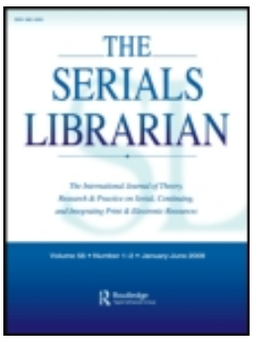

The Serials Librarian

From the Printed Page to the Digital Age

ISSN: 0361-526X (Print) 1541-1095 (Online) Journal homepage: http://www.tandfonline.com/loi/wser20

Gap Analysis by Subject Area of the University of Houston Main Campus Library Collection

Jackie Bronicki, Irene Ke, Cherie Turner & Shawn Vaillancourt

To cite this article: Jackie Bronicki, Irene Ke, Cherie Turner & Shawn Vaillancourt (2015) Gap Analysis by Subject Area of the University of Houston Main Campus Library Collection, The Serials Librarian, 68:1-4, 230-242, DOI: 10.1080/0361526X.2015.1017717

To link to this article: http://dx.doi.org/10.1080/0361526X.2015.1017717

The Serials Librarian, 68:230–242, 2015
Published with license by Taylor & Francis
ISSN: 0361-526X print/1541-1095 online
DOI: 10.1080/0361526X.2015.1017717

Routledge
Taylor & Francis Group

Gap Analysis by Subject Area of the University of Houston Main Campus Library Collection

JACKIE BRONICKI, IRENE KE, CHERIE TURNER, and SHAWN VAILLANCOURT
Presenters

This article outlines preliminary results of a complete gap analysis of the collections at the University of Houston Libraries. Methodology for a large-scale collection analysis project is explored here, including necessary collaborators for successful and accurate data collection. As well, the preliminary results of phase one of the project, studying usage by subject area and comparing to available interlibrary loan data, are outlined. Limitations and future directions for the project are also discussed.

KEYWORDS gap analysis, usage data, interlibrary loan, print monographs, project management

INTRODUCTION

The University of Houston (UH) Main Campus Library collection is a large, multiformat, and evolving research collection that supports the teaching and research needs of the entire campus. Located at the center of the City of Houston, the university is one of three tier-one public research universities in the state of Texas. It has 13 colleges and schools, currently offering over 120 undergraduate majors, close to 200 graduate and professional degree programs, and 40 research centers.[1] The Main Campus Library has built a collection exceeding 2.5 million items to meet the ever-evolving teaching and research needs on campus. As the collection continues to grow and library personnel change, it becomes more and more challenging for current selectors to gain a good grasp of our collection's strengths and weaknesses. At the same time, librarians at UH are working in a campus environment that demands evidence-based practices and data-driven decision making. In order

to ensure that the collection successfully meets the campus's research and teaching needs, the collections unit at the library decided to embark on a large-scale collection assessment project to gather collection data for future planning and purchase decisions.

A project team of four librarians was formed and tasked with designing and developing a high-level collection assessment project to assess the breadth and coverage of both print and electronic resources at the University of Houston M.D. Anderson Library. The team focused on developing methodology to gain an understanding of the current collection level across all subjects in the Library of Congress (LC) classification system, and to identify collection gaps through analysis of circulation and interlibrary loan data. Therefore, all data, including collection, circulation, and interlibrary loan data, was analyzed at the subject level.

Because of the scale of this assessment project, the project was carried out in two phases. The first phase was focused on print monographs in the library collection, whereas the second phase will focus on journals and serials. The first phrase, print monographs, was completed in the spring of 2014. This article reports on the methods and results of the first phase of this project. It will describe development of a research matrix, identified data sources, and our data analysis strategies. It also presents the initial findings, how the results can inform our collection decisions, and future areas of research.

METHODOLOGY

The main objective of this project was to determine the depth and relevance of the current UH Main Campus Library collection. The collection could then be validated against usage and Interlibrary Loan data as a way of determining if the needs of patrons are being met and if any gaps by subject area currently exist in the collection. The team used LC call numbers as the main proxy for subjects to allow for descriptive statistical analyses such as frequency counts and trends across time in each subject area.[2] For research purposes, the team determined that the first letter of the 21 main LC subject classifications would be used for initial subject analysis, followed by a further in-depth analysis of the full alphabetical subclass. To this end, call number data would be captured from ILLiad and Sierra, the integrated library system (ILS) currently used at UH.

A research matrix was developed (see Table 1) to identify the data sources that would best answer our research question, as well as determine the most appropriate type of analysis to undertake for each resource type. This matrix also served to keep us from deviating too far from our original purpose and give us something to refer back to whenever our project scope was slipping. Recognizing the unique nature of different resource

TABLE 1 Research Matrix

Resource type	What do we want to find out? (What is the research question?)	Why we are asking this question? (How the answers to this question can help us in practice?)	Where can we get the answers to these questions?		What types of analyses are required in order to get the answers? (Learning new skills, analytical tools, etc.?)
			Raw data	Source	
Monographs	The team would like to determine completeness of coverage per subject area in order to identify possible gaps in coverage.	1. To help make future purchasing decisions 2. To benchmark the collection for future analysis 3. To determine if we are meeting the needs of our patron population	1. Number of titles per subject 2. Circulation counts per subject 3. Acquisition counts for time range per subject 4. ILL requests for time range per subject	1. Sierra/Catalog 2. Sierra/Catalog 3. Sierra/YBP—Gobi ordering system 4. Iliad	1. Systems Analyst 2. SPSS to analyze results 3. Understanding of JR1 structure and definitions 4. Excel—Basic Analytical Functions for descriptive statistics
Serials	The team would like to determine completeness of coverage per subject area in order to identify possible gaps in coverage.	1. To help make future purchasing decisions 2. To benchmark the collection for future analysis 3. To determine if we are meeting the needs of our patron population	1. Number of titles per subject 2. Circulation counts per subject a. Limited circulation counts—print b. Usage reports—e-journals 3. Acquisition counts for time range per subject 4. ILL requests for time range per subject	1. Sierra/Catalog 2. Sierra/Catalog a. Sierra b. Usage reports—JR1 3. Sierra/EBSCONet 4. Iliad	1. Systems Analyst 2. SPSS to analyze results 3. Understanding of JR1 structure and definitions 4. Excel—Basic Analytical Functions for descriptive statistics

types, the matrix was divided into two groups: monographs and serials. The team identified raw data needs and sources for each resource type. The ILS and Illiad were identified as the main data sources for monographs. The ILS, EBSCONet, ILLiad, and Project COUNTER (Counting Online Usage of Networked Electronic Resources) were identified as the data sources for serials, both print and electronic.

SAMPLING

Each format type was evaluated for inclusion into the sampling pool, including: electronic monographs (e-books), print monographs, electronic serials, and print serials. All other format types were excluded from the sampling pool. Due to lack of Library of Congress call numbers for a large portion of the e-books, the project team later decided to not include e-books in the sample or in phase one. The data for this phase of the project, print monographs, was mined from Sierra by a systems analyst using a Structured Query Language (SQL) script to parse the database. The project team carefully outlined the input criteria for inclusion based on codes from the ILS, with the understanding that the sample should not be restrictive at the raw data collection stage. We wanted to start with the widest possible sample of records and filter down, with the understanding that cleaning and deleting records could be done using output variables. Therefore, the team purposely used only a few parameters to create the sample for fear that records would be lost for various reasons (they were miscataloged, not parsed correctly, etc.). Recorded output variables were also predefined and used in conjunction with call number to substantiate that each record truly met criteria for inclusion. For version control, the team implemented a naming convention for the data files as the data transformed from raw to process through cleaning and review.

DATA COLLECTION AND FINAL SAMPLE

The UH Libraries' ILS was first parsed for the print monograph raw data for this phase of the project. The team determined that print monographs would provide the cleanest and easiest data export to understand how the raw data would appear. Monograph record data would also provide the least complexity, an opportunity for scalability, and help shape further processes for cleaning the raw data. Also, students at the main campus have access to collections from other campuses via a delivery service as our catalog is shared with several campuses among the UH System. However, we made a conscious decision to focus our research only on the UH main campus collection because the purpose of the study was to inform our main campus collection development and we have no influence on the collections at other campuses.

TABLE 2 Input Criteria for Study Inclusion and Output Variables for Analysis

Format type	Input criteria	Output variables	Types of analysis
Print monographs	bcode=a location=UH Main status= (-)	1. Title 2. Call Number 3. Publication Year 4. Record # 5. Publisher/Vendor 6. Catalog Date 7. ISBN	1. Last 2 years circulation data 2. Last 2 years of ILL data
Print serials	bcode=s location=UH Main status= (-)	1. Title 2. Call Number 3. Coverage (Catalog) 4. Record # 5. Publisher/Vendor 6. Catalog Date 7. ISSN 8. Subject Headings	1. Last 2 years circulation data (if any) 2. Last 2 years of ILL data
Electronic serials	bcode=3 location=UH Main status= (-)	1. Title 2. Call Number 3. Coverage (Serials Solutions, SFX) 4. Record # 5. Publisher/Vendor 6. Catalog Date 7. ISSN 8. Aggregator	1. Last 2 years of Counter usage statistics 2. Last 2 years of ILL data

Therefore the collection data of all other UH campuses were excluded with the parameters implemented. As shown in Table 2, the systems analyst ran a script in Sierra to pull all records that met the three parameters: bcode = a (print monograph), location = UH main campus, and status = In use (-). 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. The project team then reviewed the raw data for each output variable and assigned outlier criteria to flag records for review and subsequent removal from the sample. Records had to be removed for most government documents and dissertations because of a lack of proper LC call numbers. There were also a number of records that did not have any call numbers at all. After review of data for each variable, the project team included 889,825 records in the final print monograph sample that we felt met the criteria for inclusion and supported the research question. The team is following this process for the print and electronic serials in the next phase of the study scheduled for early May 2014.

Interlibrary loan (ILL) data provided interesting challenges in terms of data collection by call number. The research team was interested in only the ILL borrowed requests, those requests by UH patrons for material from

another institution. We felt these requests from UH patrons for materials outside of our library provided insight into call number ranges/subjects lacking full collection development, thus representing a potential gap in collection coverage. The Information and Access Services Department, where ILL is handled, collaborated with the project team to supply the raw data. The initial research into the structure of the reports provided by ILLiad showed a lack of call number information for borrowing requests from UH patrons. However, the ILL department discovered two reports for UH specific requests, OCLC Online Computer Library Center, Inc. (OCLC) Usage Statistics and WorldCat Resource Share, which provided call numbers for the borrowed items requested. A time parameter of January 1, 2012 through December 31, 2013 was used to generate a borrowed request report and only completed requests were included in the final data. The report was then filtered to include just print books to map to the call number ranges for print monographs. The OCLC record number is included in this report and provides us with a unique identifier for further statistical analyses.

RESULTS

A wide variety of analyses were conducted on the collected and cleaned data, particularly focused on giving us a view of differences between LC classes and subclasses. A major goal of our project was to capture a complete picture of our current collection, and in order to do so we looked at the distribution of our monograph collection by LC class and subclass, and the age of our collection (Figure 1).

The largest areas of our collection are in language and literature (P), social science (H), and science and technology (Q and T). The average age of our collection, shown by LC class in Figure 2, ranges from 1965 (auxiliary sciences of history) to 1985 (medicine).

A second major goal was to identify gaps in our collection. We related this to usage by taking the usage for each LC class, and comparing to our holdings for each LC class, as is shown in Figure 3. Several call number

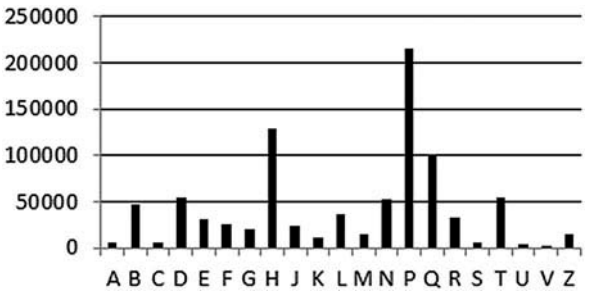

FIGURE 1 Total number of monographs by call number.

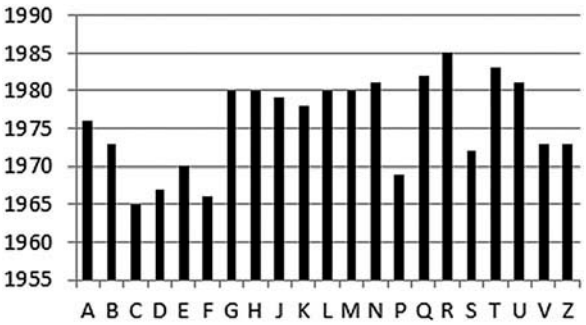

FIGURE 2 Average age of monographs in each LC class.

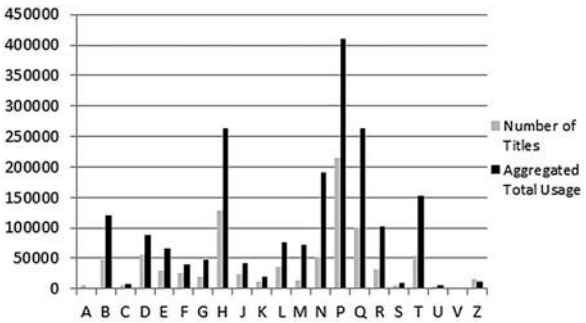

FIGURE 3 Usage of items in each LC class and number of titles in each LC class.

ranges show large amounts of usage, including the largest areas of our collection: language and literature (P), social science (H), and science and technology (Q and T). Some smaller areas of the collection are also notable, including fine arts (N), philosophy, psychology, and religion (B), medicine (R), and music (M).

While examining usage, we took the opportunity to look at overall usage across the entire collection. We found that of our 889,825 item monograph collection: 48%, or 425,865 titles, have never circulated; 88%, or 787,590 titles, have circulated five or fewer times; and 97%, or 861,910 titles, have not circulated in the last year.

Next, we compared the ages of the items in our collection that are being used, including both total circulations, and total year-to-date (YTD) circulations (Table 3). In particular within the YTD numbers, representing usage within the past year from the download date, we see that there is preferential use of newer items. Even including all circulations, we find that of all items that have never circulated, 76% were published prior to 1991.

TABLE 3 Usage of Our Collection by Age

		Total circulations						Total YTD circulations					
		≤0	1–1	2–5	6–50	51–100	101+	≤0	1–1	2–5	6–50	51–100	101+
Publication year	≤ 1950	19%	10%	6%	3%	1%	0%	13%	3%	2%	1%	0%	0%
	1951–1970	25%	20%	15%	10%	5%	5%	20%	8%	4%	4%	0%	0%
	1971–1990	32%	34%	35%	35%	29%	13%	34%	20%	13%	10%	0%	0%
	1991–2000	9%	16%	25%	39%	52%	42%	17%	24%	21%	14%	0%	33%
	2001–2005	6%	10%	12%	11%	7%	13%	8%	17%	19%	17%	0%	0%
	2006+	10%	11%	8%	3%	6%	28%	8%	28%	41%	54%	100%	67%
Total items		422,843	156,129	204,948	101,355	622	160	858,178	24,503	3,211	158	4	3

$$PEU = \frac{Percent\ Usage}{Percent\ of\ Holdings}$$

FIGURE 4 PEU calculation.

$$RBH = \frac{Percent\ of\ ILL\ Borrowing}{Percent\ of\ Holdings}$$

FIGURE 5 RBH calculation.

Finally, we analyzed use of our collection alongside our ILL borrowing.[3] From our collected data, we calculated the percentage of the collection in each LC class or subclass, or the percent of holdings. Similarly we calculated the percentage of our total usage for each LC class or subclass, or the percent usage. We then calculated a ratio of usage to holdings, percentage expected use (PEU), with those numbers (Figure 4).[4]

In an ideal situation we would assume that each item in the collection is used with equal frequency, so we can say that the percent usage should be equal to the percent of holdings, making the PEU equal to 1. So, when the actual percent usage is greater, the PEU is greater than 1, and we can say that that class or subclass is comparatively overused. Alternatively, if the PEU is less than 1 we can say that that class is underused.

From our collected ILL data, we calculated a percent of ILL borrowing, which is the percentage of our total ILL borrowing for each LC class or subclass. We then calculated the ratio of borrowing to holdings (RBH) (Figure 5).[5]

In this case, we cannot reasonably expect the percent of ILL borrowing to be equal to the percent of holdings, since it does not relate to our own collection. Instead, we compared the RBH for each class or subclass to the mean RBH for our collection. Our mean RBH is equal to 1.54 ± 5.18, indicating a lot of variation in ILL usage across different LC classes and subclasses. When the RBH is greater than 1.54 we can say that ILL is overused for that range, and when it is less than 1.54 it is underused (Table 4).

By investigating use of our holdings, and use of ILL for each class or subclass, we now have a large-scale view of user demand in that area. If our collection is overused, but ILL is underused then we know that our collection is meeting most needs. If our collection is overused and ILL is overused for a call number range, then we have a demonstrated demand and might consider purchasing more in that area. If our collection is underused and ILL is underused then we can assume that there is little demand in that area. Finally, if our collection is underused, and ILL is overused, then we may be collecting the wrong items or need to update our existing resources.

TABLE 4 Detailed Analysis of LC Class B, Comparing Holdings Usage to ILL Usage

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	0.60	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

DISCUSSION

Having completed phase one of this project, we have a much better understanding of our print monograph collection. We now have a clear picture of the subject distribution across our print monographs, and with our analysis we have learned several important things. We learned that our print monograph collection is relatively aged. This may be because we have never had to weed our collection. Also, in the past the Libraries have enjoyed cheaper journal costs and better monograph funding so the investment in print monographs 20 or 30 years ago could be assumed to have been better than more recent allocations. Further investigations into our acquisition data could help us determine if this is true. Another factor that might have skewed the age analysis is that we had to exclude the e-book collection from this study. This is primarily because our catalog records for e-books do not contain LC call numbers, so many of our recent science related collections have been excluded. We will need to identify an alternative way to incorporate the e-book collection data to develop a more complete picture of the age of our monograph collection.

Based on numbers, our largest collection areas are in language and literature (P), social science (H), and science and technology (Q and T), which mirrored the subject areas that have the largest amount of usage. Thus, overall, our collection matches the demands when considering it at the highest LC subject classification level. However, we also found that close to half of our materials have never circulated after their addition to the collection. Although it is heartening to know that newer materials have higher circulation rates than our older materials, which means we are at least buying more of the items our patrons want. At our library, we have started a small scale patron-driven acquisitions (PDA) model. It might be worthwhile to look at a more granular data level to see if and how we would like to expand the PDA plan and possibly improve this rate of usage.

When the collection data was compared with the circulation and interlibrary loan data, we identified subject areas that require more attention and demand probable investment. These results also indicate the need for a much more granular level study of our collection and user behavior. For example, we found that the BF classification area has a higher than average circulation rate and more interlibrary loan requests. In order to strengthen the collection in the area to meet the user needs, the subject librarian needs to learn what specific topics and types of monographs are in demand. To get these answers, selectors will need to expand upon this study and further investigate the circulation and interlibrary loan data at a granular level.

We also need to collect data and perform analyses looking at the patron status (i.e., faculty, graduate or undergraduate, etc.) of our users to understand differences in the behaviors of these groups. In addition, it may be necessary to combine this research with other types of collection analyses

(e.g., citation analysis based on research outputs) and user studies. This would allow us to form an action plan for collection development, as well as have a greater understanding of the nature of our collection usage.

In addition, we have learned some essential components for making a project of this nature succeed. We cannot overstate the importance of having an ILS expert, our System Specialist, on board. We relied on our specialist to download the entire catalog data set for us. At the beginning, our research team had tried to harvest catalog records from our ILS without assistance. However, after much confusion and fruitless effort, we realized that we could not be totally confident in the data we had gathered. We finally decided that we needed the person with unique expertise in using, accessing, and retrieving information from this system. The same situation arose with interlibrary loan request data. It was our Assistant Head of Access Services and ILLiad expert who downloaded and configured the ILL data in a useful format that we could work with. For a project of this nature, it is essential to work collaboratively. Publicize the work and seek out the colleagues with the expertise needed. Letting them know what you would like to achieve can not only help speed up the process, but also ensure the work is done accurately with the right options.

We also learned that data cleaning requires a strategic examination of the purpose and scope of the study; the decision of how to clean and what to exclude are dependent on the research goals and scope of a project. We made data exclusion decisions based on the nature of the data available and how relevant they were to our ultimate goals. Those decisions in turn have an impact on the interpretation of results. For us, LC call numbers were the essential data point for analysis. We therefore excluded all formats that did not have LC call numbers, including e-books, government documents, theses and dissertations, microfilm and microfiche. We considered the lack of LC call numbers in e-book records a significant deficiency of the monograph analysis because we have been purchasing a considerable amount of e-books in recent years. Therefore, we need to identify other data sources for this part of the study.

Since one of our research goals was to get a big picture of the current state of the collection, we obviously still have a lot of work ahead of us. Phase one completed the analysis of print monographs across subject areas. Our next task is to tackle the serials collection, both print and electronic. While the information we have discovered about the collection thus far has been interesting, these preliminary results cannot be the sole basis of future collection decisions. We are pleased that we are getting close to having a better understanding of our collection, but more work at deeper, more granular levels needs to be done in order to gain a complete understanding of the collection. This has been an essential project to begin developing an accurate overview from which we can derive further research. Given the

greater focus on accountability in higher education, this work can lead to measureable ways to ensure our collection money is well spent.

NOTES

1. University of Houston. "UH at a Glance," accessed June 19, 2014, http://www.uh.edu/about/uh-glance/.

2. Much of this analytical approach and structure was heavily influenced by the study at the Cornell University Libraries and it can be seen clearly in our results reporting that our types of analysis have been brought as possible in line with their reporting as our data allowed. Rich Entlich et al., "Report of the Collection Development Executive Committee Task Force on Print Collection Usage Cornell University Library" (Ithaca, NY: Cornell University, 2010), accessed April 3, 2014, http://staffweb.library.cornell.edu/system/files/CollectionUsageTF_ReportFinal11-22-10.pdf.

3. John H. Oocha, "Use of Circulation Statistics and Interlibrary Loan Data in Collection Management," *Collection Management* 27, no. 1 (2003): 1–13, doi:10.1300/J105v27n01_01.

4. Terry R. Mills, "The University of Illinois Film Center Collection Use Study" (Urbana, IL: University of Illinois, 1982), accessed April 16, 2014, http://files.eric.ed.gov/fulltext/ED227821.pdf.

5. William Aguilar, "The Application of Relative Use and Interlibrary Demands in Collection Development," *Collection Management* 8, no. 1 (1986): 15–24, doi:10.1300/J105v08n01_02.

CONTRIBUTOR NOTES

Jackie Bronicki is Collections and Online Resources Coordinator, University of Houston, Houston, Texas.

Irene Ke is Psychology and Social Work Librarian, University of Houston, Houston, Texas.

Cherie Turner is Chemistry Librarian, University of Houston, Houston, Texas.

Shawn Vaillancourt is Education Librarian, University of Houston, Houston, Texas.