

Mapping the Inside of a Collection: ArcGIS as Content Analysis Tool

B. Grantham Aldred

University of Illinois at Chicago, USA

Abstract: This article explores the potential of geospatial visualizations of subject matter content in urban planning books at the University of Illinois Chicago. Using ArcGIS, the visualizations explore gaps in the collection and provide analysis of the strengths and weaknesses of the collection and explore the collection at different levels. The analysis is then used to create ordering and weeding priorities and highlight parts of the collection. Overall, the study finds weaknesses in metadata standards do create limitations, but the visualizations make comprehensible many aspects of the collection that can drive collection development decisions.

Keywords: geographic information systems, collection development, content analysis, urban planning, subject headings

1. Introduction

Collection development faces a challenge, the relationship between data driven decision making and the qualitative nature of textual information. In an age when decisions increasingly need to be justified by data, the textual content within a collection defies easy quantification. Failing to address this challenge means decisions may be made only using the easily counted aspects of a collection, failing to consider the quality of the content within. Fortunately, there are tools available in the modern computing environment that allow for the transformation of qualitative content into data, and the further development of those data into visualizations designed to support decision making and communication.

To do this effectively, a librarian may be well served by making use of analytical tools available related to the content being analyzed. A tool designed to analyze information within a discipline may be usable to perform a meta-analysis of a collection that examines that information. To test this question, whether a discipline's tools could be used to analyze a collection of material in a way that would support data focused decision making, this study attempts to examine a collection of material on urban planning using geographic visualizations, revealing geographic patterns in the data to help guide decisions



made on ordering, weeding and communicating about the collection. This process, transforming the qualitative content into analytical data, then using the data to develop geographic visualizations, should provide a model for how to address this challenge.

2. Literature Review

There is minimal literature addressing the exact proposition at the heart of this study; no articles addressing the transformation of qualitative information into data for analysis or in the use of data visualization of geographic qualitative subject content for collection development decision making. However, there are a number of questions that can be answered by the literature that connect to aspects of the central research question and help guide the inquiry.

1. Is Data Visualization useful in library decision making?
2. Is geographic subject content useful for library patrons?
3. Is GIS a usable tool for Libraries?

These questions show how different facets of the question build on prior research to expand library practice and examining how those subjects have been addressed helped guide the design of this study.

The question at the heart of the study involves the use of data visualization for library decision making and the literature on library use of data visualization is extensive. Studies have been done examining aspects such as dates of publication within a collection(Taylor & Mitchell, 2016) tracking library budget expenditure by department(Finch & Flenner, 2016) and the examination of topical coverage in an institutional repository(Polley, 2016) among many others. These studies and others present data visualizations as effective tools for library decision making, allowing the comparison of disparate data and easing communication of those decisions. Kilb summed this up well, “Visualizations themselves and the data they summarize should be tailored to their intended audience and should clearly answer questions that drive decision making. (Kilb & Jansen, 2016): 200).” These decisions can be extended directly to the subject of collection development. “On a broad scale, librarians can use the information gathered to decide or confirm subject strengths and weaknesses and acquisition emphases.(Blake, 2004: 463)”

While the literature agreed on the effectiveness, there were a number of particular points that emerged to guide data visualization studies. One of the clearest guidelines was the importance of planning in the overall data visualization process. Ben Fry’s *Visualizing Data* lays the process out in a number of steps--Acquire, Parse, Filter, Mine, Represent, Refine, Interact-- and recommends careful planning for addressing each step as part of the process. Beyond that, however, Fry counsels that a data visualization planning should start with a need. “Great information visualization never starts from the standpoint of the data set; it starts with questions. Why was the data collected, what’s interesting about it, and what stories can it tell?(Fry, 2008)”

These design principles proved helpful in the design of the study and the development of the methods for the study followed these principles and focused on questions that could be asked in terms of collection development, expanding beyond existing data. “As more librarians become increasingly skilled at turning library data into visualizations, we should look for ways to integrate other data sources into our visualizations for collection development.(Taylor and Mitchell, 2016: 92)”

Another facet of the research that is present in the library literature is the relevance of geographic subject matter for users. When addressing questions that are potentially relevant to users, the question of geography appears prominently. Multiple studies have addressed the importance of geographic information for user search behavior in multiple contexts. Sanderson and Han’s conference presentation “Search Words and Geography” examined a corpus of searches from a major search engine and found that geographic terms were among the most common recurring terms(Sanderson & Han, 2007). This built on similar findings from Jones et al. that found geo-specification a common aspect of query re-writes or requests for more specific information(Jones, Zhang, Rey, Jhala, & Stipp, 2008). These articles show that geographic information is a known need for users, with users seeking information based on location. From a libraries and archives perspective, Clough et al. explored the importance of enhancing the UK government’s National Archives by improving the geographic information attached to historic data for findability purposes(Clough, Tang, Hall, & Warner, 2011). This was additionally supported by the study “Improving the Geospatial Consistency of Digital Libraries Metadata” by Renteria-Agualimpia et al, which looked at the problems that emerge when geographic information is inconsistently applied to maps and atlases. “Metadata sharing and reuse in digital libraries should include richer geospatial consistency validation: to ensure the data retrievability; to improve the quality of the entire library processes; and also to improve the entire user experience.(Renteria-Agualimpia, Lopez-Pellicer, Lacasta, Zarazaga-Soria, & Muro-Medrano, 2016): 521)”

In considering the use of geographic visualization tools to address these problems, it is important to consider whether these tools are accessible to libraries. The library literature can answer this question as well. While there have not been library studies that use GIS to analyze qualitative subject matter in a collection, there have been a number of studies that have used GIS in other ways. The most relevant study to the direct question in the literature is from Shonn M. Haren’s “Data Visualization as a Tool for Collection Assessment” in which Haren produces a number of different visualizations to analyze a Latin American Studies collection. One of the visualizations in the study is a map of different volumes categorized by publication location(Haren, 2014). While this does not involve the transformation of qualitative information into mappable data, it does show that a geographic representation of a collection can be done.

There also exist a number of studies that use GIS for analysis of library related issues. These studies, such as John Ottensmann's study of library branch utilization(Ottensmann, 1997) and Roya Pournaghi's GIS analysis of library space usage (Pournaghi, 2017) point to the fact that librarians are using GIS for other sorts of analysis. The book "Integrating Geographic Information Systems into Library Services" explores the importance of GIS analysis to libraries in a way that highlights the general point, that there are librarians out there who are competent in GIS and able to use it for analysis(Abresch, 2008). Bradley Wade Bishop and Lauren H. Mandel examined the breadth of this literature back in 2010 and found 34 different specific LIS articles using GIS technology, which points to the general accessibility of GIS as a tool for library data visualization(Bishop & Mandel, 2010).

Together, these lessons from the literature guided the development of the analysis in the following ways. The decision was made to acquire data specific to a limited subject matter(Urban Planning) in order to create a tailored visualization. Geographic information was chosen as a subject relevant to user needs. ArcGIS was selected for the project because of the broad availability of different tools, allowing a variety of maps to be produced for different sorts of analysis. While other mapping tools exist, they are more limited in types of output, and the literature indicated ArcGIS was accessible for library use.

3. Methods

3.1. Project Design

To achieve the ends of creating data visualizations based on a collection, a study was designed to approach the various aspects of the central research question. The goal was to develop a bounded set of materials that could be analyzed for geographic data and for which collection development actions could be taken. Based on this, the set had the following specific characteristics, LC Call number range HT1-HT395, generally inclusive of urban groups and cities, in the UIC main library collections. The call number range was chosen over the use of subject headings for two reasons, because it could be more easily bounded for collection development decisions and because preliminary review of the library collections revealed that there were relevant materials that lacked relevant subject headings and would have been excluded from the study. The determination to focus on the main library collections was made to exclude a local collection of local materials from the Department of Housing and Urban Development that was incompletely cataloged and would have skewed the results of the data.

With the characteristics of the dataset determined, the data was pulled from the library's Voyager ILS. The dataset included the following information for all volumes within the call number range, along with associated MARC fields. Title(245), Author(100), Table of Contents(505), Notes(500), Subject Headings(6XX).

3.2. Annotation Process

The process of transforming the qualitative text information from the collection into mappable geographic data involved a process of human annotation. This process included multiple examinations of each volume in the dataset to parse out relevant textual references to geographic locations and involved multiple steps.

The first step in the process was deduplication. The dataset as pulled from the ILS included multiple copies of several books, and the list was reviewed to minimize duplication. These duplicates fell into a two categories, multiple copies of the same volume, and both electronic and physical versions of the same volume. Given that the study design focused on examining diversity of content, volumes with identical content were excluded.

The second step in the annotation process involve the annotation of titles for geographic content. The was done through a human annotation process, as titles were examined for whether they included explicit or implicit references to geographic subject matter within the text of the title. When content was identified related to a geographic location, notes were made in the spreadsheet tracking geographical content at multiple scales, Region, Nation, State or Subregion and City terms were annotated.

Explicit geographic references in titles were easy to annotate. Many geographic locations were mentioned in the full title statement, sometimes in the title, sometimes in the statement of responsibility. In those cases, the annotation was simple. In some cases, the geographic locations were referenced more implicitly, either through the use of nicknames or specific known locations within a city. As shown in table 1 and table 2, an explicit reference to Chicago is as positive a mention as a mention of the windy city, and both are annotated for region, nation, state and city, accordingly.

TITLE	Region	Nation	State/Subregion	City
Future of Chicago : a blueprint for political change / by Dick Simpson.	North America	USA	Illinois	Chicago

Table 1: Explicit Geographic Reference in Title

TITLE	Region	Nation	State/Subregion	City
When architecture meets activism : the transformative experience of Hank Williams village in the windy city / Roger Guy.	North America	USA	Illinois	Chicago

Table 2: Implicit Geographic Reference in Title

After annotating the title, a similar process was applied to the annotation of the Table of Contents field as seen in tables 3 and 4 which illustrate explicit and implicit references. Many of the volumes from the dataset featured information related to the individual chapters, primarily in edited books. 831 volumes had content in the Table of Contents field, and these were annotated similarly to the title field. For these, all geographic references were annotated, including each chapter. In many cases, this led to volumes being annotated with multiple geographic terms in each category. Implicit references were double checked against other evidence to best clarify any ambiguous details.

TOC	Region	Nation	State/Subregion	City
Atlanta, Georgia : collaboration addresses regional concerns -- Austin, Texas : smart growth zones direct growth, spur revitalization	North America	USA	Georgia; Texas	Atlanta; Austin

Table 3: Explicit Geographic Reference in Table of Contents

TOC	Region	Nation	State/Subregion	City
Brilliant corners -- Magnificent spectacle -- The new spaces of Times Square	North America	USA	New York	New York City

Table 4: Implicit Geographic Reference in Table of Contents

The next step of annotation was the examination of the notes field. Only 205 volumes had information in the notes field, but in some cases, the information was useful for annotation. The example in table 5 provides a good example of the clarifying information that was available in some cases. While the volume annotated was titled *Cities are People*, the notes field gave geographic details unavailable in the ambiguous title.

Notes	Region	Nation	State/Subregion	City
Traces the development of such cities as Rome and London, Philadelphia, Boston and other United States cities;	Europe; North America	Italy; UK; USA	England; Pennsylvania; Massachusetts	Rome; London; Philadelphia; Boston

Table 5: Geographic Reference in Notes

The final step of annotation involved the examination of Library of Congress Subject Headings as seen in Table 6 below. Unlike the other categories of annotation, the controlled vocabulary of subject headings meant that only explicit references were annotated. Subject headings were a rich source of geographic information, though inconsistently applied. Subject headings contained geographic information most frequently at larger scales, though many included information at the state or city level.

Subject_headings	Region	Nation	State/Subregion	City
City planning New York (State) New York. Regional planning New York Metropolitan Area.	North America	USA	New York	New York City

Table 6: Geographic Reference in Subject Headings

Once each volume was evaluated for geographic content, volumes that included information for multiple geographic locations were separated into multiple rows, to enable mapping of individual locations. In total, the annotated results by step resulted in the following numbers of results. Table 7 gathers together the overall number of volumes from the full dataset through the various annotation steps.

Annotation Step	Total Results	Annotated by Nation	Annotated by City
Full Dataset	6343	--	--
Deduplication	5435	--	--
Title	5435	1795	654
Table of Contents	5435	1960	798
Notes	5435	1973	807
Subject Headings	5435	3318	1115
Divided into rows	5435	4743	2047

Table 7: Annotation Process Numbers

3.3. GIS Preparation and Analysis

Once annotated, the dataset was separated into two files for mapping purposes. One file consisted of every entry that had data at the nation level, the other consisted of every entry that had data at the city level. Once separated, additional columns were added to add latitude and longitude information for each entry. These data were drawn from Wikimedia's Geohack tool, when that data was available, when it was not, Google Maps was used to identify location data. In several cases, the geographic center of a nation was at a location that did not exist within the land borders of the nation, for example, the Philippines. In these cases, Google Maps was used to determine an alternate geographic data point that aligned with land borders to fit within the national shapefiles.

These spreadsheets were converted to csv files, then imported into ArcMap within ArcGIS suite version 10.6. For global analysis, the Countries WGS1984 shapefile was used¹. For analysis of North American states and provinces, data was downloaded from GADM.org and a combined shapefile layer was built from the states and provinces of the USA, Canada and Mexico. These files were combined to create multiple maps, visualizing the geographic content of the collection using multiple tools.

4. Results

This analysis provided several results that addressed the central questions of the study. The first map produced for analysis was a proof of concept map, a test of whether it was possible to produce a visualization using geographic data produced from the qualitative content of the collection. Figure 1 shows that

¹ https://hub.arcgis.com/datasets/a21fdb46d23e4ef896f31475217cbb08_1

proof of concept, a map of the cities that have at least one entry in a volume within the collection. The production of this visualization was additionally used to check for transcription errors in the dataset and several errors were corrected for this and later visualizations.

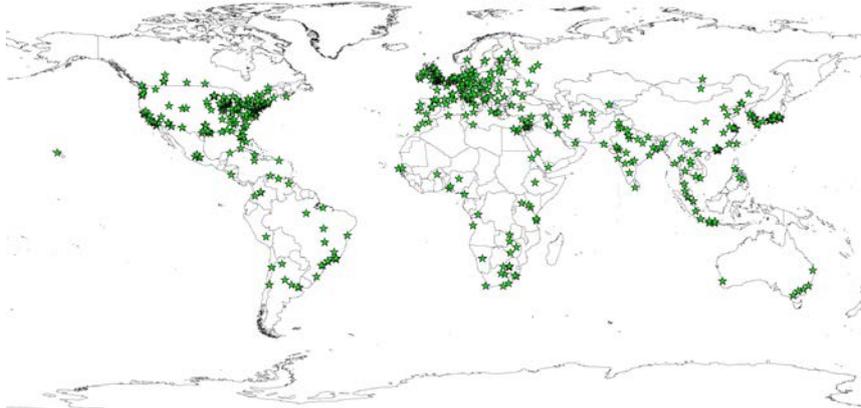


Figure 1: A map of cities in the world

The second map produced for analysis addressed a deeper question, whether it was possible to transform that data into a visualization that would be useful for collection development. Figure 2 approaches that question with a map that examines the dataset detailing countries using a choropleth map. Intersections between the points representing data on individual volumes and the shapefiles of countries were tracked. These intersections were counted to evaluate how many volumes had information on each country.

This visualization provided useful information to guide collection development. While the visualization shows a number of countries with a significant number of volumes, one of the most evident aspects of the visualization comes in terms of the visible gaps, countries that have zero volumes. These gaps are worthy of attention in collection development, each one a question worthy of address, should an urban planning collection have material from each country?

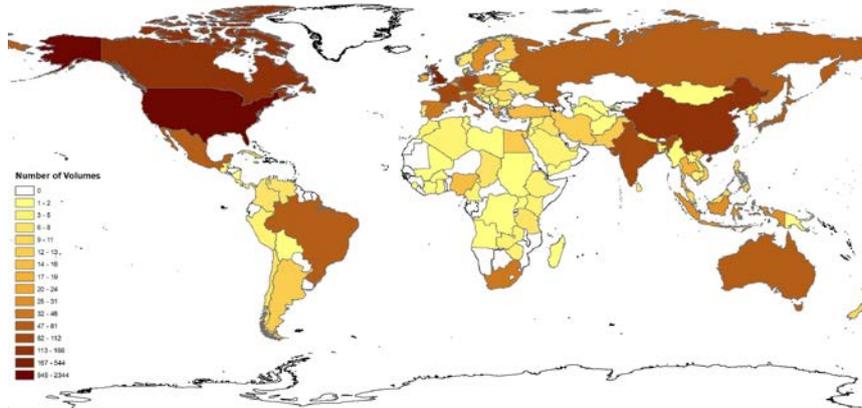


Figure 2: A choropleth map of world countries analyzed by number of volumes

One advantage of ArcGIS as an analytical tool is that information can be visualized at scale. To demonstrate that, a map was produced analyzing the same dataset at a different scale to get different information. Figure 3 is a similar map to figure 2, but instead of focusing on countries around the world, it focuses on the states and provinces of the United States, Mexico and Canada, looking at how many volumes were available for each individual state or province. This map showed gaps across the continent, allowing an examination of such information locally. However, this map also showed information about the local strengths of the collection, the states that had large collections standing out from others.

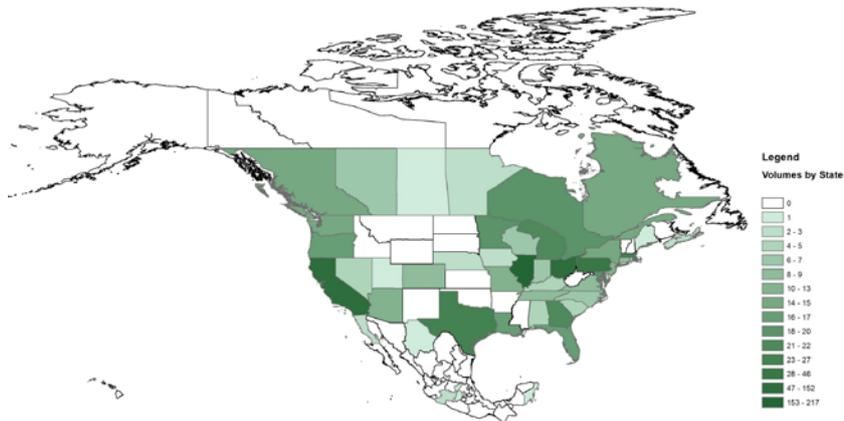


Figure 3: A choropleth map of North American states and provinces analyzed by number of volumes

GIS visualizations can provide larger scale information, as well. The first visualization is merely a map of individual cities, the second and third provide information about the analysis by country, which allows for specific country based decisions, but GIS can be used for larger scale collection strategy as well. To explore this, a map was produced to discern patterns outside the boundaries of individual countries. Figure 4 is a heatmap, showing cities data mapped across the world to determine areas where the collection included more data.

This visualization shows zones of strength and weakness, showing the areas where the collection is strongest, but also areas of general weakness. The wide empty areas of South America, Central Africa and Central Asia when compared with the most concentrated focus on the Eastern United States and Northwestern Europe show a clear bias towards material from those areas. This map shows that the results are concentrated most around English speaking areas, with concentrations in the United States, the UK, Anglophone Africa and South Asia. As the majority of the collection at the University is in English, this partially shows the linguistic bias of the collection. This map also raises an important question about the priorities of the collection and whether it serves to perpetuate a first world/third world mentality. Not every question that spurs collection decision making is a comfortable one, but reflection is important nonetheless.

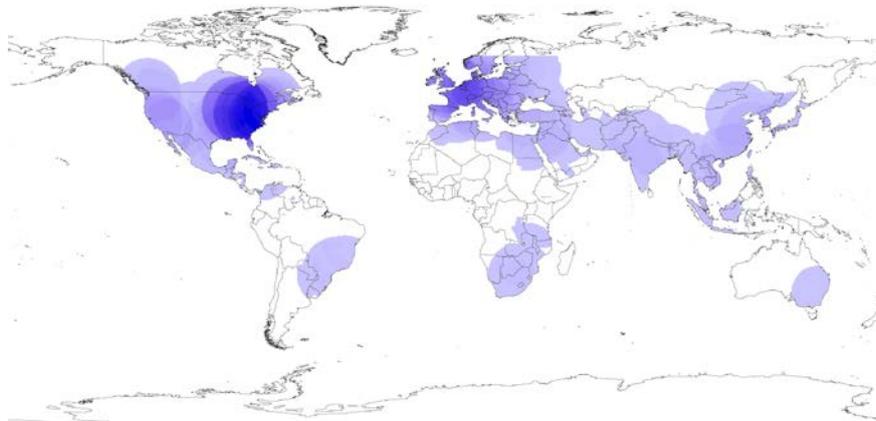


Figure 4: A heat map of world cities

At their heart, these maps all show that the qualitative content within catalog records can be transformed through annotation and technology into data visualizations of the collection of material. The source for these is not explicitly data, but through annotation, it could be converted to data that could be used in visualization.

5. Discussion

Once these visualizations were made, several actions were taken to test the applicability of the information suggested by the maps to library collection management decisions. These decisions show that the visualizations allow for an efficient and usable form of analysis.

One of the ways in which the visualizations proved useful for collection management involved the filling of gaps in the collection. The map from Figure 2 showed that there were a number of countries that did not have any representation within the collection. Materials were ordered for the collection within the call number range of the survey to increase the general coverage of the collection. The following countries had materials available to be added.

HT Volumes Ordered

Botswana
Central African Republic
Ecuador
Iceland (2 volumes ordered)
Mozambique
Niger
Oman
Rwanda
Senegal

For countries that did not have readily available content within the range, a wish list was put together to allow the setting of future ordering priorities to fill gaps. This list also included materials at the State and Province level to fill gaps there as the collections budget allowed.

The map in Figure 3 was also used to highlight areas of the collection that were very well represented within the collection, specifically Illinois, California, Pennsylvania and Texas and target those for selective weeding. A weeding list was created of books from those states for examination based on circulation and year, with a plan to streamline the collection to better suit user needs. This visualization demonstrated that localized strengths can be observed and those strengths can be honed.

These maps also allowed direct coordination with other collection priorities related to curriculum and student demographics. One important aspect of the Urban Planning curriculum at the University of Illinois at Chicago where this study was done is that many of the classes involve projects that directly look at Chicago and its neighborhoods. It makes sense from a collections point of view to coordinate the collection with that curriculum, and having the visualization allows that comparison. Another aspect of the Urban Planning program involves a large number of international graduate students, many of whom study aspects of international urban planning. Visualizations of international subject content

can allow the library to better support these students and their projects and coordinate collection priorities accordingly.

The other priority highlighted in these maps is that of social justice. Figure 4 shows a general bias towards Western Europe and the United States, with significant weakness in the overall collections in some of the poorest areas of the world. Collection development is about more than discrete actions, it is also about philosophy. A principle of balance between the West and the East, between the North and the South, can be part of a collection development philosophy, and a visualization can help track the issues at its heart.

Overall, these actions, ordering materials to fill gaps, targeting weeding to sections that can bear it, coordinating collection priorities between the subject content of the collection and that of the curriculum, and adapting the collection development philosophy to address global biases, show that visualizations of this data can guide collection decisions. While measuring the numerical impact of specific collection decisions is difficult, these are tangible actions that were enabled by the process of transforming this content into mappable data and producing data visualizations.

6. Conclusions

Based on the process and the outcomes, there are a number of relevant takeaways. For future research, this process could be streamlined in ways that allows the application to other collections more efficiently. A follow-up study to this one will make use of automated text mining to improve the annotation process and to increase replicability.

One obstacle to the process was a lack of subject headings covering geographic content at very specific detail. There are two dimensions related to this, subject headings for table of contents related information and subject headings and geographic scale. For the first of these, Library of Congress cataloging guidelines suggest that subject headings only be applied for the overall volume rather than individual chapters. This limits findability for chapters that have specific geographic content that is not present in an entire volume. For the second, there is limited guidance in the Library of Congress about when to apply geographic terms and at what scale. (Library of Congress, 2013). This guidance limits the number of city specific subject headings that are applied, especially to edited volumes, which restricts findability. As seen in the literature review, geographic terms are a priority for many users, and the lack hampers findability. This issue was identified well by Buckland “Library catalogs are well-designed for searching by author, by title, and by topic, but not for searching by place....One can search for place names in titles or in subject headings, but the geographical headings and the geographical subdivisions of the Library of Congress Subject Headings (LCSH) tend to be political jurisdictions--primarily countries, states, and cities.(Buckland & Lancaster, 2005) ” Additionally, better data can lead to better analysis. By increasing the

application of specific geographic subject terms, the collection could be better analyzed at scale, and improved visualizations could be produced.

Beyond this specific application, a similar process could be extended to qualitative content from other disciplines with other visualization tools. A history collection could be annotated by timeline and visualized for era coverage, a literary collection could be examined ethnographically by the annotation of author data with geographic content to help understand geographic bias in authorship, an anthropology collection could be annotated based on the four fields and visualized to assess balance between fields.

At its most fundamental level, this process is about increasing librarian knowledge of a collection. Understanding a collection is important, it is impossible to make strategic decisions without understanding, and the process of annotation and visualization ultimately makes a collection more understandable. The act of transforming this qualitative content into trackable data makes a collection much more understandable. With greater automation and improved content for annotation, this process can help solve the challenges facing library science and open up qualitative content for more scientific analysis.

References

- Abresch, J. (2008). *Integrating geographic information systems into library services: A guide for academic libraries*. Hershey: Information Science Pub. doi:10.4018/978-1-59904-726-3
- Bishop, B. W., & Mandel, L. H. (2010). Utilizing geographic information systems (GIS) in library research. *Library Hi Tech*, 28(4), 536-547. doi://dx.doi.org/10.1108/07378831011096213
- Blake, J. C., & Schleper, S. P. (2004). From data to decisions: Using surveys and statistics to make collection management decisions doi://doi.org/10.1016/j.lcats.2004.09.002
- Buckland, M., & Lancaster, L. (2005). Combining place, time, and topic: The electronic cultural atlas initiative. *D-Lib Magazine*, 10(5) Retrieved from <http://www.dlib.org/dlib/may04/buckland/05buckland.html>
- Clough, P., Tang, J., Hall, M. M., & Warner, A. (2011). Linking archival data to location: A case study at the UK national archives. *Aslib Proceedings*, 63(2/3), 127-147. doi://dx.doi.org/10.1108/00012531111135628
- Finch, J. L., & Flenner, A. R. (2016). Using data visualization to examine an academic library collection. *College & Research Libraries*, 77(6), 765-778. doi:10.5860/crl.77.6.765
- Fry, B. (2008). *Visualizing data* (1st ed.). Sebastopol, CA: O'Reilly Media, Inc. Retrieved from <https://learning.oreilly.com/library/view/visualizing-data/9780596514556/>
- Haren, S. M. (2014). Data visualization as a tool for collection assessment: Mapping the Latin American studies collection at University of California, Riverside. *Library Collections, Acquisitions, & Technical Services*, 38(3), 70-81. doi:10.1080/14649055.2015.1059219
- Jones, R., Zhang, W. V., Rey, B., Jhala, P., & Stipp, E. (2008). Geographic intention and modification in web search. *International Journal of Geographical Information Science*, 22(3), 229-246. doi:10.1080/13658810701626186

- Kilb, M., & Jansen, M. (2016). Visualizing collections data: Why pie charts aren't always the answer. *Serials Review*, 42(3), 192-200. doi:10.1080/00987913.2016.1207479
- Library of Congress. (2013). Geographic subdivision to the city level H832. Retrieved from <http://www.loc.gov/aba/publications/FreeSHM/H0832.pdf>
- Ottensmann, J. R. (1997). Using geographic information systems to analyze library utilization. *Library Quarterly*, 67(1), 24. Retrieved from <http://proxy.cc.uic.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=lxh&AN=9708106250>
- Polley, D. E. (2016). Visualizing the topical coverage of an institutional repository with VOSviewer. In L. Magnuson (Ed.), *Data visualization: A guide to visual storytelling for libraries* (pp. 111-125). New York, NY: Rowman & Littlefield.
- Pournaghi, R. (2017). GIS as a supporting instrument for making decisions about the library sources collection management. *Collection Building*, 36(1), 11-19. doi:10.1108/CB-06-2016-0014
- Renteria-Agualimpia, W., Lopez-Pellicer, F., Lacasta, J., Zarazaga-Soria, F., & Muro-Medrano, P. (2016). Improving the geospatial consistency of digital libraries metadata. *Journal of Information Science*, 42(4), 507-523. doi:10.1177/0165551515597364
- Sanderson, M., & Han, Y. (2007). Search words and geography. *Proceedings of the 4th ACM workshop on geographical information retrieval* (pp. 13-14) ACM. doi:10.1145/1316948.1316952
- Taylor, R. S., & Mitchell, E. (2016). Minding the gap: Utilizing data visualizations for library collections development. In L. Magnuson (Ed.), *Data visualization: A guide to visual storytelling for libraries* (pp. 77-93). New York, NY: Rowman & Littlefield.