ISSN 2324-7878 (online)
palrap.org

PaLA College & Research Division (CRD)

Pennsylvania Libraries: Research & Practice

Practice

Data Visualization for Collection Analysis

Using Tableau to Visualize an Academic Library Collection

Sylvia Orner

Sylvia Orner is an assistant professor and Collections and Resource Management Librarian at the University of Scranton's Weinberg Memorial Library, sylvia.orner@scranton.edu.

Due to the increasingly digital nature of library resources and collections, it is sometimes difficult to envision a library's unified holdings and to understand how they have changed over time. Conducting a collection analysis and applying data visualization techniques can be an excellent way to get a top-down view of the collection as a whole. This article outlines the author's process for a collection analysis of the Weinberg Memorial Library's entire catalog of print and electronic resources. It explores the rationale behind some key collection analysis decisions and discusses approaches for data extraction and clean-up as well as visualization using the Tableau software.

Introduction

As library collections expand into an increasingly digital presence, there is a disconnect between print and electronic collections. Because of the ways they are separately stored and accessed, it can be difficult to conceive of a single unified collection amongst the conglomerate of e-books, e-journals, print books, print journals and streaming media. This becomes an issue when it is time to examine the overall state of our collections. For example, one area where a collection may be deficient in print resources may have an overabundance of electronic resources, one subject may dominate a print collection but have poor representation in electronic format, or other subject areas may be absent or lacking in timely materials.

Using Tableau, a visual analytics platform, and data extracted from the library's ILS (Sierra) a snapshot of the Weinberg Memorial Library's catalog was created. It was an excellent way to get a top down view of what the unified collections look like. Additionally, the use of Tableau allows for the creation of visualizations that are easily sharable and accessible for faculty and administration outside of the library. At this point, it is important to distinguish between collection analysis and collection assessment. An analysis is simply a scientific, data-driven process. It is an accurate depiction of holdings by the numbers. It is by no means an assessment of the collection's appropriateness concerning the informational needs of students and faculty nor is it a model to determine a collection's diversity or well-

roundedness. A good analysis is necessary for assessment as it gives insights into past collection development practices and may help inform future purchasing decisions (Walton & Bunderson, 2020).

Literature Review

Both the need and importance of collection analysis has been well documented. As Johnson noted, "libraries continue to work within the confines of a budget" (2016, p.490). Therefore, a deeper understanding of collection development priorities over time and a solid grasp of library holdings across various subjects and formats are helpful when it comes time to allocate funds. Johnson further stresses the importance of bridging the cost benefits that can be gleaned with analysis to the cost effectiveness of assessment, treating the reflective act of analysis as opportunity to explore the questions that will serve as catalyst for assessment. (2016) As a Collections Management Librarian new to this particular collection, collection analysis was imperative for the author to gain a deeper understanding of what exactly was being managed and to see how collection development priorities may or may not have shifted over time.

Coupling analysis with data visualization methods has been examined and successfully executed by many librarians in many different capacities. Murphy (2013) and Datig and Whiting (2018) have found success using Tableau to track data on library programs, spaces, and services in order to inform future decision making. Finch and Flenner (2016) and Walton and Bunderson (2020) used collection analysis and visualization to identify and track budget and collection development trends over time, while Haren (2014), Eaton (2017), and Wissel and DeLuca (2018) all used visualization methods to explore sets and subsets of data within the library's catalog. Even though many of today's ILSs have more sophisticated analytic capabilities than their predecessors, they often cannot provide the context necessary to make the data truly valuable. As stated by Datig and Whiting, "if statistics are not used to provide context or factored into decision-making processes, then there is no purpose to taking them in the first place" (2018, p.6).

However, beginning a data driven analysis requires considerable thought and planning since "library data is inherently messy, especially as libraries change platforms, upgrade systems, and archive diverse data sets over a long period of time" (Murphy, 2013, p.466). This vast wealth of data must be collected, parsed, and crafted into something that is both useful and meaningful. Visualization of this data can help impart meaning effectively and efficiently. It can also provide additional insight that one couldn't get from a spreadsheet and make valuable connections between data and real life (Yau, 2013).

The task of visualization is not without its challenges. As previously noted, data can be messy, and harvesting and storing the data can be challenging. In order to conduct an analysis, it is essential to consider what tools will be used. Here, several factors can come into play.

One of the first considerations is the analyst's knowledge and comfort of current technologies. If they are comfortable with coding or using APIs, then building a web application like Eaton's SeeCollections can lead to a highly customizable mechanism for visualizing collection data. (Eaton, 2017). It also requires a significant amount of time and programing skill. SeeCollections, which was created by Eaton as a prototype data visualization tool for libraries, relied on the catalog vendor's API as well as Flask, a Python based micro-framework, to transform API data into visual displays (Eaton, 2017). If the created application relies on data from a vendor API, it may also include significant upkeep as vendors frequently make improvements to their systems. Alternatively, it may be rendered inoperable due to breaking changes in vendor API, as was the case with the SeeCollections (https://github.com/markeeaton/Primo-Visualization, last accessed December, 8th, 2022). If ease of use and reproducibility are more important, then visualization using simpler tools like Microsoft Excel or Tableau may be a better fit. Both Haren (2014) and Finch and Flenner (2016) found tools like these to be most useful in their collection analysis and mapping due to their flexibility and relatively low learning curve.

Whatever tools are used, graphically representing a library collection can yield information and insight that raw data alone could not. Murphy stated that it "gives libraries the ability to query, blend, explore, discover, and then analyze and present data in new and compelling ways" (2015, p.482). Furthermore, while Tableau and other visualization tools are powerful it is also important to understand how to present data in ways that provides appropriate context and complexity.

While simplicity can be important for user comprehension, Yau (2013) reminds us that clarity, above all, is paramount. If a data set is complex, then it may require complex visuals. However, it is the responsibility of the creator to ensure that this complexity is clearly conveyed and that the content is easily accessible to all. This may be as simple as making informed choices about color to ensure readability for users with color blindness or other visual impairments, or it may be more involved like critically thinking about how best to represent gaps or missing data. (Wong, 2010).

Because visualization can be such a powerful tool, it is equally important to understand what the data represents and how to present data in a way that is not misleading, intentionally or unintentionally. Cairo tells us that "no chart can ever capture reality in all its richness. However, a chart can be made worse or better depending on its ability to strike a balance between oversimplifying reality and obscuring it with too much detail" (2019, p.213). In order to effectively produce good visual data for collection analysis, it is important to put effort into the thoughtful collection of data and to have a deep understanding of what exactly that data represents and how best to present it in a way that is easily interpretable to the average user. With these things in mind, the next step is developing a methodology.

Methodology

This project consisted of two phases. The first included selecting and normalizing the data for analysis, and the second consisted of visualizing the data in Tableau.

Data Clean-Up

The first step for any data project is gathering and cleaning data. Because our library uses MARC records for all print and electronic materials, the data was gathered from all existing bibliographic records in Sierra. This was done using Sierra's "Create List" function and exporting selected data points to a .csv file. Since this constituted nearly 1 million records, only a few specific data points were ultimately utilized in this analysis. Because the purpose of this analysis was to gain a general understanding of the age, formats, and subjects held, the author chose Library of Congress call number, location code, item type, date of publication, and language as data points for extraction from the ILS.

After gathering the data, the next step was the data clean-up process. In order to conduct a successful analysis, it is important that the data be normalized in a way that creates uniformity but is also meaningful for the overall analysis. Data clean-up was conducted using a series of Excel functions and the Find and Replace feature. In retrospect, given the size of the dataset, it would have been more efficient to use a more powerful tool like OpenRefine or RStudio.

The first data point to be addressed was the call number. Since call numbers are unique, the field was truncated so that only the letter representing the LC class and subclass remained. This was accomplished using the LEFT function to return all characters to the left side of the string. This returned the first two characters of the call number string. Find and replace was then used to remove any remaining numerical values.

Depending on the intended audience for the final report, location codes and material types may not need any further cleaning. However, since the intention was to share this data with people who might not be familiar with that

type of data coding, they were converted back to natural language (for example, all material types of "a" were converted back to "Book"). See figures 1 and 2 for a comparison of the data before and after clean up.

	A	B	C	D	E
1	**Call #**	**Language**	**Location**	**Type**	**Date of Publication**
2	HC59.72.I55 T96 2006	eng	mcc	a	2006
3	S494.5.I5 E575 2007	eng	mcc	a	2007
4	HD3611 .D58 2006	eng	mcc	a	[2006]
5	HC59.7 .E295 2005	eng	me	s	2005?
6	HC59.72.C3 I54 2006	eng	mcc	a	2006
7	HJ7902 .S64 2008	eng	me	s	2008
8	HC59.7 .M36 2007	eng	me	s	2007
9	HC79.B38 G56 2008	eng	me	s	2008&x5d
10	HG195 .B355 2009	eng	me	s	Â©2009
11	HC79.E5 W48 2005	eng	mcos	a	2005
12	HC60 .G745 2005	eng	me	s	2005
13	RA643.86.A357 W67 2008	eng	mcc	a	2008-09
14	SD387.P74 F67 2008	eng	mcc	a	2008
15	HC244.Z9I52 U65 2008	eng	mcc	a	2008
16	HE8635 .F56 2005	eng	mcc	a	2005
17	RA395.D44 N34 2007	eng	me	s	Â©2007
18	RA643.86.A357 A35 2007	eng	me	s	Â©2007
19	HG3881.5.W57 A66 2008	eng	me	s	Â©2008
20	LA596 .T69 2007	eng	me	s	2007
21	HC59.7 .G765	eng	me	s	s.d
22	HD75 .G54 2006	eng	me	s	2006
23	LA1516 .E36 2005	eng	me	s	[2005]

Figure 1

Sample Data Set Prior to Clean Up

	A	B	C	D	E
1	**Call # (Clean)**	**Language**	**Location**	**Type**	**Date of Publication**
2	HC	eng	Circulating Collection	Book	2006
3	S	eng	Circulating Collection	Book	2007
4	HD	eng	Circulating Collection	Book	2006
5	HC	eng	Electonic Access	eBook	2005
6	HC	eng	Circulating Collection	Book	2006
7	HJ	eng	Electonic Access	eBook	2008
8	HC	eng	Electonic Access	eBook	2007
9	HC	eng	Electonic Access	eBook	2008
10	HG	eng	Electonic Access	eBook	2009
11	HC	eng	Oversized Stacks	Book	2005
12	HC	eng	Electonic Access	eBook	2005
13	RA	eng	Circulating Collection	Book	2008
14	SD	eng	Circulating Collection	Book	2008
15	HC	eng	Circulating Collection	Book	2008
16	HE	eng	Circulating Collection	Book	2005
17	RA	eng	Electonic Access	eBook	2007
18	RA	eng	Electonic Access	eBook	2007
19	HG	eng	Electonic Access	eBook	2008
20	LA	eng	Electonic Access	eBook	2007
21	HC	eng	Electonic Access	eBook	
22	HD	eng	Electonic Access	eBook	2006
23	LA	eng	Electonic Access	eBook	2005

Table 2

Sample Data Set After Clean Up

Date of publication proved to be the most challenging field to clean. At the start of the project, a decision was made to extract date of publication information from the MARC 264_c field. This seemed suitable as an option since the catalog had been retroactively converted to RDA format. However, given that some of the date of publication entries were inconsistently structured or structured under older cataloging rules, many data points needed to be cleaned as best as possible to provide an exact 4-digit date. Unfortunately, this was the least successful area of clean-up and resulted in approximately 400 unnormalized dates that could not be included in the final analysis. For future projects, extracting a date of publication from the fixed field information might be more viable.

During visualization, date of publication needed some further consideration when it became apparent that there were far too many unique dates for the type of big picture analysis that was intended. For that reason, it was decided to create an additional attribute for decade of publication. So, in future exploration of the data, 1990 was understood to represent any date of publication between 1990 and 1999. This binning of data may be considered problematic for the purpose of assessment. For instance, there may be a significant difference in information between a title published in 1990 and a title published in 1999, but perhaps not so much difference between a title published in 1999 and one published in 2000. However, since the purpose of this project was a top down analysis of the collection, it made sense to organize publication data by decade for the sake of ease.

Ultimately, cataloging is a human endeavor with processes and rules that have changed over time as well as systems that house and parse data in many different ways. For that reason, cataloging data may be subject to inconsistencies. In this case, 100% data normalization was not possible, but the results were close enough that the author was reasonably comfortable continuing.

Visualization

Once data clean-up was completed, the data tables were joined to a Tableau workbook where various data categories were manipulated into visualizations. In keeping with data visualization experts like Knaflic (2015) and Wong (2010), the author chose to employ charts and visuals that were simple and focused. The nice thing about Tableau is that visualizations are highly customizable and can be made interactable by easily adding filters and limiters. With this in mind, several filterable bar charts were created to highlight and explore different facets of the collection more easily.

For instance, the initial chart organized items by the number of titles per Library of Congress class and subclass. Of course, this is quite a large chart and not very easy to comprehend at a glance (see Figure 3). While it is easy enough to pare results down to the top 5 or 10 (see Figure 4), adding the ability to filter on things like location, material type, LC class, and date of publication can give a good snapshot of a particular scenario. For example, if one of the university's departments needs information on the titles available in their field for accreditation purposes, the data could easily be filtered on the relevant LC classes. Similarly, if the library wanted to see only materials of one item type (e-books or print books) or in a particular date of publication range, that is easily achievable with filter (see Figure 5).

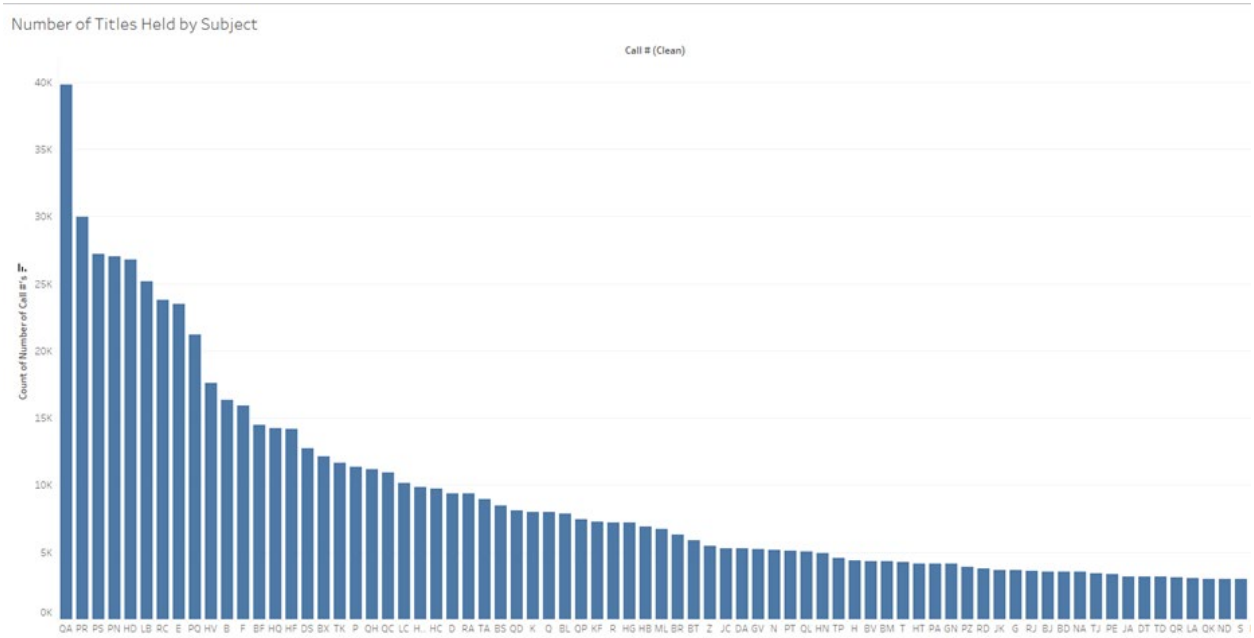

Figure 3

Number of Titles held by Subject (No Filters)

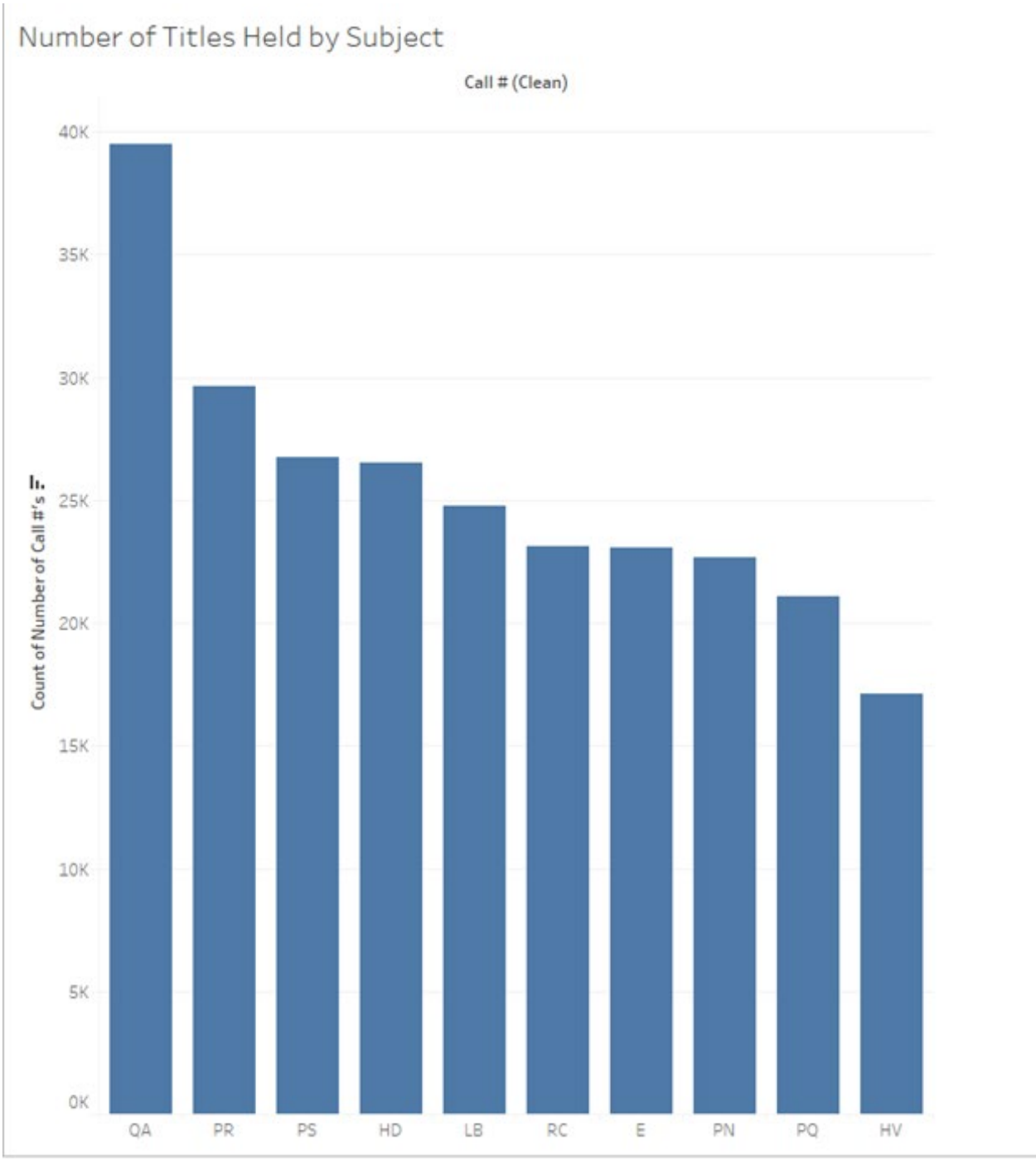

Figure 4

Number of titles Held by Subject (Filtered to Top Ten Responses)

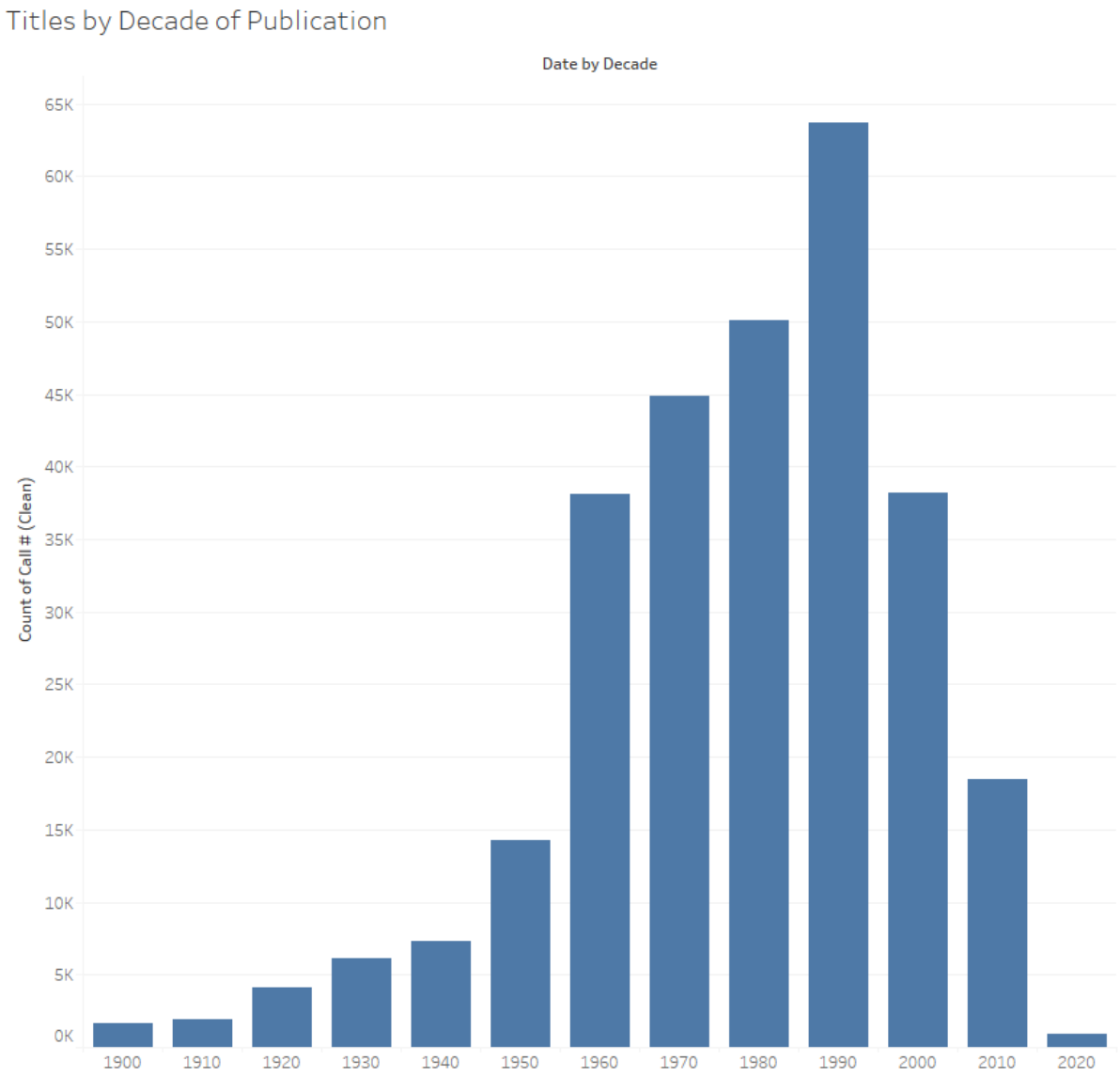

Figure 5

Titles by Decade of Publication (Filtered to Print Only)

More complex queries may necessitate more complex visualizations. For example, Figure 6 shows a snapshot of print materials in the Library of Congress class L (Education) organized by decade of publication. The bar represents the total number of L titles published in a particular decade while the different colors within the bar represent how many titles are present in each LC subclass (LA, LB, LC, etc.).

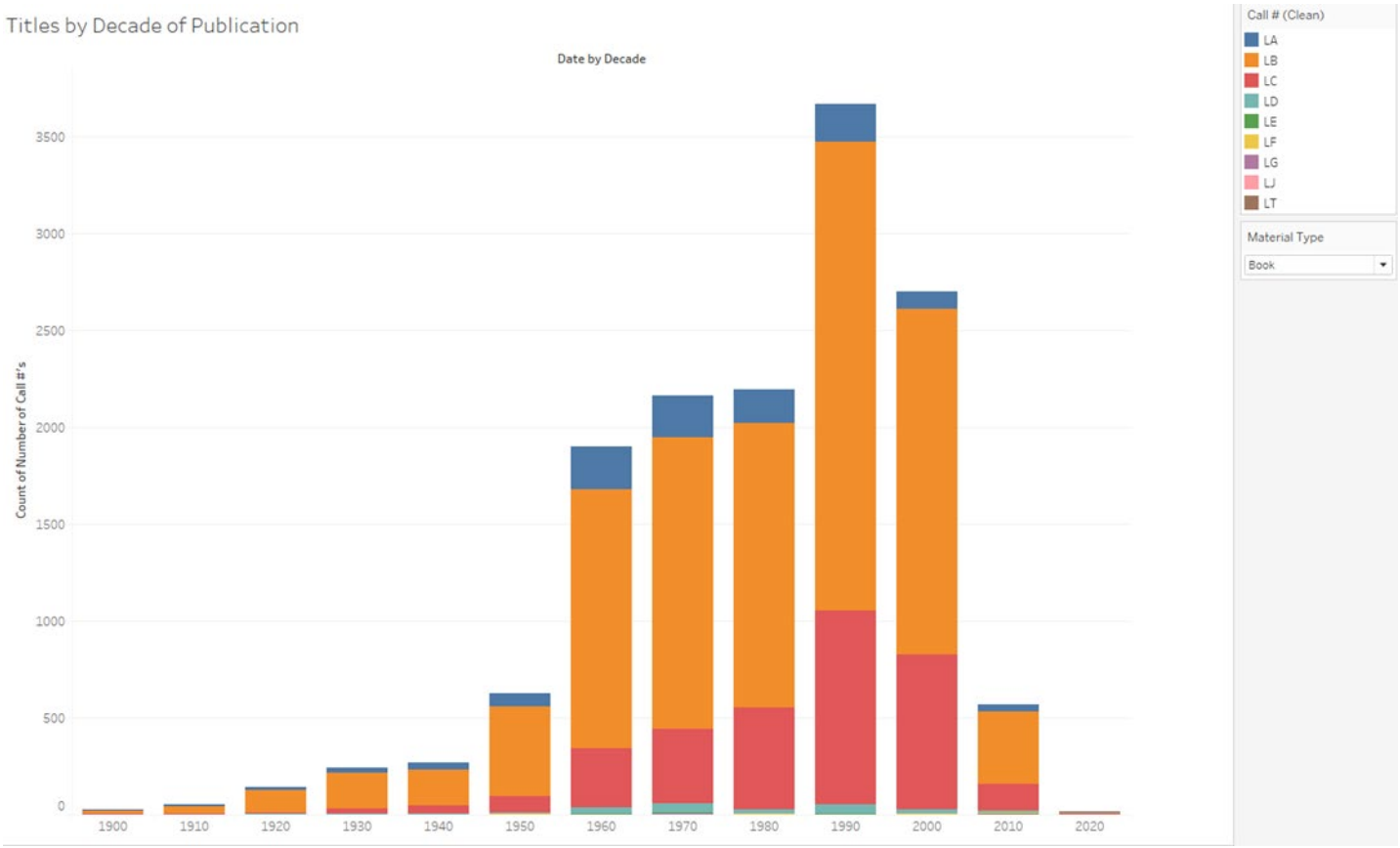

Figure 6

An Analysis of Class L Titles by Decade of Publication

Because of its flexibility, Tableau offers opportunities to easily view our data using different types of visualizations. It will even suggest different types of charts to use based on the number and types of attributes used in its rows and columns. This can offer a way to easily explore new and different ways of looking at the data. For example, in Figure 7, a bubble chart was used to explore which LC classes and subclasses in the collection's Spanish language material were most represented. Here, the data is filtered to include only Spanish language materials with each bubble representing a specific LC class or subclass. The larger the bubble and darker the color, the more titles are represented in a particular class or subclass.

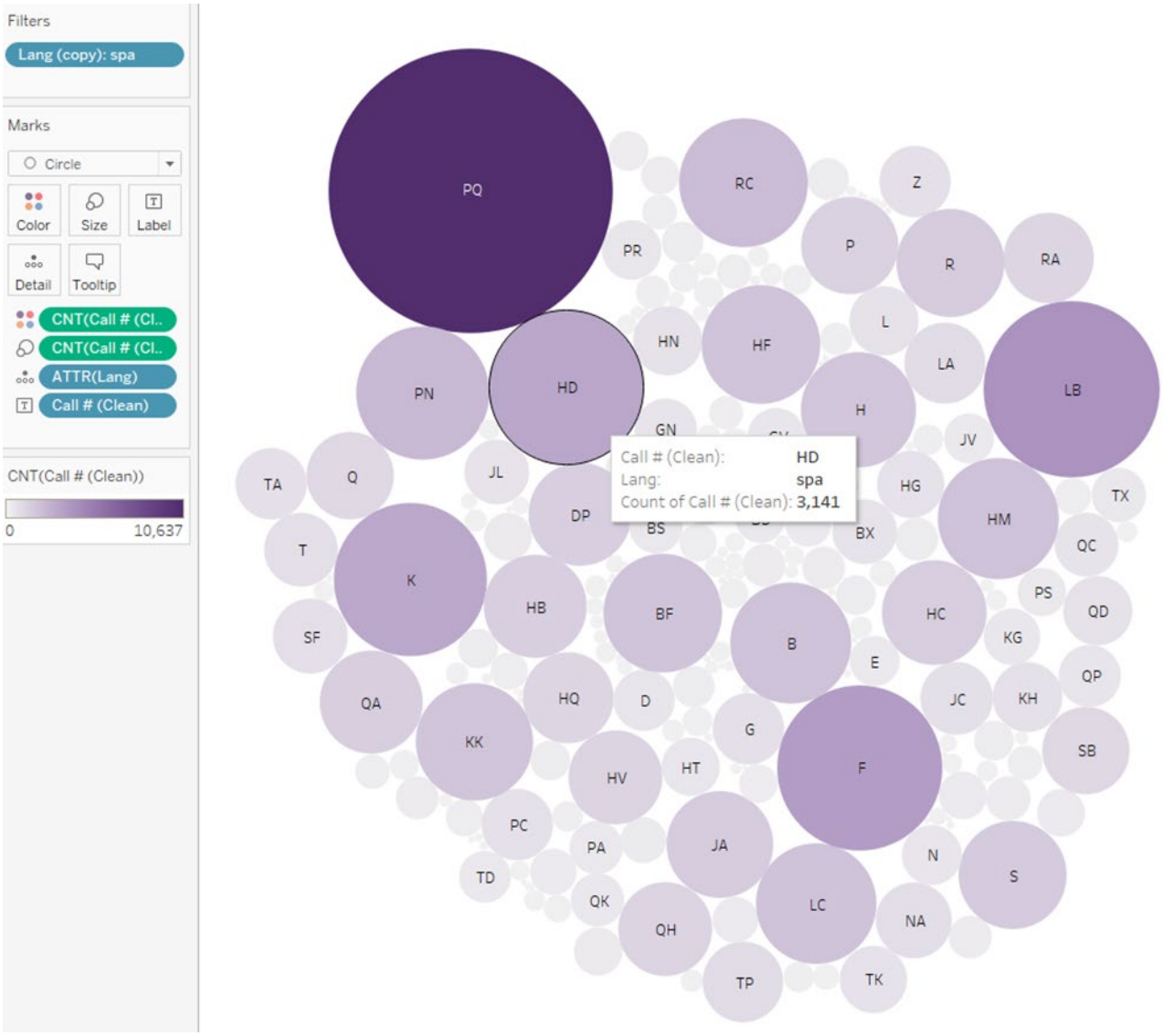

Figure 7

General Representation of Subject Areas in Spanish Language Material

Conclusion

After conducting an initial analysis, the findings were enlightening. Overall the collection leans heavily towards electronic resources, and overall subject representation leans slightly towards science and mathematics with literature and other humanities having more representation in print resources. Most of the print collection is from the 1980s-1990s, but this makes sense in light of collection development priorities slowly shifting from print to electronic materials in the mid-2000's with the majority of monograph purchases now being made in electronic format.

If a picture is worth a thousand words, then using data visualization for collection analysis can be an excellent way to effectively create graphical representations of collections for both informational and analytical purposes. This project allowed the author to not only gain some insight into what exactly lives in the library's vast collections, but it

has also provided an opportunity to explore the shifting priorities over time and identify any subject areas that may have been overlooked in recent years.

The next area of research is the identification of areas in the collection that may need further analysis and/or subsequent assessment, as well as continuing to reevaluate the analysis process to enhance accessibility and reproducibility. Ultimately, this workflow only provides a snapshot of the collection at a given moment in time. Given the time involved with data extraction and clean-up, it may prove worthwhile to explore ways in which the process could be more automated. This would allow more agility with the overall analysis process and help create a more dynamic picture.

References

Cairo, A. (2019). *How charts lie: Getting smarter about visual information.* W.W. Norton and Company.

Datig, I. & Whiting, P. (2018). Telling your library story: Tableau public for data visualization. *Library Hi Tech News, 25*(4), 6-8. https://doi.org/10.1108/LHTN-02-2018-0008

Eaton, M. (2017). Seeing library data: A prototype visualization application for librarians. *Journal of Web Librarianship*, *11*(1), 69-78. https://doi.org/10.1080/19322909.2016.1239236

Finch, J. L. & Flenner, A.R. (2016). Using data visualization to examine an academic library collection. *College and Research Libraries*, *77*(6), 765-778. https://doi.org/10.5860/crl.77.6.765

Haren, S. M. (2014). Data visualization as a tool for collection assessment: Mapping the Latin American studies collection at University of California, Riverside. *Library Collections, Acquisitions, and Technical Services*, *38*(3-4), 70-81. https://doi.org/10.1080/14649055.2015.1059219

Johnson, Q. (2016). Moving from analysis to assessment: Strategic assessment of library collections. *Journal of Library Administration*, *56*(4), 488-498. https://doi.org/10.1080/01930826.2016.1157425

Knaflic, C. N. (2015). *Storytelling with data: A data visualization guide for business professionals*. Wiley.

Murphy, S. A. (2013). Data visualization and rapid analytics: Applying Tableau desktop to support library decision making. *Journal of Web Librarianship, 7*(4), 465-476. https://doi.org/10.1080/19322909.2013.825148

Murphy, S. A. (2015). How data visualization supports academic library assessment: Three examples from the Ohio State Libraries using Tableau. *College and Research News, 76*(9), 482-486. https://doi.org/10.5860/crln.76.9.9379.

Walton, R. A. & Bunderson, J. (2020). Measuring the past to guide the future: Takeaways from a retrospective analysis on print books and ebooks. *Collection Management*, *46*(2), 80-90. https://doi.org/10.1080/01462679.2020.1841701

Wissel, K. M. & DeLuca, L. (2018). Telling the story of a collection with visualizations: A case study. *Collection Management*, *43*(4), 264-275. https://doi.org/10.1080/01462679.2018.1524319

Wong, D. M. (2010). *The Wall Street Journal guide to information graphics: The do's and don'ts of presenting data, facts, and figures*. W. W. Norton and Company.

Yau, N. (2013). *Data points: Visualization that means something.* Wiley.