular expressions (and websites like Regex 101) to clean up data. While it is extremely powerful, I think for most practical purposes I can envision, I would turn to Open Refine first, as it is much more user friendly, while still being powerful enough.

Additionally, this assignment revealed something that I have been thinking about as we have learned about the variety of metadata formats that are used in digital libraries. The goal of a unified and standardized schema seems both very close and very far away. In this case, moving the data into a MODS format from Dublin Core was not challenging, and much of it could likely be handled programmatically - mapping fields like "Date" to "Date Issued" is not a challenge, for example. The problem comes with fields like "Personal Role" and "Corporate Role" which must be carefully determined for every item in a digital library. It seems unlikely that a single unified format will be implemented because of labor-intensive changes like that, but if it could be done, and especially if it could be linked to MARC records and other types of library metadata it could be a huge improvement for the field.

One issue that was discussed in class was errors with the Date issued field, but based on the screenshots below, it looks like I was able to import that data successfully. It appears with other publication information in the 260 field of the MARC record, as viewed in the Islandora sandbox screenshots below. I did use Google Sheets to edit the spreadsheet data, so I wonder if there is some difference in the way Microsoft Excel converts dates to text that introduces some error. This was a fascinating assignment that really appealed to me perfectly. As someone who spent hours and hours standardizing the metadata and album art of my excessive iTunes collection back in the day, I think I have an affinity for this sort of work!