A Visualization Tool for Atlas Collection Assessment

A Visualization Tool for Atlas Collection Assessment

A collection assessment tool is described which allows the visualization of a collection in order to determine the depth, scope and needs for the collection. Using an online tool, ManyEyes™, the degree of complexity and the relationships within a collection can be explored and assessed according to the specifics of the collecting institution's own criteria. The tool provides images of the data and is sufficiently flexible to permit multiple analyses as needed by the user.

RUNNING HEADER: Visualization Tool for Atlas Collection Assessment

A Visualization Tool for Atlas Collection Assessment

Visualization tools have become increasingly common in the past ten years for a variety of uses, both in library science and other information science disciplines. These tools are means for turning large amounts of data into pictures, charts, or other images. These images increase the understanding of the user by reducing the information load from multiple pages (or screens) of tabular data into images that highlight the most important information and relationships among the elements. As Carr has noted, "Applying visual analytics to a mountain of data is one of the fastest, easiest, and most comprehensive ways of culling that data to a more manageable, mole-hilled size set".[1]

Geospatial materials such as maps and atlases, for instance, seem especially apt for visualization tools for a number of reasons. Primary is the fact that maps and atlases are formats that convey multidimensional information. They not only tell us where something is in space, they also tell us information about that space such as the population, economic status, geology, etc. It is difficult to get a good understanding of a collection of spatial materials when they are presented in a tabular form or in a listing format partly because this format is essentially unidimensional. A list or table is organized either by objects (geographic places) or by topical information. Thus the materials are ordered from top to bottom either in an order related to the space (e.g., by continents, then countries within the continent, then divisions within the country, etc.) or in order of the type of information (e.g., economic-related maps, then geology, then population, etc.). The full nature of a collection is obscured because one does not have a good feel for all of the relationships among the materials in the collection.

Whether the librarian who seeks to understand their collection is new to the job or has been in their position for a number of years, it is important to understand the collection for which they are stewards. What is proposed in this paper is a description of an approach that allows the capacity for "zooming in" on particular areas of the collection of atlases for a closer inspection. A web site called ManyEyes™ has a variety of visualization tools that enable the user to assess a collection of information. With such a tool, a close inspection allows for an evaluation of the materials included in the collection. Because the visualization approach used is essentially multidimensional, it will allow a more complete understanding of the relationships in the collection, both geographic and topical. One can see for instance that a collection has linguistic atlases in Germany, Great Britain, but does not have them in other countries.

LITERATURE REVIEW

How can one determine if a collection of atlases is both in the guidelines of the collection development policy of an institution and at the same time is recent enough that it is a useful resource? Perry and Weber observed that "it has become near impossible … to evaluate the effectiveness and the adequacy of library collections".[2] Their comment is in response to the many changes in the nature of collections in the early 21st century, changes such as electronic databases and journals, increased Internet use, and agreements within consortia.

Change is also a characteristic that can make the situation more complicated in geographic information collections. For example, maps and atlases from twenty years ago are insufficient to

reflect changes in Eastern Europe and the break up of the Soviet Union. These are not the kinds of changes that user statistics, such as circulation numbers, can readily address.

Borin and Yi have said that, "the literature on evaluation can be grouped into two camps – traditional (criteria based) and new (usage based)."[3] Various approaches have been suggested for collection comparison and assessment within the criteria based methods. For instance, a criteria based assessment of a collection would be to compare it to a standard bibliography. A standard bibliography is a listing of materials that are considered to be the basic, essential items for a collection. There are some standard items that Larsgaard recommends for general-reference world atlases and she says that there are bibliographies of state atlases though the ones mentioned are no later than 1988.[4] As to standard lists or bibliographies that could be used for topical areas or for specific disciplines, a general search has not yielded any standard listing of recent date in WorldCat.

To determine what is the appropriate standard listing with which a particular collection should be compared is then difficult. As we have seen there is a paucity of up-to-date standard bibliographies that cover the range of geographies from world to local atlases. Collections of maps and atlases in a particular library are strongly influenced by the geographic location of the holding institution, the collection development policies of the institution, and the specifics of the academic curriculum, in the case of colleges and universities. Finding a comparable library or set of libraries which have the same general characteristics (size, type of locale, public vs. private, discipline strengths) to make a peer comparison can then be difficult as well.

WorldCat Collection Analysis (WCA) can provide a comparison listing of holdings based on the Library of Congress (LC) classification number. The WCA is an online program that allows the user to create reports of frequencies with which certain LC ranges of items appear in WorldCat. The reports can be across the whole of the WorldCat data base or include sets of institutions that are of interest to the user. In addition the frequencies can be subdivided by such categories as format, language, and publication date. If one were interested in the atlas collection for example, a listing of the LC "G" series from G1000 to G3171 materials (which includes all of the materials classified as atlases of various locations) could be produced and compared to the holdings in the atlas collection in the whole of the World Cat database and/or with peer institution groups. This can generate a set of reports about holdings in general geographic regions (Asia, North America, etc) which can be broken down further by language, format or audience. Additionally once a level of specific geography, format, and language have been reached to suit the purpose of the analysis, the listing of titles can be seen. Circulation statistics can also be generated to add further information to the report.

Another usage-based assessment method is one that examines borrowing and lending patterns based on the institutions' library records. User-centered measures such as circulation data are a primary approach recommended for collection evaluation (American Library Association).[5] As Agee points out, "Most online management systems collect circulation data that may be organized in report form to provide frequency of individual title or classification-area loan information."[6] However, geospatial materials, such as atlases, are likely to be used as a

reference material in a library and are often not borrowed from the library.[a] So these usage records are not going to reflect actual usage and thus are not especially helpful for determining collection needs.

Both the traditional criteria-based and the newer usage-based assessments are essentially ordered lists. With the standard lists or bibliographies there is a listing of suggested or recommended items; in the usage measures we have lists of items used and not used. The lists from either of these do not account for the multidimensional ways that the atlases are related. This research proposes the use of a visualization tool which charts the atlases using both their geography and their information topic.

It is necessary to understand a visualization tool because the word visualization can be understood on many levels. Charts such as pie charts and bar charts which show percentages and frequencies are forms of visualization, which is turning data into an image. Libraries are using forms of visualization in information retrieval; the AquaBrowser® is a visualization of a catalog search. Likewise, Internet search engines such as Quintura (www.quintura.com) use visualization in the form of a word cloud to help you narrow down or redirect the results. Other instances of visualization range from mundane uses such as floor plans of a library for assisting patrons to "node and link" network diagrams that represent the collaboration among groups of researchers based on co-authorship on books and articles. The Harvard Catalyst Profile (http://catalyst.harvard.edu/people.html) provides these types of visualizations for researchers.

METHOD

The purpose of this research is to explore the use of a particular visualization tool, Phrase Net from ManyEyes™ in providing information about a collection. The end result should allow the user to better understand the materials that are in their institution's collection and suggest areas in which the collection can be improved. To start the collection assessment it was decided that this initial test of a procedure would begin with a limited collection that could be handled readily. The atlas collection at the University of Illinois at Chicago (UIC) consists of 2,013 titles that have been entered into the catalog. A complete listing of all atlas materials catalogued for the University Library at UIC was obtained from the catalog records in Voyager. These records were downloaded to an Excel file to make manipulation and formatting easier. Each record consisted of the local identification number, the LC call number, the title, and the MARC 500 field (Notes), 650 field (Topical term), and 651 field (Geographic name). The specific titles in the collection were used to provide a more complete description of the latter MARC fields. While the 650 and 651 fields should be sufficient data to perform the analysis, it was found that missing information was of such an extent that neither of these fields alone could be used for the test data. It would have been optimal to use the 650 and 651 fields as they represent a controlled vocabulary and would have made the cleanest analysis of the data.

After checking that all records were atlases based on both a call number and a title check, 1,918 records remained after the removal of 86 items that were irrelevant or incorrectly classified. These 1,918 atlases were used to create atlas-related terms from the title and MARC headings.

[a] By way of example, there were only 48 atlases checked out in the calendar year 2010 at the University of Illinois at Chicago, less than one per week.

The following procedures were used to prepare the data for use in the analysis.

- The data were cleaned of stop terms including "the", "of", "and", "in" and other such grammatical terms that were not essential to organizing the information.
- Words "atlas" and "map" and their related terms such as "mapping" and "atlases" were removed because it was known that the materials were atlases and had maps so these terms provided no new information.
- Certain terms that were deemed to represent a single concept in multiple words (e.g., Great Britain) were edited to include an underline between the separate words (Great_Britain) in order to keep them as a single term in the data. Initial runs had shown that the individual terms (Great and Britain) occur close together in the analysis. These joined terms were limited to geographic names. Other terms such as "city" and "transportation" were retained as individual terms so that they might be seen in relation to topical terms and geographic entities which were related to them by frequency.

The cleaned data file was saved as a text file and was used in the next step of the analysis. In terms of time, the author took approximately three hours to create the data set from receiving it as a report from the cataloging system report to being ready to enter the data file into the online program.

The data were analyzed using the online service, IBM® Many Eyes™ (http://manyeyes.alphaworks.ibm.com/manyeyes/) which provides several types of visualization tools including Phrase Net which is a tool for looking at relationships in text. The website has options to upload a data set or to use one that has been uploaded by someone else. With the data set you may create a number of types of visualizations, including traditional data charts such as pie, bar and line graphs. It also has US county and state maps as well as world maps in which data can be shown such as income levels. Finally there are a set of text analysis diagrams, including word clouds and tags, word trees and Phrase Net.

The website describes Phrase Net: "A phrase net diagrams the relationships between different words used in a text. It uses a simple form of pattern matching to provide multiple views of the concepts contained in a book, speech, or poem." The tool is considered to be somewhere between a "tag cloud" and a "word tree". The relationship between terms that are to be diagramed may be defined in a number of ways such as words connected by "and", "is", "at", or other customized relationships. The analysis for the data here required a space between the words. Data from the text data file are pasted into a data box in Many Eyes™ and saved online. These data can then be visualized using any of the appropriate tools.

Another option in the Phrase Net is to show how many individual terms or words are in the diagram. In this analysis several levels of words were chosen including the ten most frequently repeated words, then the 25 most frequent, then the 50, and finally the 75 most frequent words. These different numbers of terms provide increasing depth to the analysis and the user is not limited to specific numbers of terms, so if they desired, the analysis could be on the 43 most frequent terms. Overall Phrase Net provides a means to analyze the text from the atlas geographic and topical terms based on the proximity of the geographic name and the topic so that the relationships are maintained in the analysis.

Because this is an open website, the readers can access the data discussed in the article and perform visualizations of whatever type they wish. The data can be found by searching in data sets for atlas and selecting "revised atlas title and subject terms from UIC" and by clicking on the Visualize button to the right hand side of the listing, the selection of a type of tool can be made.

RESULTS

After cleaning and entering the data into Many Eyes, the Phrase Net tool was then used to create four initial sets of images of relationships. These images are shown in Figures 1 – 4. In Phrase Net diagrams there are a few conventions that make interpretation of the diagram easier. First, the words are connected by arrows if they are related by a space between them in the data file. So only first-order connections (side-by-side terms) are shown with arrows. The size of a word in the diagram (i.e., its font size) is proportional to the number of times it occurs. Larger words represent more frequent use of those words. The arrow between words has varying thickness depending upon how many times those two words were related. The color of a word varies depending upon whether it was more likely to be found in the first or second position. Position is simply determined by which word came first in the pair of words. Given that the words are simply terms in no syntactical order, position does not play as important a part as it would in sentence-based diagrams. The darker the word, the more often it appeared in the first position. Table 1 gives the frequency of the top 25 terms used in the analysis.

[INSERT TABLE 1 HERE]

Figure 1 illustrates the top ten terms and provides a high-level view of the atlas collection. It is clear from this diagram that the atlas collection at UIC is highly oriented towards materials for Illinois and specifically Cook County, where the university is located. Also the words "county", "real", and "property" suggest that the materials are weighted towards items that relate to local property and real estate matters. These indications are in line with the collection development goals for the library in general. The terms "historical", "history" and "geography" especially in relation to "united_states" indicate that there are more than just local or current records in the collection. Even at this high level view, it is clear that we are going to be able to see relationships between geography and topic, and that we are going to have a diversity of levels in geography, from local to national, and so on.

[INSERT FIGURE 1 HERE]

Expanding the PhraseNet diagram to include the top 25 commonly-used words shows that even the local aspects of the collection are not restricted to Cook County, but also include other nearby Illinois counties (e.g., DuPage County) and other aspects of local atlases especially related to outdoors, natural, resources, and recreation (Figure 2). The "history" and "historical" terms with "geography" are now expanded to include both the United States and Europe, particularly British and German materials. There is a sub-branch appearing on the right side which indicates that language and dialects related to the German materials exist. There are also "economic" and "conditions" that connect to "geography". Clearly the local geography is separated from larger areas and the topics for the local vs. the larger areas seem quite different at this point.

[INSERT FIGURE 2 HERE]

Increasing the Phrase net diagram to 50 terms (Figure 3), one can see further expansion of the local area terms such as additional counties (McHenry, Lake, and Kane counties) as well as the city of Chicago. There are also additional countries in Europe (such as Netherlands, Finland, and the Soviet Union) as well as other parts of the world (e.g., Africa, Pacific, Canada, and China) that now appear in the chart. It should be noted that places which one might logically think would be closely related geographically such as England, British, and Great Britain, are not directly connected to each other in the diagram. This suggests one of the limitations to the relationship rule used, i.e., the immediate adjacency connection by one space between the terms in the data. Another interesting type of item to come to the fore is dates, such as 1945 and 1800, suggesting something about the recency as well as interests reflected in the collection.

[INSERT FIGURE 3 HERE]

The expansion to 75 terms (Figure 4) becomes almost too complex to elicit helpful data. However, we will examine this diagram more closely using some of the Phrase Net tools. While the drawing may seem complex and a list of an equivalent number of individual terms may be relatively easy to grasp, it should be remembered that this charts represents not just 75 terms. These are the top 75 most frequent terms from a list of thousands of words and phrases, and are representative of many more specific atlases. Also we are not looking just at frequencies here but also at relationships among these terms. It is important then to be able to examine this chart more closely, which we will do in the next chart. In general the same types of additions hold here as have been seen in the previous expansion of terms – more countries and regions along with increased terms that define the types of atlases such as “administrative”, “streets”, and “population”.

[INSERT FIGURE 4 HERE]

Moving the mouse over a term in the screen display will result in a box appearing that indicates the number of occurrences of that term. On the other hand moving the mouse over the arrows provides the number of occurrences of the two terms connected by the arrow as well as the first ten instances of the word pairs (Figure 5). Given the emphasis on the Chicago area in the UIC collection, it is somewhat surprising that the first ten occurrences show a variety of “metropolitan areas” for which there are atlases in the collection. However given knowledge that the University has a College of Urban Planning and Public Affairs, it is more understandable that there would be collections from a number of metropolitan areas. Knowing about the relationship also allows us to begin making use of the chart’s result. While there are a number of metropolitan areas, it would be good to approach the faculty of the College of Urban Planning to determine if there are additional metropolitan area atlases that would be useful to them.

[INSERT FIGURE 5 HERE]

In examining the arrow between housing and Netherlands shown in Figure 6, we see that the first ten occurrences are identical and warrant further examination. Going to the library catalog one

finds a 20 volume atlas of the Netherlands that includes both housing and population data. This points to an issue that must be considered when examining the data and the diagrams. The data from the catalog had shown each of the volumes as a separate entity rather than one title, so we have many instances in the diagrams. While that many volumes do represent a big space in the collection, a larger question is left. Do we want to count it that many times in our data? Multi-volume atlases generally represent a major work on the geography and the topic involved so a decision has to be made by the user of this approach about the extent to which they wish multiple volumes to be included as separate items.

[INSERT FIGURE 6 HERE]

Figure 7 indicates that there is only one occurrence of a pairing between England and Great Britain, a seemingly small number. There are limitations to the relationship definition in this type of visualization. Titles that refer to Great Britain will often not include the specifics of the countries such as England that are included in the broader category of Great Britain. From that point it may be useful to determine if there are differences in the nature or types of atlases that fall under each of the terms. Specific topics may be of more interest in England, Scotland, or Wales than would be in Great Britain, such as an atlas of surnames where there is reason to believe that the surnames are more likely to be found in those specific countries.

[INSERT FIGURE 7 HERE]

In order to maintain the depth of relationships that occur in the 75-item diagram without being overwhelmed by the complexity of the diagram, the PhraseNet service allows zooming to and panning of portions of the diagram. Figure 8 shows an example of a diagram where it has been enlarged in order to look more closely at the lower left quadrant of the diagram shown in Figure 5. In addition to being much easier to read, one can examine some specific relationships. For there is a connection from “1945” to “world” and from “world” to “war” suggesting a portion of the collection that is related to the Second World War. “Poland” has arrows indicating relationships out of this quadrant but when we pan across, we would find they are relationships to “geology” and “historical”. In addition the user may select an area of the diagram by clicking and dragging with the left mouse button and then zooming in on that area.

[INSERT FIGURE 8 HERE]

DISCUSSION AND CONCLUSIONS

The use of freely available tools to assist in collection assessment can be a fruitful procedure. In this case all that is required is a small set of computer tools that are available at virtually all libraries – a spreadsheet program (such as Excel), a text editor like Notepad, and Internet access. Additionally, this approach is relatively simple and straightforward. The librarian works with materials that they are familiar with and that they have some knowledge of. Titles of atlases and the subject headings associated with them are the basic materials used. Finally, the procedure provides information that is not readily available in other assessment procedures; it shows relationships between geographic entities and the topics that are addressed in the atlases. Where there are gaps in a collection there will be like “holes” in the chart. The librarian knows that the

library needs to support a program in a specific discipline and by looking for the types of information or the geographic entity involved, the presence or absence of a type of atlas can be determined.

Because the program is freely available via the Internet, the user can easily manipulate and alter the data set. Multiple runs with corrections and changes to the data set (such as combining terms to more correctly represent a single entity) do not "cost" anything. Instead the user learns from these iterations. For example, the writer, in working with a different set of materials from a collection finding aid found after the first run that the dates (in years) were overwhelming that analysis. When the charts were examined, very few terms beside years were showing up. But one is free to change the data files and make new ways to better represent or analyze the information.

Some of the results were surprising at first glance and needed analysis, such as the non-Chicago "metropolitan areas" in the analysis. But this will serve to make the librarian more aware, not only of the collection, but also of the institution that it serves. In examining the visualizations, the strengths of the collection are fairly obvious; they literally appear in large print. The shortcomings of the collection will require more thoughtful examination. The materials that are not in the chart but should be there or should be more prominent are the areas where the collection needs to be enhanced. This of course requires the librarian to know not just the collection, but also the collection plan of the library and the needs of the predominant disciplines of the institution.

For example, in the 75-word diagram of the UIC collection, many Western European countries have shown up individually as well as both the United States and Canada along with North America. However, very little is seen concerning specific African and Latin American countries. Given both the diverse nature of the UIC student body and some of the academic specialties of the university (such as African-American Studies, Latin American Studies, and Slavic and Baltic Languages and Literature) one would expect to see more than Mexico and Poland in the phrase diagram of the atlases. Clearly this points to collection priorities that need to be addressed. The question is raised, "Can you create a hypothetical, idealized chart? A chart that represents what one wants the collection to look like and that includes all geographic entities in proportion to their importance as well as atlas topics in the areas of importance to the institution's disciplines?" Such a chart is possible but it would require a specific knowledge of not only the geographic entities and the topics to be ideally included but also the proportions of each. In a sense this leads us back to the standard bibliographies and the issues associated with them. Perhaps this is an area to be considered in the future about how we can best assess a collection. How can we readily get the information?

This approach offers a new tool to visualizing collection assessment. Several of its shortcomings have already been noted. First among them is the nature of the relationships diagramed. Whenever relationships require side-by-side (first order) connections, the information is limited to the "neighborhood" that the title words and subject headings occur in. More statistically advanced procedures such as multivariate clustering can provide groupings of materials that are related using many variables which are not limited to adjacent data items. In the long run such techniques may be more amenable to this type of project but they require more complex

procedures to create the visualizations of the results that are as easily understandable as our diagrams of the atlas terms. Kim, Lee, and Park (2009) for instance have used pathfinder scaling, which has nodes and links based on the subject-usage data. However these data are circulation based and are therefore not as satisfactory for materials such as atlases that have a strong reference use component in the library.

Additionally, the testing of this procedure was conducted on a limited sample of related materials. How well this can be used in different materials and with a greater number of items remains to be seen. Given these constraints, this approach seems to provide a basic first step in the assessment of an atlas collection and provides the user with a means of grasping the breadth and depth of their collection. This approach can allow both a broad vision and a more in-depth assessment. As a collection assessment tool, it is a streamlined method for analyzing the strengths and needs of the collection where the scope of a collection can be delimited to a specific type of item.

The use of the visualization procedure in a limited situation does not preclude its use in a number of other types of analyses. For other parts of the library's collection obviously the same type of approach can be used. The nature of its multidimensional analysis and presentation makes it amenable to other types of materials as well. As indicated earlier, it is being tested using materials from finding aids for a manuscript collection where there is no real limitation on the vocabulary as there is within a limited range of the LC numbering system. The intent of that assessment is to determine if the procedure has potential as a visual form of "finding aid". Using the charts of the terms from the finding aid, the expectation is that relationships between different parts of the manuscript collection can be seen that might have been overlooked or at least were hard to find in a listing format. A user could follow connections from the initial vocabulary item(s) of interest via the arrows to other portions of the manuscript collection.

No one procedure – whether for listing or visualizing data – is the only correct approach. Having several means of looking at the data, especially means that are suited to the multiple dimensions of the data, are important for us to advance our understanding of the data. They can provide us with a toolbox of approaches that address not only the question at hand but are flexible enough to be used in a variety of situations bringing additional insights to other data as well.

Notes

The author would like to thank Steve Brantley of the University of Illinois at Chicago Daley Library for bringing the Many Eyes website to the author's attention and encouraging its use.

REFERENCES

[1] Carr, K. 2008. "Techniques for making molehills out of unstructured data mountains". *The Information Management Journal* September/October:43-48.

[2] Perry, S.L. and D.C. Weber. 2001. "Evaluating academic library quality today". *Advances in Librarianship* 25:97-131.

[3] Borin, J and H. Yi. 2008. "Indicators for collection evaluation: a new dimensional framework". *Collection Building* 27(4):136-143.

[4] Larsgaard, Mary Lynette. 1998. *Map Librarianship: An introduction* (3rd ed.). Pp.35-36.

[5] Lockett, Barbara (ed.). 1989. *Guide to the evaluation of library collections*. Chicago: American Library Association.

[6] Agee, Jim. 2005. "Collection evaluation: a foundation for collection development". *Collection Building* 24(3):92-95