Using Data Visualization to Examine an Academic Library Collection

Jannette L. Finch and Angela R. Flenner

The authors generated data visualizations to compare sections of the library book collection, expenditures in those areas, student enrollment in majors and minors, and number of courses. The visualizations resulting from the entered data provide an excellent starting point for conversations about possible imbalances in the collection and point to areas that are either more developed or less developed than is needed to support the major and minor areas of study at the university. The methodology used should offer a template to follow for others wishing to examine their collection and may prove valuable for adjusting expenditures, suggesting service opportunities or for marketing pieces of the collection that had been hidden before graphical analysis.

Gathering and displaying data in visual representations helps inform the brain faster and more effectively than reading textual lines of information. "One picture is worth a thousand words" is the cliché we use to describe this phenomenon. The classic works of both Edward Tufte and of Information Science professor and scientist Katy Börner provide beautiful examples of what excellent design principles applied to information and data may graphically reveal.

Visualizations provide "overviews about general patterns and trends" and allow discovery of "hidden structures."[1]

Edward Tufte, professor emeritus of Yale University and a well-known advocate and creator of elegant graphical display of complex data, explains that "graphics reveal data." Tufte asserts that the most "effective way to describe, explore, and summarize a set of numbers is to look at pictures of those numbers."[2]

Used in libraries, information gathered into graphical impressions can reveal patterns hidden in lines of text. Xu et al. remind us that, "[i]n the context of large-scale and heterogeneous collections, the different layers of information cannot be easily comprehended if presented linearly and sequentially, and there is a risk of getting buried in details or lost in generalities."[3]

Visualizations of library data have been used to:
- reveal relationships among subject areas for users.
- illuminate circulation patterns.
- suggest titles for weeding.
- analyze citations and map scholarly communications.

Future emphasis, as suggested by Eden, could be in replicating whole libraries in 3D printouts, making predictions of growth and space easier to visualize.[4]

Definition of Terms

As defined by Börner et al., the broad concept of visualization "refers to the design of the visual appearance of data objects and their relationships." Börner explains that well-designed visualizations improve our interaction with large volumes of data, providing comprehension, understanding and "revealing relations otherwise not noticed."[5]

To decide what kind of graphical display is appropriate to reveal the data analyzed in this study, the authors used categorizations suggested by Börner and Polley in their 2014 text, *Visual Insights: A Practical Guide to Making Sense of Data*. Börner and Polley suggest units of analysis as Meso/Local, containing 101 to 10,000 records. The units of analysis examined for this study are the purchases for one year, the number of courses offered in each major and minor, student enrollment for one year, circulation since purchase, and the expenditures for books supporting each department. Each unit of data analyzed can be described as topical, asking "what."[6]
- What is the number of courses offered in each major and minor?
- What is expended in each subject area?
- What is the size of the physical collection in each subject area?
- What is student enrollment in each area?
- What is the circulation in specific areas for one year?

Börner and Polley describe a graph as the most common visualization used to examine Meso/Local topical data. Within the context of graphing visualizations, we display the results as circular visualizations. Further explanation of the visualizations can be found in the methodology section.

Literature Review: Collection Building and Budgeting

The library in this study supports a liberal arts and sciences curriculum of undergraduate and limited masters programs, and a student population of about 12,500. Like many libraries, the library that is the focus of this study does not have rigid criteria for ordering materials and setting budgets. A 2013 study by Catalano and Caniano finds that, when libraries examine the collection and expenditures, there is little formal rationale for allocating funds.[7]

Presently, for the library in this study, there is no set formula to determine budget except an initial expense to support new majors ($2,000) and new minors ($1,000). Once a year, subject liaisons are asked to justify budget increases or decreases. The Collection Development team also looks at past ordering and spending history. As new services such as patron-driven acquisition become available, some of the firm order budget is redirected to support those efforts. Of course, the book budget is only available after the serials costs are met.[8]

When academic libraries use deliberate methods to allocate funds, Catalano and Caniano find that most libraries use the following five methods: percentage-based, weighted multiple-variable, factor or regression analysis, historical spending plus use percentage of new formula, and circulation-based statistics.[9]

A 2007 study by Canepi includes a thorough literature review of various methods of allocating funds. Canepi acknowledges that all libraries allocate funds but vary in their

approach. Out of seventy-five different formulas used by libraries, Canepi pulls a final total of twenty-three formulas. The top four most frequently used factors in Canepi's study are student enrollment, cost of materials, circulation use, and number of faculty.[10]

Other studies call for the application of "more rigorous statistical methods" to create a "more equitable balance across departments," for using ROI techniques to assess institutional value, or for making use of objective data in decisions regarding collection development.[11]

In Rick Anderson's article, "Collections 2021," Anderson states that libraries, if they are to survive, must rethink their collecting and service strategies in radical and possibly scary ways and to do so sooner rather than later. Anderson predicts that, in the next ten years, the "idea of collection" will be overhauled in favor of "dynamic access to a virtually unlimited flow of information products."[12]

The library collection of today is changing, affected by many factors, such as demand-driven acquisitions, access, streaming media, interdisciplinary coursework, ordering enthusiasm, new areas of study, political pressures, vendor changes, and the individual faculty member following a focused line of research. If libraries do not allocate based on data, there could be subjective distribution of funds, affecting the perception of fairness and damaging the library's reputation on campus.[13]

As described by Blake and Schleper, when librarians think "more and more about the cost of information," new opportunities appear based on findings grounded in real data analysis. Knievel, Wicht, and Connaway suggest that subject librarians may see opportunities in looking more closely at the relatively unexplored "intersection of circulation, interlibrary loan, and holdings." Many studies are starting to examine using circulation and patron use data to support service, tying in with instructional outreach. Morrisey reminds us that collections data can inform decisions regarding services. Select databases that are heavily used or high-circulation areas may suggest a change in staffing concentrations or opportunities for outreach. Finnel et al. propose that reference transactions may point to scholarly conversations that are taking place both for students and faculty. Using data analysis on the local level may illuminate indicators of quality in much the same way the Leiden Ranking (http://www.leiden-ranking.com/) indicates scientific impact and scholarly collaborations worldwide.[14]

Literature Review: Using Visualizations to Address Library Problems

Much of the current research concerning library data visualization efforts address digital library collections, most often the interface and user environment. Two major sources for visualizations within libraries includes the entire 2005 January and February issue of *Library Technology Reports* and Sage's journal, *Information Visualization*. The 2005 *Library Technology Reports* issue addresses 2D and 3D visualizations and includes practical applications, resources, organizations, and a short bibliography. *Information Visualizations*, published by Sage, offers many examples of data visualization crossing multiple subjects.

In a 1999 visualization article, Beagle defines the difference between graphical representations of environments and knowledge visualization, which generates graphical representations of meaningful relationships among retrieved files or objects. In a 2003 work, Beagle applies data visualization to a digital library collection to foster the serendipitous discovery enjoyed by many while wandering physical stacks. Beagle's physical depiction of the collection based on LC subject area holdings is based on VisualNet, called "Scholastica," and depicts the relative size of the holdings in each class. Visually available to patrons is the type of material: print book, video, or e-book.[15]

Also working in the area of visualizing collections, we find the work of Zang, Junliang, and Mostafa, who use concepts and clustering to produce graphs of what is

available in a document collection. Major subtopics appear in the document collection as concept clusters.[16]

Pousman, Stasko, and Mateas describe the emphasis on using interactive visual models as attempting to provide amplified cognition and "deep insight for expert user populations." Along with user behavior and information seeking, many library data visualization studies address citation analyses.[17] Included in this focus is the important work of Katy Börner, who is a major influence in the visualization field. Börner's work within library literature concerns many areas of interest, including distinguishing patterns in scientific communication through citation analysis.

Other research that diverges from digital collections to analyze the physical library collection focuses on usage statistics and collection analysis. Lima describes how student Syed Reza Ali mapped transaction data from the Seattle Public Library to illuminate circulation trends. As reported by Brown and Stowers, knowledge gained from visualizing the physical collection is used most often to support assessment, decisions on cancellations, and proposals about which items to store remotely or to weed.[18]

A few studies use the term "mapping" the collection. Bailey suggests analyzing a collection by constructing a matrix of prominent authors, keywords, and public figures within a particular subject area.[19]

The visualizations produced in this study will provide a snapshot of the current collection, with room for further analysis as gaps appear. The authors hope to gain insight through looking at graphical representation of the number of physical books and a small number of e-books purchased in a single year in each collection area and expenditures in those areas, compared to the number of course hours offered, which reflect the number of students enrolled.

This study's primary focus is not on circulation numbers, although the authors provide some visualizations of circulation for one year, compared with expenditures and student enrollment. There are many variables in examining circulation, which may offer opportunity for a separate study. As stated by Bradford,

> Very often, a circ is not just a circ. Does that number include renewals or is it just first-time circulation? Those numbers can be, and often are, significantly different. Are you comparing items with different loan lengths? If your DVDs circulate for three days or one week, take that into account when comparing them to books that may circulate for three or four weeks.[20]

Literature Review: Tools

There are myriad tools available, described in the literature in the context in which they are used. An entire issue of *Library Journal* (March 2005) names tools for text clustering, topical browsing, and information mapping. In citation analysis studies, Dunne et al. name other visualization tools such as CiteSpace, Network Workbench, and the SocialAction Network analysis tool. For visualizations and chart making, Chapman and Woodbury describe open source products Protovis (http://vis.stanford.edu/protovis/), Highcharts (www.highcharts.com), Google Chart API (http://code.google.com/apis/chart/), and Microsoft Excel.[21] Other major players in the field and inspiring examples can be found on the site www.infovis.net.

Some tools are bundled with library products already owned. For example, Watters reports that WorldCat has an Identity Map that can be used for relationships among subjects, authors, and characters. Bradford offers tools for collection analysis that are bundled with other common library products: collectionHQ from Baker & Taylor, Decision Center from Innovative Interfaces, Inc., and Intota Assessment.[22]

Word Clouds appear in the literature[23] as analyzing social media but, in a more limited library setting, could be used to examine user searching behavior.

A study by Zang and Mostafa focuses on semantic relationships between words and describes the concept of a digital library. In the literature, there are many studies addressing visual interfaces and digital libraries, an area that is outside the scope of this study. However, looking at studies like Xu et al. suggest even more tools to use beyond the scope of digital collections.[24]

Exhaustive lists of data visualization tools include:

- the DIRT Directory (http://dirtdirectory.org/categories/visualization)
- Kathy Schrock's educating through infographics (www.schrockguide.net/infographics-as-an-assessment.html)
- Dataviz list of online tools (www.improving-visualisation.org/case-studies/id=5)

Visualization tools explored for this study include Plotly, Microsoft Excel, Python programming language, and D3.js, a javascript library for creating documents based on data.[25] Because the process should be easily replicated without special knowledge of programming language, the authors generated some visualizations using Tableau Public©, which is freely available. Tableau charts are easily customizable using drag-and-drop, which allows flexible and intuitive generating of data. Tableau accepts both text and Excel files.

Plotly, a free online data visualization tool, was explored with limited success. The need to know Python programming is the disadvantage in using Plotly. The advantage of using Plotly is the interactive visualizations that result, making engagement with the data very dynamic. Plotly is also social: the program is web-based, and visualizations may be shared among the community for insight and feedback.

In the end, the authors found most success generating data bubbles using Microsoft Excel (version 2010), which is probably familiar to the widest audience and requires no special programming skills. In using Excel, the authors could plot multiple variables in various combinations: department name; books purchased within a year, expenditures, course hours, student enrollment, and circulation since purchase. An excellent tutorial by Eugene O'Loughlin, National College of Ireland, is very helpful in composing the charts and is found here: https://youtu.be/4FyImh2G7N0.[26]

Methodology

For this study, data on the number of course hours by major and undergraduate enrollment by major was retrieved from the institution's Office of Institutional Research, Planning, and Information Management website. The authors chose to include semesters for one academic year: 2013–2014.

The input data on purchases is found in the library catalog and from records held by the Collection Development department. Collection Development provided a list of titles purchased from firm order, approval plan, and demand-driven acquisition (DDA) budgets, separated by fund code. The authors totaled firm order, approval plan, and DDA by fund code to get a total number of purchases by fund for Fiscal Year (FY) 2014.

It is important to note that the numbers provided by the Collection Development department includes both print and a small number of DDA books. The e-book collection included is DDA and firm orders purchased for perpetual access. The number is very small, too insignificant to skew the numbers.

Not included in the study are e-books purchased as part of a large subscription package such as e-brary or EBSCO e-books. The institution subscribes to seven different platforms with major e-book holdings and about six more with smaller e-book holdings. E-book collections cover many subject areas. It was thought that examining the e-books collections by subject would not add to this study, for several reasons. E-

book use is calculated primarily at the chapter or page level rather than the title level, some books allow full-title downloads while others do not, and download statistics are only generated if the entire book is downloaded, leaving out partial viewings. At this time, e-book usage is calculated so differently that it was not included in this study.[27]

The purchase data also excludes databases and journals, which are purchased from a different budget and are often interdisciplinary in subject.

Aligning library fund codes with the majors offered, the authors compiled the data into one large Excel spreadsheet. The authors manually entered the corresponding fund code in the column next to the major, then used Excel's VLOOKUP function to bring the data into one sheet. A few special discretionary funds were excluded because they did not correspond to a major. These exclusions were very small funds, less than 1 percent of the total purchases. The figure for Expenditures is taken from the total firm orders, books from the approval plan, and DDA. DDA purchases were a pilot program in 2014 and comprised a small part of the total purchases.
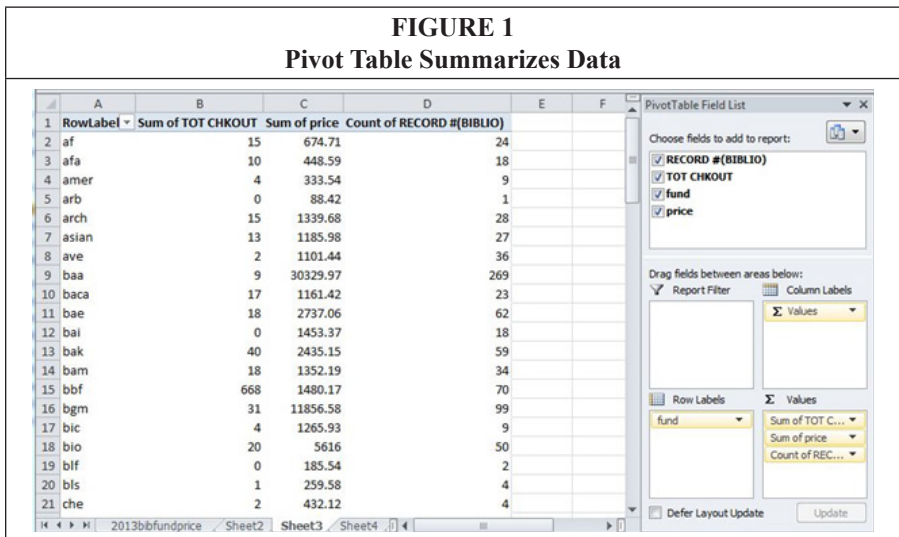
Three lists were generated to get raw data:

- List 1 includes firm book order records with paid date covering FY 2013–2014. No fund was specified.
- List 2 includes all bibliographic records attached to the List 1 orders.
- List 3 is composed of all items attached to the bibliographic records from List 2, limited to item type = books and status = available.

Exported from List 1 is bibliographic record number, fund code, and price. Exported from List 3 is bibliographic record number and total circulation. Both text files generated were imported into Excel, then combined using the VLOOKUP function to pull the circulation figures into a column in List 1. A pivot table was used to summarize the data, as seen in figure 1. The values from the pivot table were copied into a new Excel sheet for editing. The export initially using Tableau was performed several times, as the authors encountered varied results that occurred due to items held by Special Collections, which are in-house only and don't circulate. Another variable that muddies an accurate grasp of expenditures is the fact that some disciplines buy fewer, more expensive books, while others purchase inexpensive books in larger quantities.

The authors looked at the data for 2013 first, then compared data from 2014 using the same methods to see if similar patterns emerged.

FIGURE 1
Pivot Table Summarizes Data

	A	B	C	D	E	F
1	RowLabel	Sum of TOT CHKOUT	Sum of price	Count of RECORD #(BIBLIO)		
2	af	15	674.71	24		
3	afa	10	448.59	18		
4	amer	4	333.54	9		
5	arb	0	88.42	1		
6	arch	15	1339.68	28		
7	asian	13	1185.98	27		
8	ave	2	1101.44	36		
9	baa	9	30329.97	269		
10	baca	17	1161.42	23		
11	bae	18	2737.06	62		
12	bai	0	1453.37	18		
13	bak	40	2435.15	59		
14	bam	18	1352.19	34		
15	bbf	668	1480.17	70		
16	bgm	31	11856.58	99		
17	bic	4	1265.93	9		
18	bio	20	5616	50		
19	blf	0	185.54	2		
20	bls	1	259.58	4		
21	che	2	432.12	4		

Some departments didn't align with fund codes, so that represents some extra work in areas such as Environmental Geoscience and Astronomy. It could be that other disciplines, like Environmental Studies, are close enough to provide adequate coverage, but that opens up another area for research.

In other cases, to simplify the data bubbles, the authors chose to limit. For example, course hours for BioChemistry were missing, so that department was not included. Any omissions can be explored more thoroughly in the future.

Findings/Discussion

By looking at the data, more questions are revealed, much like archaeological excavation, good detective work, or the research process. The three-dimensional data bubble visualizations offer a starting point for discussing the collections in support of the curriculum and what is expended in each area. The visualizations provide greater comprehension than the two-dimensional "flatland" of the spreadsheets, in which valuable questions and insights are lost in the columns and rows of data. A screenshot of a portion of the Excel spreadsheet containing library fund codes, course hours, books purchased, percentages, expenditures, and enrollment in each discipline is seen in figure 2.[28]

Using data visualization instead of a spreadsheet, figure 3 offers a much more vibrant depiction of books purchased within one year, expenditures and course hours for most of the schools. The data bubbles are easy to understand at a glance. A large school not included in figure 3 is the School of Education, Health and Human Performance, because of Excel limitations of displaying that much data in one chart. The School of Education, Health and Human Performance is compared with the School of the Arts in figure 7.

FIGURE 2
Screenshot of an Excel Spreadsheet Containing Library Fund Codes, Course Hours, Books Purchased, Percentages, Expenditures and Enrollment

Microsoft Excel - 2014 title count by fund.xlsx [Read-Only]

	School	Fund	Fund Code	Course Ho	Books Purc	% Course Hours	% Books Pi	Expenditur	% Expendit	Undergrad	% Enrollment	Estimated checkout rate
2	School	African Stu	af	126	30	0.000811511	0.002704	1005.58	0.001434	42	0.000859564	0.62963
3	School	African Am	afa	408	20	0.002627749	0.001803	1031.78	0.001472	136	0.002783349	0.529412
4	School	Arabic Stu	arb	280	18	0.001803357	0.001622	950.69	0.001356	75	0.001534935	0.083333
5	School	Archaeolo	arch	4	30	2.57622E-05	0.002704	2109.9	0.00301	1	2.04658E-05	0.413793
6	School	Asian Stud	asian	515	17	0.003316888	0.001532	1090.01	0.001555	37	0.000757235	0.444444
7	School	Accountin	baa	4695	87	0.03023843	0.007842	5063.89	0.007223	1453	0.02973681	0.473684
8	School	Computing	baca	192	29	0.001236588	0.002614	2200.12	0.003138	64	0.001309811	0.789474
9	School	Econ & Fin	bae	5209	360	0.033548877	0.03245	19408.02	0.027684	1699	0.034771397	0.278689
10	School	Internatio	bai	255	44	0.001642343	0.003966	3105.23	0.004429	85	0.001739593	0
11	School	Marketing	bak	1980	41	0.012752309	0.003696	2603.33	0.003713	660	0.013507429	0.661017
12	School	Managem	bam	5186	106	0.033400745	0.009555	4540.91	0.006477	787	0.016106586	0.545455
13	School	Biology	bio	11063	360	0.071251916	0.03245	18884.38	0.026937	5160	0.105603536	0.349206
14	School	Chemistry	che	4652	97	0.029961485	0.008743	12130.38	0.017303	2238	0.045802464	0.5
15	School	Classics	cla	1734	154	0.011167931	0.013881	12904.84	0.018407	560	0.011460849	0.787611
16	School	Communic	cmm	6189	179	0.039860626	0.016135	19615.6	0.02798	2066	0.042282346	0.479508
17	School	Computer	com	2960	172	0.019064058	0.015504	14825.42	0.021147	938	0.019196922	0.5625
18	School	Elementar	ede	2879	207	0.018542372	0.018659	12125.23	0.017295	708	0.014489788	0.273292
19	School	Education:	edf	5152	430	0.033181765	0.03876	22273.6	0.031771	773	0.015820065	0.15942
20	School	English	eng	7305	1105	0.047048291	0.099603	49827.47	0.071074	1850	0.037861733	0.926471
21	Interdi	Environme	env	713	179	0.004592119	0.016135	14696.97	0.020964	59	0.001207482	0.318966
22	School	European	eur	33	13	0.000212538	0.001172	1235.31	0.001762	11	0.000225124	0.785714
23	School	Exercise S	exs	1263	10	0.008134427	0.000901	2002.93	0.002857	444	0.009086816	0.428571
24	School	Art History	faa	2334	474	0.015032267	0.042726	29085.57	0.041488	759	0.015533543	0.529954
25	School	Theatre	fad	3896	362	0.025092422	0.03263	14251.73	0.020329	1320	0.027014858	0.26087
26	School	Dance	fadc	319	23	0.002054539	0.002073	1601.34	0.002284	130	0.002660554	0.214286
27	School	Music	fam	3309	284	0.021311813	0.025599	16673.04	0.023782	1259	0.025766444	0.23913
28	School	Studio Art	fas	2313	96	0.014897015	0.008653	4852.59	0.006922	771	0.015779133	0.296875
29	School	Arts Mana	fat	1353	56	0.008714078	0.005048	4160.94	0.005935	451	0.009230077	1
30	School	French	fre	2439	83	0.015708526	0.007482	4875.63	0.006955	793	0.016229381	0.144928
31	School	Geology	geo	3883	130	0.025008695	0.011718	12466.38	0.017782	1835	0.037554746	0.326389
32	School	German	ger	1164	138	0.007496812	0.012439	6651.83	0.009488	355	0.00726536	0.19403
33	School	Hebrew	heb	95	19	0.000611853	0.001713	1253.24	0.001788	24	0.000491179	0.761905
34	School	Hindi	hin	48	21	0.000309147	0.001893	338.07	0.000482	12	0.00024559	0.4
35	School	Historic Pr	hip	900	69	0.005796504	0.00622	2797.07	0.00399	291	0.005955548	0.44
36	School	History	his	8385	1336	0.054004096	0.120425	58122.15	0.082905	2638	0.053988785	0.647727
37	School	Hospitality	hos	1869	34	0.012037407	0.003065	2705.4	0.003859	565	0.011563178	0.65
38	School	Internatio	int	1567		0.010092358				513	0.010498956	
39	School	Italian	ita	987	22	0.006356833	0.001982	617.6	0.000881	329	0.006733249	0.5

Big Funds Table / 2013 # of titles by fund / 2013 firm / Course Hours by Discpline / Course Hours / Fund Codes / Course Hours by dept / 7yearsCd

FIGURE 3
Data Bubbles Representing Number of Books Purchased, X-Axis
Showing Expenditures & Y-Axis is Course Hours for School of Humanities
& Social Sciences, School of Sciences & Mathematics, School of Business,
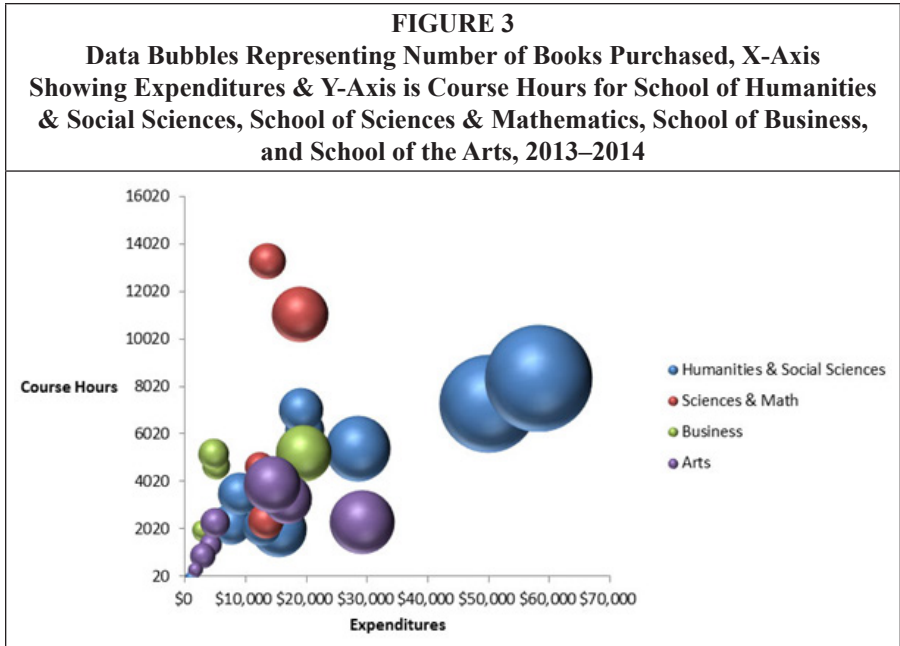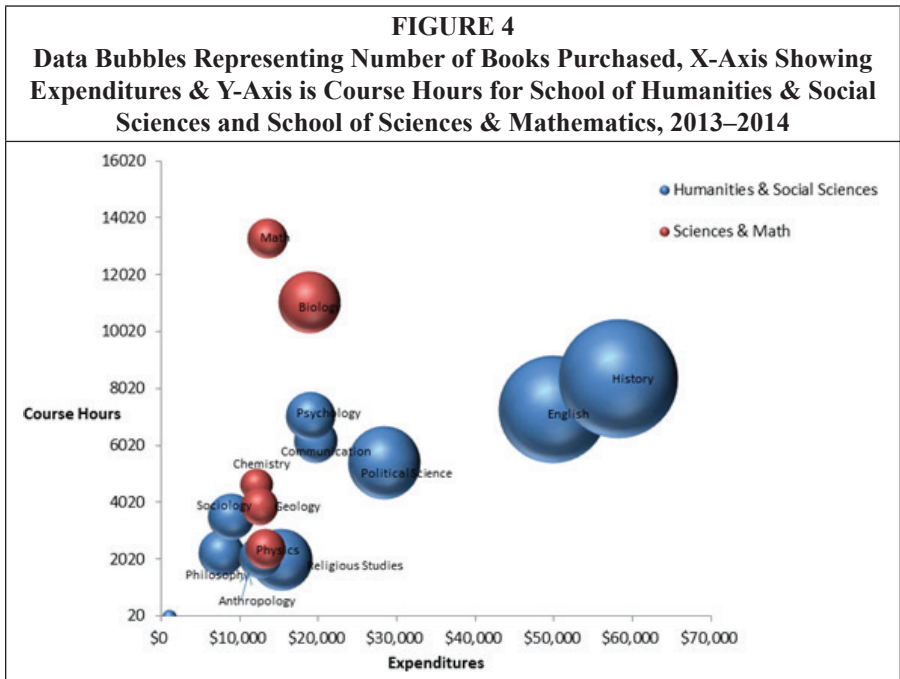**and School of the Arts, 2013–2014**

Figure 4 offers several visualizations that ignite opportunities for discussion. For example, math's course hours are huge, which were unexpected, although explained by high enrollments for required general education requirements and increasing math and statistics requirements from other courses.[29] Math's small physical collec-

FIGURE 4
Data Bubbles Representing Number of Books Purchased, X-Axis Showing
Expenditures & Y-Axis is Course Hours for School of Humanities & Social
Sciences and School of Sciences & Mathematics, 2013–2014

tion is probably typical of many universities. However, a closer look at support for that department is warranted. The undergraduates fulfilling requirements may not be conducting research, but do the faculty teaching the high number of classes need additional support or specialized databases for their research?

Further study of figure 4 suggests that communication and psychology may benefit from discussion about an increase in book budget. They both have small collections and fewer expenditures but are large majors. On the other hand, religious studies has a large collection but low course numbers and low expenditures. What is the reason? Active ordering? Bargain books? Increased communication with the library and departmental liaisons and with the Collection Development team is needed to answer these questions.

In figure 4, there is a small bubble near the origin that is only partially shown, with a small budget and no course hours. This bubble represents the special discretionary fund of a faculty member with relatively narrow and rare research interests. This scholar has ordered twenty books, but their collection is growing and represents a unique niche the library can advertise to other scholars conducting similar research. As Steele suggests, data visualization is a useful tool to reveal these collection oddities and perhaps provide a marketing opportunity for libraries. The library could highlight the collection for Interlibrary Loan or begin a miniconference for visiting scholars.[30]

Other topics of interest illustrated in figure 4 are the large collections for English and history. These areas don't have the largest enrollments or course hours, but they have huge collections and healthy expenditures. When university budgets are threatened and shrinking, a look at circulation statistics can justify the numbers in these two areas if there is any challenge.

As seen in figure 4 and figure 5, any area in the top left quadrant of the charts needs review. Are the collection needs of the subjects that have high course hours and healthy student enrollments being met? If not by expenditures, then are there other resources not included in the figures? Could funds from areas with high expenditures but low course hours be redirected to support low budget departments?

FIGURE 5
Data Bubbles Representing Number of Books Purchased, X-Axis Showing Expenditures & Y-Axis is Course Hours for School of Business and School of the Arts, 2013–2014
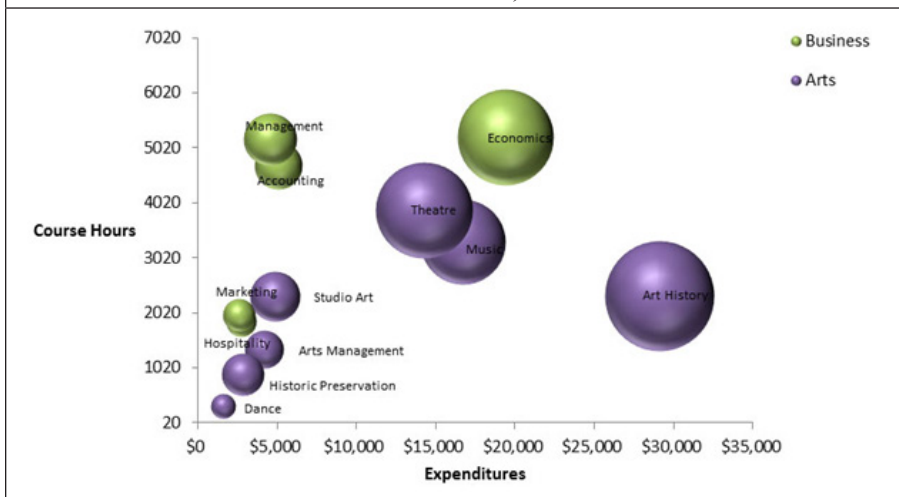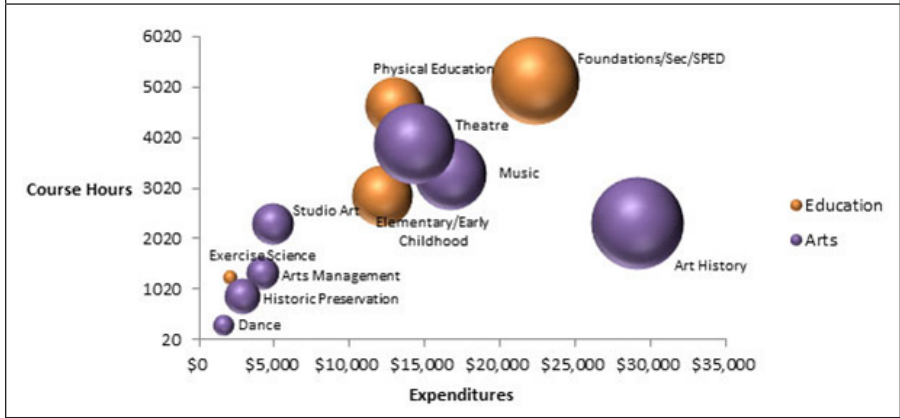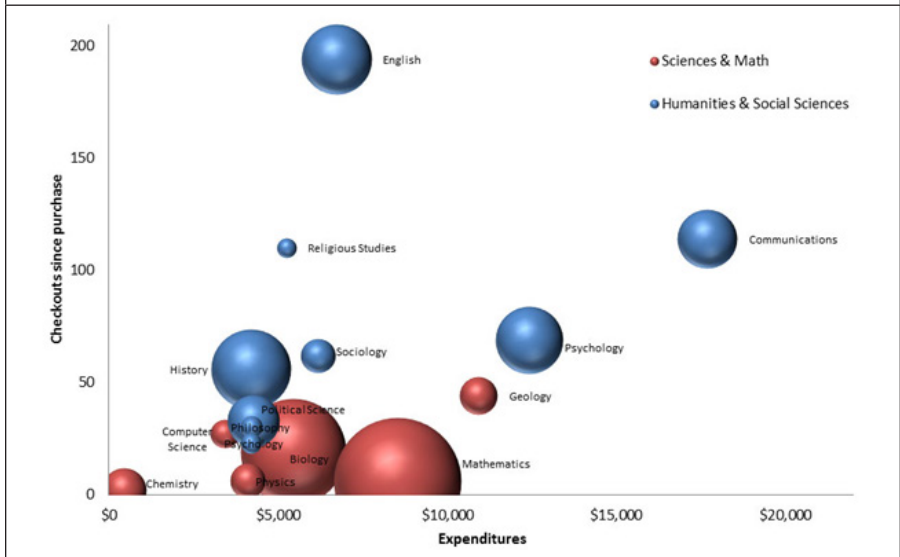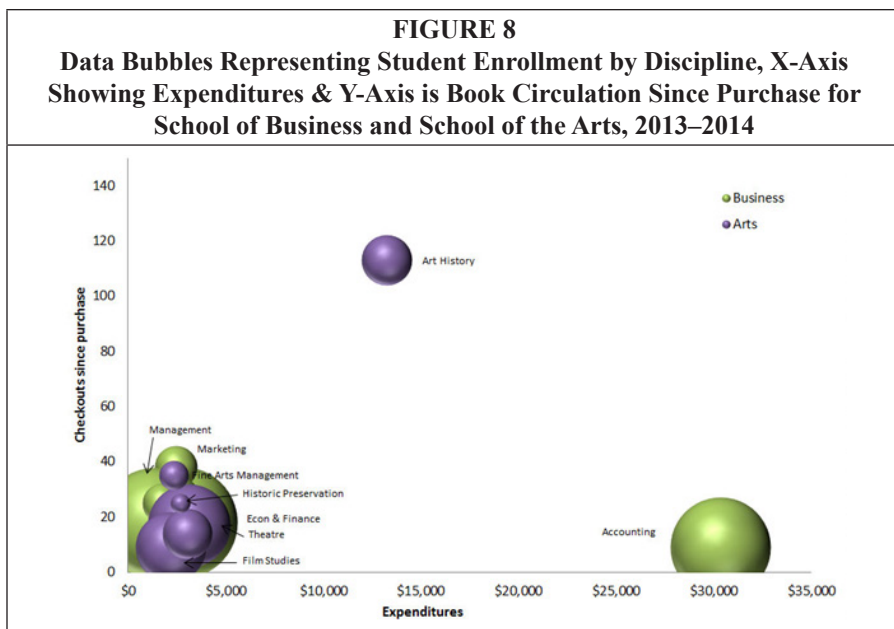
FIGURE 6
Data Bubbles Representing Number of Books Purchased, X-Axis Showing Expenditures & Y-Axis is Course Hours for School of Education, Health, and Human Performance and School of the Arts, 2013–2014

Also of interest are circulation numbers. In figure 7 and figure 8, the size of the data bubbles represent student enrollment and physical book circulation since purchase is graphed with expenditures. Again, it is easy to imagine conversations taking place about the significance of departments that have healthy student enrollment, robust circulation, but small expenditures, or conversely, areas in which healthy expenditures

FIGURE 7
Data Bubbles Representing Student Enrollment by Discipline, X-Axis Showing Expenditures & Y-Axis is Book Circulation Since Purchase for School of Humanities & Social Sciences and School of Sciences & Mathematics, 2013–2014

FIGURE 8
Data Bubbles Representing Student Enrollment by Discipline, X-Axis Showing Expenditures & Y-Axis is Book Circulation Since Purchase for School of Business and School of the Arts, 2013–2014

are occurring, with very little circulation of materials and low student enrollment. In figure 8, the data bubbles are too overlaid when scaled to 100 to be of much use, but they still provide important clues about the appropriateness of the collection. Right away, we can visualize economics and finance, accounting and film studies as outliers that need discussion and possible attention. Using Excel, the data fields may be adjusted or changed on the fly during a meeting to foster meaningful conversation about implications.

By looking at data visualized in different combinations, library collection development teams can clearly compare important considerations in collection management: expenditures and purchases, circulation, student enrollment, and course hours. Library staff and administrators can make funding decisions or begin dialog based on data free from political pressure or from the influence of the squeakiest wheel in a department. The visual depiction of information revealed in data bubbles represents an opportunity for conversation among collection development teams, subject liaisons, and other interested parties.

Implications for Future Research

Future research areas call for experimenting with different data visualizations using alternate tools or in additional areas. An obvious first step is to try to compare the size of the book collection for each area beyond purchases for a single year. While looking at the collection for a single year may hint at supporting subject areas, more definitive gathering of collection numbers is needed. Looking at the entire collection using data visualization may provide a new way of performing collection assessments. Comparing the collection with circulation figures may be used with other variables to suggest weeding decisions. Libraries may look at other items, such as DVDs, and determine how they circulate. An examination of e-book usage statistics, if possible given the many variables discussed earlier, may reveal interesting trends. Could patterns of interlibrary loan requests of materials be easily understood through data visualization, suggesting solutions for lending?

Once the visualization tool is selected, data are gathered and cleaned up, a workflow is created and the process delineated, data combinations may be studied as needed. For example, what patterns might appear when figures are compared from Interlibrary Loan, patron use, and instructional sessions? How about amount allotted in discretionary funds compared to expenditures for new majors? What happens when the collections for majors and minors are compared with the collections of popular interdisciplinary subjects?

Conclusion

The need for examining collection data clearly extends beyond simply buying materials to support curriculum or to meet the requirements of the most vocal faculty. Accurate visualizations of library data suggest avenues for staffing and service, resource expenditures, scholarly relationships and instructional outreach as well as opportunities for excellent collection development.

Groups to help with data visualization include The Office for Creative Research, found at http://o-c-r.org/abstract/. This group includes Jer Thorp, data artist. A description of Thorp should sound familiar, as it also describes librarian and information scientists. In a *National Geographic* interview, Jer states that his biggest thrill comes from engaging with a completely new topic.

"I get to become a little expert in a lot of different things," he says. "We work on projects that are in all kinds of categories and all types of subject areas, and we really make an effort to become as educated about all of them as we can."[31]

Librarians and data artists like Thorp are alike. We can benefit from becoming better versed with the tools of data analysis. The call for librarians to become more comfortable with data is echoed in the literature. Brown and Stowers suggest justifying collections expenditures with data analysis, important since collection expenditures is "second only to personnel in the library's budget." As supported by Morrisey, a "thorough data analysis will let you know how people are using your book collections and may inform you as to adjusting collections dollars among the disciplines."[32]

The changing landscape of collection development calls for a more accurate, unbiased, and objective view of library holdings using a combination of data gathering to give an overall picture of the strength or weakness of the collection.[33]

In creating data visualizations that are clearly understood at a glance, without extravagant explanation, librarians will be able to have meaningful conversations resulting in free and impartial decision making.

Notes

1. Katy Börner, Chaomei Chen, and Kevin W. Boyack, "Visualizing Knowledge Domains," in *Annual Review of Information Science and Technology,* vol. 37 (2003) ed. B. Cronin (Medford, N.J.: Information Today, Inc.), doi:10.1002/aris.1440370106: 209.

2. Edward R. Tufte, *The Visual Display of Quantitative Information*, 2nd ed., (Cheshire, Conn.: Graphics Press, 2001).

3. W. Xu, M. Esteva, S.D. Jain, and V. Jain, "Interactive Visualization for Curatorial Analysis of Large Digital Collection," *Information Visualization* 13 (2013): 159–83, doi:10.1177/1473871612473590.

4. Brad Eden, "Practical Applications of 2D and 3D Information Visualization for Information Organizations," *Library Technology Reports* (2005), available online at https://journals.ala.org/ltr/article/view/4599/5427 [accessed November 6, 2014].

5. Börner, Chen, and Boyack, "Visualizing Knowledge Domains," 209.

6. Katy Börner and David Polley, *Visual Insights: A Practical Guide to Making Sense of Data* (Cambridge, London: MIT Press, 2014), 7.

7. Amy J. Catalano and William T. Caniano, "Book Allocations in a University Library: An Evaluation of Multiple Formulas," *Collection Management* 38 (2013): 192–212, doi:10.1080/01462679.2013.792306.

8. Katina Strauch, e-mail message to authors, December 3, 2014.

9. Catalano and Caniano, "Book Allocations in a University Library," 5, 193.

10. Kitti Canepi, "Fund Allocation Formula Analysis: Determining Elements for Best Practices in Libraries." *Library Collections, Acquisitions, and Technical Services* 31 (2007): 12–24, doi:10.1016/j.lcats.2007.03.002.

11. George Stachokas and Tim Gritten, "Adapting to Scarcity: Developing an Integrated Allocation Formula," *Collection Management* 38 (2013): 33–50, doi:10.1080/01462679.2012.730495; Denise Pan, Gabrielle Wiersma, Leslie Williams, and Yem S. Fong, "More Than a Number: Unexpected Benefits of Return on Investment Analysis," *Journal of Academic Librarianship* 39 (2013): 566–72, doi:10.1016/j.acalib.2013.05.002; Robin Bradford, "Getting Data Right," *Library Journal* 139 (2014): 26.

12. Rick Anderson, "Collections 2021: The Future of the Library Collection Is Not a Collection," *Serials* 24 (2011): 211–16.

13. Katina Strauch, e-mail message to authors, June 17, 2015; Canepi, "Fund Allocation Formula Analysis," 2.

14. Julie C. Blake and Susan P. Schleper, "From Data to Decisions: Using Surveys and Statistics to Make Collection Management Decisions," *Library Collections, Acquisitions, and Technical Services* 28 (2004): 460–64, doi:10.1016/j.lcats.2004.09.002; Jennifer E. Knievel, Heather Wicht, and Lynn Silipigni Connaway, "Use of Circulation Statistics and Interlibrary Loan Data in Collection Management," *College & Research Libraries* (2006): 35–50; Locke Morrisey, "Data-Driven Decision Making in Electronic Collection Development," *Journal of Library Administration* 50 (2010): 283–90, doi:10.1080/01930821003635010; Joshua Finnel and Walt Fontane, "Reference Question Data Mining: A Systematic Approach to Library Outreach," *Reference & User Services Quarterly* 49 (2010): 278–86.

15. Donald Beagle, "Visualization of Metadata," *Information Technology and Libraries* (1999): 192–99; Donald Beagle, "Visualizing Keyword Distribution across Multi-Disciplinary c-Space," *D-Lib Magazine* 9 (2003), doi:10.1045/june2003-beagle.

16. Junliang Zang and Javed Mostafa, "Information Retrieval by Semantic Analysis and Visualization of the Concept Space of D-Lib® Magazine," *D-Lib Magazine* 8 (2002), doi:10.1045/october2002-zhang.

17. Zachary Pousman, John T. Stasko, and Michael Mateas, "Casual Information Visualization: Depictions of Data in Everyday Life," *IEEE Transactions on Visualization and Computer Graphics* 13 (2007): 1145–52; V. Nikolaevich, G. Nikolai, and A. Mazov, "Detection of Information Requirements of Researchers Using Bibliometric Analyses to Identify Target Journals," *Information Technology and Libraries* (2013): 66–77.

18. Manuel Lima, *Visual Complexity: Mapping Patterns of Information* (New York: Princeton Architectural Press, 2011), 211; Jeanne M. Brown and Eva D. Stowers, "Use of Data in Collections Work: An Exploratory Survey." *Collection Management* 38 (2013): 143–62, doi:10.1080/01462679.2013.763742.

19. Lea Bailey, "Does Your Library Reflect the Hispanic Culture? A Mapping Analysis," *Library Media Connection* (2009): 20–24.

20. Bradford, "Getting Data Right," 26.

21. Cody Dunne, Ben Shneiderman, Robert Gove, Judith Klavans, and Bonnie Dorr, "Rapid Understanding of Scientific Paper Collections: Integrating Statistics, Text Analytics, and Visualization," *Journal of the American Society for Information Science and Technology* 63 (2012): 2351–69; Joyce Chapman and David Woodbury, "Leveraging Quantitative Data to Improve a Device-Lending Program," *Library Hi Tech* 30 (2012): 210–34, doi:10.1108/07378831211239924.

22. A. Watters, "Visualization of the Week: Visualizing the Library Catalog" *Radar: Insight, Analysis, and Research about Emerging Technologies* (Aug. 2011), available online at http://radar.oreilly.com/2011/08/visualization-of-the-week-visu.html [accessed 14 July 2014]; Bradford, "Getting Data Right," 26.

23. H. Andrew Schwartz et al., "Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach," *PloS One* 8, ed. Tobias Preis (2013): e73791, doi:10.1371/journal.pone.0073791.

24. Junliang Zang and Javed Mostafa, "Information Retrieval by Semantic Analysis and Visualization of the Concept Space of D-Lib® Magazine," *D-Lib Magazine* 8 (2002), doi:10.1045/october2002-zhang; Xu, Esteva, Jain, and Jain, "Interactive Visualization."

25. "DHO:Discovery (fionnachtain)," last modified n.d., http://discovery.dho.ie/discover.php.

26. "How to Draw and Format a Basic Bubble Chart in Excel 2010," YouTube video, 7:34, posted by Eugene F.M. O'Loughlin (Apr. 5, 2013), https://youtu.be/4FyImh2G7N0.

27. Michelle Sellars and Lindsay Barnett, e-mail message to authors, December 10, 2015.

28. Edward R. Tufte, *Envisioning Information* (Cheshire, Conn.: Graphics Press, 1990).

29. Robert J. Mignone, e-mail message to authors, June 12, 2015.

30. Kirstin Steele, "Visualizing the Value of Library Content," *Bottom Line: Managing Library*

Finances 26 (2013): 14–17, doi:10.1108/08880451311321537.

31. R. Schleeter, "Data Artist: Jer Thorp," *National Geographic Education* (2013), available online at http://education.nationalgeographic.com/education/news/data-artist-jer-thorp/?ar_a=1 [accessed 6 November 2014].

32. Jeanne M. Brown and Eva D. Stowers, "Use of Data in Collections Work: An Exploratory Survey," *Collection Management* 38 (2013): 143–62, doi:10.1080/01462679.2013.763742; Locke Morrisey, "Data-Driven Decision Making in Electronic Collection Development," *Journal of Library Administration* 50 (2010): 283–90, doi:10.1080/01930821003635010.

33. Julie C. Blake and Susan P. Schleper, "From Data to Decisions: Using Surveys and Statistics to Make Collection Management Decisions," *Library Collections, Acquisitions, and Technical Services* 28 (2004): 460–64, doi:10.1016/j.lcats.2004.09.002.