Collections Data, Tools, and Strategy: Applying R, Tableau, and Excel to Print Assessment

Lori M. Jahnke
Emory University, ljahnke@emory.edu

Chris Palazzolo
Emory University, cpalazz@emory.edu

Follow this and additional works at: <https://docs.lib.purdue.edu/charleston>

Part of the [Collection Development and Management Commons](#)

An indexed, print copy of the Proceedings is also available for purchase at:

<http://www.thepress.purdue.edu/series/charleston>.

You may also be interested in the new series, Charleston Insights in Library, Archival, and Information Sciences. Find out more at: <http://www.thepress.purdue.edu/series/charleston-insights-library-archival-and-information-sciences>.

Lori M. Jahnke and Chris Palazzolo, "Collections Data, Tools, and Strategy: Applying R, Tableau, and Excel to Print Assessment" (2019). *Proceedings of the Charleston Library Conference*.
<http://dx.doi.org/10.5703/1288284317141>

Collections Data, Tools, and Strategy: Applying R, Tableau, and Excel to Print Assessment

Lori M. Jahnke, Emory University, ljahnke@emory.edu

Chris Palazzolo, Emory University, cpalazz@emory.edu

Abstract

As is the case at most academic libraries, collection assessment has become an essential component of collection management and development work. Although much of the assessment focus has disproportionately fallen on e-resources, print collections remain fruitful areas for evaluation and review. At Emory, print collections, including a complex approval plan, continue to be a significant component of our overarching collection strategy (in volume and expenditure). However, shifting priorities for library space and the growth of interdisciplinary programs and centers within the university are placing a higher demand on subject librarians for communication and coordinated decision-making regarding print acquisitions. As a result, we are currently preparing for a comprehensive print collection review, of which the approval plan is an integral component. This assessment will inform a more coherent print strategy, which effectively and efficiently meets research and teaching requirements as well as administrative needs. Using data cleaning and visualization tools, such as R, Excel, and Tableau, we have enriched our local usage data with detailed GOBI approval data (e.g., series, publisher, subject, etc.) and profile parameters. Merging these data types and enriching local use data will allow us to analyze the print collection in a more nuanced fashion and ask questions that do not require the LC classification framework. This analysis considers the development of additional tools and approaches that facilitate subject specialist communication with collection management and overall collaborative decision-making, especially in cross-disciplinary areas.

Introduction

This material was originally presented as a poster (Figure 1) and the following text is an elaboration of some of the elements in the poster to provide additional context.

Our overarching goal for this project was to merge three data sets that each describe different aspects of print acquisition and management: (1) LC parameters from the GOBI Print Approval MOA, (2) GOBI expenditure data, and (3) local use statistics from Alma. Creating one data set from these separate sources will provide us with more flexibility in analyzing the print collection and allow us to ask subject-driven questions that cannot be answered by the more traditional categories of the LC classification framework. This project lays some of the groundwork for an upcoming comprehensive print collection review, of which the approval plan is an integral component. A goal of this print assessment is to develop a more coherent print strategy that supports the growth of interdisciplinary programs within Emory University and allows us to balance shifting priorities for library space. As a parallel goal we are using this work to develop tools that will facilitate communication among the subject specialists who are responsible for overlapping areas

of the collection. We also hope the development of user-friendly tools will promote more active monitoring of the approval plan and the on-site collection. Although we will expand these processes to our other approval plans, we started with GOBI since it is Emory's primary vendor and all selectors work with this plan.

We view the GOBI expenditure report as enriching the other data sets since it includes item-level profile data such as aspect, interdisciplinary topic, and select or content level. As we move toward another comprehensive assessment of the on-site collection, and a possible reduction in footprint, the content-level data may be particularly useful in setting priorities for materials that remain in the on-site collection. Our local use data collected in Alma includes in-house use and circulation data. The in-house statistics will provide us with another means of determining which materials are best utilized on-site.

Process

As we allude to in the poster, our GOBI print approval plan is highly granular and through its many revisions over the years, it has accumulated exceptions at multiple levels of the plan hierarchy and throughout the LC parameters (Figure 2). This has created a few

Figure 1. Thumbnail of the poster presented at the Charleston Conference, "Issues in Book and Serial Acquisition," November 6, 2019.

A	QR					
	Action		LC Range	Description	Fund	Exceptions
B	S	QR	1-502	Microbiology		
A	R					
	Action		LC Range	Description	Fund	Exceptions
B	X	R	1-920	Medicine (General)		Where excluded...
	B	R	1-722	General. History		
	S	R	127.5	Other ancient works		
		R	601-603	China		
		R	723-726	Philosophy. Ethics. Psychology. Terminal care		
	B	R	724-726.2	Medical ethics. Medical etiquette		
	S	R	726.8	Terminal care. Palliative care		Exclude Clinical...
		R	727-849	Practice of medicine. Medical education		e-preferred

Figure 2. Structure of the original GOBI MOA spreadsheet. Each LC class is a separate table within the same sheet that begins with a highlighted row, followed by a header row, a subclass description row, and then the range descriptions. Each LC section is followed by one empty row. The levels of hierarchy are labeled as A, B, and C. This structure is replicated throughout the LC instructions spreadsheet. This is a representation of the Emory MOA and it does not include accurate data.

class	subclass	range	subclass description	range description	subclass action	range action	subclass exceptions	range exceptions
Q	QR	352-354	Microbiology	Mycoplasmas. Rickettsias	S	S	None	None
Q	QR	355-502	Microbiology	Virology	S	S	None	None
R	R	1-722	Medicine (General)	General. History	X	B	Where excluded...	None
R	R	127.5	Medicine (General)	Other ancient works	X	S	Where excluded...	None
R	R	601-603	Medicine (General)	China	X	X	Where excluded...	None
R	R	723-726	Medicine (General)	Philosophy. Ethics. Psychology....	X	X	Where excluded...	None
R	R	724-726.2	Medicine (General)	Medical ethics. Medical etiquette	X	B	Where excluded...	None
R	R	726.8	Medicine (General)	Terminal care. Palliative care	X	S	Where excluded...	Exclude...
R	R	727-849	Medicine (General)	Practice of medicine. Medical education	X	X	Where excluded...	e-preferred

Figure 3. A section of the GOBI MOA file after it has been cleaned and transformed by the R script. Rows in the QR section were hidden in the original spreadsheet (Figure 2). Although the script cleans and transforms all rows, several were omitted here for space. This is a representation of the Emory MOA and it does not include accurate data.

challenges locally for interpreting how changes to the plan will affect multidisciplinary areas, which has impeded decision-making for selectors who share responsibility in those areas. Our 2018 assessment of the R, S, and T classifications provided ample lessons for how we might change our local processes and improve access to collections data for selectors.

As a starting point, we used an R script, written by Josh Fjelstul, to convert the GOBI MOA LC parameters spreadsheet (Figure 2) to a “flat” table. The R script uses the structural elements in the original file, such as highlighted or empty rows and the structure of the header row, which always begins with the value ‘Action’, to locate the relevant data in multiple tables and extract it to new variables. Once the data have been extracted to the new variables defined by the script, the transformed data can be downloaded as a csv file (Figure 3). Cleaning and flattening the GOBI file resulted 3,644 separate range segments, excluding K, which could then be assigned unique IDs (gseg_id) that we used to relate approval plan actions to the expenditure and local use data sets. The K ranges were excluded from this step since this part of the GOBI file has a slightly different structure. A subsequent version of the script will account for the distinct structure of the K tables and include these data.

Our script uses the Shiny app to create a graphical interface that selectors can use to explore the LC ranges and download the cleaned data set for

alternate views or manipulation. The app is shown in the upper-right of the poster (Figure 1). Selectors can use the app to view all aspects of the LC parameters that are relevant to their subject simultaneously. The app also includes other features, such as filtering the plan by exceptions or actions, and searching the exceptions fields.

Merging the Expenditure Data and Local Use Data

We merged the GOBI expenditure data and our local use data from Alma using the Fuzzy Lookup tool in Excel. This tool performs fuzzy matching of textual data and returns a similarity score with each match. It can be used to identify duplicates within a table or to merge tables based on matching selected fields, as we did here. We achieved good match results based on the title fields in the two reports, but depending on your local data, the ISBN field could work well or a match based on multiple fields might yield better results. Fuzzy Lookup is robust to spelling mistakes, synonyms, many abbreviations, and other errors. The same task could be accomplished with Excel formulas or by writing a script.

Matching Profile Segments to the Merged Data

The most complex part of this process was to match the GOBI profile segments to the title-level data in

the combined Expenditure-Use dataset. We accomplished this through a series of formulas in Excel that extract the range number from the item's LC call number and match it to the 'bins' represented by the gseg_id. For example, this is the formula that extracts the range number from the LC classification, where 'A2' contains the full LC classification for the item.

```
=IFERROR(IF(ISERR(VALUE(LEFT(MID(A2,
SEARCH(".",A2)+1,1),1)*1)),VALUE(MID(LEFT
(A2,FIND(".",A2)-1),MIN(FIND({0,1,2,3,4,5,6,7,
8,9},A2&"0123456789")),LEN(A2))),VALUE(MID
(LEFT(A2,FIND(".",A2)+1),MIN(FIND({0,1,2,3,4,5,
6,7,8,9},A2&"0123456789")),LEN(A2))),VALUE
(RIGHT(A2,LEN(A2)-MIN(FIND({0,1,2,3,4,5,6,7,
8,9},A2&"0123456789"))+1)))
```

Once the range number has been extracted from the call number, it is relatively straightforward to use the INDEX and MATCH functions in Excel to identify the correct gseg_id for the item and add it to the data set. In future versions of this data processing we will incorporate this step into an R script, which is much more efficient than Excel at handling large data sets. This will also allow us to display item-level data within the interactive tree structure of the R app.

Results and Future Directions

The resulting data set includes all of the expenditure data (e.g., fund codes, order type, cost, etc.), standard bibliographic data, GOBI profile data (e.g., aspect, content level, YBP select level, interdisciplinary topics), local use data (e.g., number of loans, in-house loans, last loan date), and the gseg_id for

the relevant portion of the GOBI MOA. With this enhanced data set we are planning an analysis of multidisciplinary areas of the collection to identify gaps in the approval plan, as well as other areas that could be updated.

For the last couple of years we have been using the Emory University implementation of Tableau to provide access to collections data through dashboards that allow selectors to choose from a variety of preconfigured and customizable views, or to download data as needed for additional analysis (see Tableau example in Figure 1, middle-right side). Moving forward, we plan to make this enhanced data set available through this platform as well, which will allow selectors to model the cost implications of changes to the approval plan, working as a group or individually.

As mentioned above, we plan to revise the R script to accommodate the unique structure of the K ranges and to incorporate the other data processing tasks that are currently performed in Excel. In addition to reducing the number of steps that could introduce error, extending the R script will allow us to integrate the item-level data with the interactive R app. Some of the selectors have requested this feature, but it will be challenging to display the detailed data in a manner that is still legible. Possible work-arounds could include automatic filtering, such as by fund code, budget year, or order type, which would select a more digestible subset of the data. Future iterations of this project will also explore applying these tools to our other approval plans; however, we have not yet decided on the approach or scope of this work.