Taming the Wilde: Collaborating with Expertise for Faster, Better, Smarter Collection Analysis

Jacqueline Bronicki
University of Houston, jbronicki@uh.edu

Cherie Turner
University of Houston, ckturner2@uh.edu

Shawn Vaillancourt
University of Houston, svaillancourt@uh.edu

Frederick Young
University of Houston, fyoung4@uh.edu

Jacqueline Bronicki, Cherie Turner, Shawn Vaillancourt, and Frederick Young, "Taming the Wilde: Collaborating with Expertise for Faster, Better, Smarter Collection Analysis" (2014). *Proceedings of the Charleston Library Conference.*
http://dx.doi.org/10.5703/1288284315581

Taming the Wilde: Collaborating With Expertise for Faster, Better, Smarter Collection Analysis

Jacqueline Bronicki, Collections Coordinator, University of Houston Libraries
Cherie Turner, Chemical Sciences Librarian, University of Houston Libraries
Shawn Vaillancourt, Education Librarian, University of Houston Libraries
Frederick Young, Systems Analyst, University of Houston Libraries

Abstract

The importance of collection assessment and evaluation has been a hot topic due to increasing budget restrictions and the need to prove worth to stakeholders through evidence-based evaluations. More robust collection analyses, like comparisons of holdings usage to ILL requests, and gap analyses, are increasingly embraced by the library community. Less thought, however, has been given to how to best conduct these analyses to ensure that the cleanest data is used and that the data tells the right story. The data to do these types of analyses often reside in complex systems and web-environments, which may not be fully understood by the collection managers or subject librarians. The University of Houston Libraries embarked on a large-scale gap analysis of the collection by subject area. The key component to success was quickly, accurately, and properly mining the data sources such as Sierra and the electronic resource management system. Our collection team contends that collaboration with expertise in the Resource Discovery Systems Department allowed the team to more quickly develop complete and accurate datasets, and helped to shape the analysis conducted. This paper discusses the challenges of defining project scope, the process of forming methodology, and the challenges of collecting the data. It will also review how experts were able to contribute to each step of this process. Finally it will outline some initial findings of the analysis, and how this research was accomplished in a realistic time frame.

Background

The University of Houston's MD Anderson Library serves a large student, faculty, and research population. The collection strives to support 12 academic colleges, an interdisciplinary Honors College, and a diverse offering of over 120 undergraduate majors. In order to ensure a consistent level of support for this broad university community some collection analysis was necessary.

The Collection Management Committee, which oversees collection development, recognized a need to better understand the content of the current collection and to identify any potential subject gaps in the collection. In September 2013 a project team was formed to define the research questions, develop appropriate methodology, and collect and analyze data to assess the breadth and coverage of both print and electronic resources.

In recent years the library has not undertaken any efforts of this scale to benchmark the print and electronic collection. The scope of this project required a large amount of data to be collected from the many systems that maintain our holdings and act as repositories for different formats. This large scale data collection proved to be especially time consuming and problematic for public services librarians who lack expertise in the systems that store the data, and made the assistance of experts necessary.

Methodology

The primary objective of this project was to determine the depth and relevance of current UH main campus library collection. The methodology was based on a print collection analysis done by Cornell University Libraries (2012), which provided a detailed benchmark for their print collection. Modifications were made in the University of Houston Libraries project to address differences in library services platform and electronic

http://dx.doi.org/10.5703/1288284315581

management systems, but the core methodology and many assumptions are identical. To determine gaps by subject area, Library of Congress call numbers were used as a proxy for subject.

Initially, the team wanted to include and benchmark all formats in the collection (print monographs, e-books, print journals, databases and electronic journals). However, based on a lack of call numbers for e-books and the multidisciplinary nature of databases, these two formats were excluded. This paper focuses solely on the phase of research dedicated to print monographs. The team chose to begin the analysis with this format with the expectation that monograph record data would provide the least complexity, the greatest opportunity for scalability, and a reusable model for other more complex formats.

Research Questions

The project team identified two main research questions to guide the research process and determine data collection variables:

1. What are the best measurements for evaluating the current scope of the collection?
2. What subject areas are not adequately covered in the current collection?

Population

In order to best answer our research questions and benchmark the collection at a point in time, data was collected for the entire population of print monograph records with valid LC call numbers. The Systems Analyst most familiar with the library services platform developed a query to generate a list of print monographs in the collection based on three parameters: a location code designating our campus from the others in the system, print monograph designation, and status indicating availability.

This data output resulted in over 1 million records (N=1,048,575) and represented the catalog as it existed at the time of parsing, January 31, 2014. The Systems Analyst provided the raw data in a .csv file. Records were removed during cleanup narrowing the final dataset to 889,825 records that met the criteria for inclusion and supported the research question.

Initial Findings

The analyses presented at this conference focused on distribution of monographs per LC class and subclass, shown as Figure 1, as well as a ratio calculation of percentage of holdings and usage between print monographs and ILL borrowing requests which is shown as Table 1. The ratio calculation, based on an analysis conducted by John Ochola at Baylor University (2003), was used to identify and flag potential “gaps” in the collection. The details of the ratio calculations and other analysis done on this dataset are explained in another forthcoming paper (Bronicki, Ke, Turner, & Vaillancourt, in press).

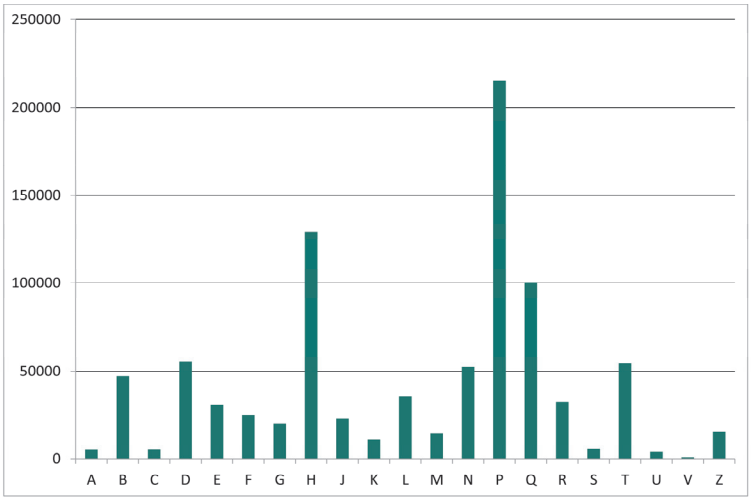

Figure 1. Distribution of monographs per call number.

LC Subclass	Percent of Holdings	Percent Usage	PEU	Holdings Usage	Percent of ILL Borrowing	RBH	ILL Usage	Action
B	1.32%	1.43%	1.08	Overused	0.79%	0.6	Underused	No Changes
BC	0.09%	0.08%	0.82	Underused	0.05%	0.51	Underused	Ease Off
BD	0.24%	0.20%	0.84	Underused	0.24%	1.01	Underused	Ease Off
BF	1.22%	1.78%	1.46	Overused	2.00%	1.64	Overused	Growth Opportunity
BH	0.07%	0.09%	1.29	Overused	0.05%	0.68	Underused	No Changes
BJ	0.22%	0.27%	1.21	Overused	0.18%	0.79	Underused	No Changes
BL	0.42%	0.65%	1.56	Overused	0.69%	1.65	Overused	Growth Opportunity
BM	0.10%	0.07%	0.67	Underused	0.09%	0.95	Underused	Ease Off
BP	0.13%	0.26%	1.95	Overused	0.34%	2.57	Overused	Growth Opportunity
BQ	0.04%	0.10%	2.63	Overused	0.32%	8.05	Overused	Growth Opportunity
BR	0.36%	0.33%	0.91	Underused	0.70%	1.96	Overused	Change Purchasing
BS	0.22%	0.16%	0.73	Underused	0.36%	1.62	Overused	Change Purchasing
BT	0.16%	0.13%	0.85	Underused	0.40%	2.53	Overused	Change Purchasing
BV	0.18%	0.15%	0.86	Underused	0.44%	2.49	Overused	Change Purchasing
BX	0.52%	0.29%	0.56	Underused	1.69%	3.23	Overused	Change Purchasing

Table 1. Holdings usage to interlibrary loan usage comparison for LC subclass B.

Challenges in the Development Phase

The team faced many challenges during the development of the methodology. Most of our initial struggles stemmed from the team's lack of expertise with the library services platform and interlibrary loan system. Records were, in many cases less complete than our inexperienced team anticipated, so priorities and expectations had to be revaluated. We had to scale our expectations about what the systems could deliver, and be cognizant of errors inherent in systems that require manual data entry, specifically the print monograph records. Developing a realistic strategy to collect this data became a major outcome of the research. In the end, collaboration with the Systems Analyst in our Resource Discovery Services department allowed us to collect the raw data in a timely and accurate manner and allowed us to focus more closely on data analysis. In many ways the process of collecting relevant and accurate data has been far more enlightening than the findings.

We presented the challenges from two viewpoints, the project team of the four public services librarians and the systems analyst doing the data mining in the library services platform.

Challenges from viewpoint of the Project Team:

1. The project team lacked necessary understanding of the infrastructure of the systems containing the data.
2. Defining input and output variables that could be reasonably and consistently parsed was more complicated than the team anticipated.
3. A deeper understanding of the MaRC record was necessary to ensure that the right data was gathered.
4. The scope of the project did not initially scale well with the timeline and the realities of obtaining this type of data.
5. Due to limitations of the library service platform, input errors in the records, and missing data in records, it was often a challenge to gather the most relevant and accurate data.

Challenges from viewpoint of the Systems Analyst:

1. The Systems Analyst was brought in after the project was underway, and after the research questions has been developed.

2. A recent migration from a Millennium ILS to the Sierra library services platform presented some benefits and challenges to gathering the requested data.
3. The scale of the proposed project did not match the types of data library system can provide.

Choosing Output Criteria

Once a method to obtain relevant and accurate data was finalized, some decisions could be made about what information to retrieve from our identified records. Our analysis needs, and therefore our output criteria, are shaped by the need for some way to meaningfully denote subject over this very large and broad collection. In addition, our research questions focus around two central ideas: scope of the collection, and adequate coverage. By researching the scope of the collection we hoped to understand what our current collection contains, and by researching adequate coverage we hoped to understand whether our collection is meeting user needs.

We initially planned to gather the fields shown in Table 2, which were selected by the project team with a very basic understanding of the data in our library's records.

Bibliographic Record	
	Call Number
	Subject Headings
	Publication/Copyright Date
	ISBN
	Record Number
	Title
Item Record	
	Copy Number
	Total Number of Checkouts
Order Record	
	Status
	Order Date

Table 2. Planned output criteria.

We planned to use the call numbers as the primary subject proxy, and if necessary utilize the subject headings to fill in any gaps. The publication date was intended to show how materials of different ages are collected and used. ISBN and the bib record number were each intended to be used as unique identifiers. The title was chosen for the benefit that it would provide to a deeper analysis within a smaller dataset, which we hoped would follow our analysis. The copy number would provide us with an understanding of how many copies of each title were feeding our usage data, and the total number of checkouts would serve as our proxy for use. Finally, we selected the status of the item to ensure that we were focusing our efforts on items that were in our circulating collection, and planned to use order date to get an idea of how the order date related with the item's age and the subject.

After having many conversations with the Systems Analyst and others with more expertise in our library's records and management systems our output criteria shifted to account for that new knowledge. In some cases we found that the fields we originally intended to capture were just not as meaningful as we had hoped. In others the data was unavailable, or not available in a form that would be useful to our analysis. The final output criteria, which are substantially different from those we began with, are displayed in Table 3.

Bibliographic Record	
	Call Number
	Publication/Copyright Date
	Record Number
	Title
	Publisher
	Catalog Date
	ISBN
Item Record	
	Call Number
	Total Checkouts
	Last Year Checkouts
	Year-to-Date Checkouts
	Location

Table 3. Final output criteria.

Subject headings, copy number, status, and order date from our final criteria. Subject headings were removed primarily because it would have been very complicated to shape the data in this field into a meaningful subject proxy. Copy number was found to be unnecessary because our final search was based on the item record rather than the bibliographic record as we originally planned. Status was also incorporated into our search

parameters, and order date was not found to add significantly to our analysis.

Several fields were added, including publisher, catalog date, item call number, year-to-date checkouts, and location. Publisher was added to allow for analysis based on publisher, which could be the basis of a future analysis by subject librarians. Catalog date was intended to serve as a more meaningful version of our prior order date category. The item call number was necessary to supplement the sometimes missing bibliographic call numbers. Last year and year-to-date checkouts were intended to provide an idea of what our current usage is like. Finally the location field provides information on where our collection is actually located and was intended to serve, in part, as a way to confirm circulating status.

Transforming the Data

Finally, with the data collection as complete as possible, our focus shifted to cleaning and normalizing the data. Once again we realized that we needed to consult with our experts to make sure we understood the data. In particular we needed clarification on how MaRC records were structured and used locally in order to ensure that we were accurately interpreting each field, and accounting for any gaps in the data based on our MaRC records.

Bibliographic record numbers initially seemed to be problematic because they appeared to be missing a digit, complicating the project team's plan to use them as unique identifiers. Fortunately the missing digit was found to be a random character which would not impact our use of the field as unique identifier. After noticing some strange numbers in our Last_yr_checkout field we sought help from our Systems Analyst for clarification, and discovered that this field was not in use and therefore that data was meaningless. This was a particularly significant discovery as we would have used this data for analysis, perhaps assuming that these strange numbers were just outliers or corrupted records. Having a stronger understanding of the data allowed us to move forward with an appropriate view of what needed to be fixed or modified to be usable.

Cleaning consisted of identifying potential issues in the record and then verifying that these issues did not cause further problems in other parts of the data for a given record. For example, if diacritics were used in the record, they would often not parse out correctly when the records were downloaded and delineated. Therefore, we checked records that had these in the title to make sure that other parts of the record were not corrupted as a result. The most common issues were titles with diacritics, dates that had not been sorted into the correct fields, and items with future cataloging or publication dates, presumed to be data entry errors. For the very few records where errors were detected, we manually checked the record and corrected the errors.

We also examined the ISBNs at this stage. The original plan had been to use them as a unique identifier for items; however, we found that there was so much extraneous information included in the field that cleaning it would have been more work than it was worth. Instead, we deleted the column and solely used bib record numbers as our identifier.

Larger issues in the data cleaning phase required the removal of larger chunks of data that were likely to become problematic in our analysis. By reviewing the location code information we were able to identify 142,823 records that should be removed. Some of these records were items located at other campuses that were not removed by the parameters of our initial search, but most were:

- Government documents cataloged using SUCOC numbers instead of LC call numbers, making analysis by subject impossible.
- Theses and dissertations cataloged with local numbers which identify the college in which the document was written rather than the subject.
- Microform materials not cataloged with an LC call number.

An additional 14,894 records were removed based on a review of the call number data. Most commonly this was because a call number was not

available from the fields which were accessed during our data export phase. We also found additional government documents, former reserve materials, and dissertations that were not in the usual locations for these materials and did not have a valid LC call number for us to use.

Because not all bibliographic records included a call number, we opted to pull call numbers from both the bibliographic record and the item record. In most cases an item would have only one of the two, or would have two matching call numbers; however, in some cases both call number fields included data but the call numbers did not match. For records with non-matching call numbers it was decided that the item call number should be used as this would be the call number with which students were most likely to interact. Once this cleaning was completed the analysis could be done.

Planning for Analysis

Given our clean data set of meaningful variables we did a wide range of analyses. Benchmarking analyses included the distribution of our collection included as Figure 1, more detailed LC subclass distributions, analyses of usage by call number and by age, and analyses of age by subject. To learn more about how our collection fits with user needs we used our usage data alongside our interlibrary loan data from a two year span. A subset of the findings of this analysis is shown in Table 1. While our analysis and our findings are not the focus of this presentation, these do help to provide some context for our analysis. Our analytical plan and our findings are closely linked to our data collection strategies.

Future Work

While our findings have been interesting, and promise to help us improve our collection, the project team has found the process of data gathering and of preparing for analysis the most enlightening. For the project team there have been many important lessons.

Lessons Learned

1. Collaborating with experts is essential to gathering data that has meaning.
2. Local practice shapes what data is available and how it can be used.
3. A deep understanding of the infrastructure of our systems is needed to effectively collect data.

The project team is currently working on a related analysis of print and electronic serials, which is encountering some of the same challenges experienced in our print monographs analysis. It is also encountering some new and exciting challenges, like adjusting to the many different systems in which the needed data is stored, understanding what information is available in a serial record, finding meaningful ways to show holdings, and deciding whether to deal with aggregated content as opposed to our own subscriptions.

References

Bronicki, J., Ke, I., Turner, C. Vaillancourt, S. (in press). Gap analysis by subject area of the University of Houston Main Campus Library Collection. *Serials Librarian.*

Cornell University Library. (2012, November 22). *Report of the Collection Development Executive Committee Task Force on print collection usage.* Retrieved from http://staffweb.edu/system/files/CollectionUsageTF_ReportFinal11-22-10.pdf

Ochola, J. H. (2003). Use of circulation statistics and interlibrary loan data in collection management. *Collection Management, 27*(1), 1-13. http://dx.doi.org/10.1300/J105v27n01_01