View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by ScholarWorks@UA

Biz of Acq — Using Tag Clouds to Visualize Circulation Patterns and Inform Acquisitions

by **H. Caroline Hassler** (Technical Services Librarian, William A. Egan Library, University of Alaska Southeast, 11120 Glacier Highway, Juneau, AK 99801; Phone 001-907-796-6345; Fax 001-907-796-6302) <hchassler@uas.alaska.edu>

Column Editor: **Michelle Flinchbaugh** (Acquisitions Librarian, Albin O. Kuhn Library & Gallery, University of Maryland Baltimore County, 1000 Hilltop Circle, Baltimore, MD 21250; Phone: 410-455-6754; Fax: 410-455-1598) <flinchba@umbc.edu>

Introduction. As Technical Services Librarian, I manage the cataloging, collection development, and acquisitions services for my library. I'm always looking for ways that these functions can work together to support better processes. For example, how can the subject classification information from the catalog record combined with circulation reports improve collection development? This article describes one relatively simple way to accomplish this, using circulation data from the integrated library system (ILS), classification information from the catalog record, and tag cloud generating services freely available on the Web.

Several years ago, while Technical Services Librarian at the **National Institute of Standards and Technology (NIST)**, I analyzed two years of circulation data to produce a report on the top 100 most popular subjects in the library.[1] My goal was to identify subject areas in the library that were heavily used, and where books purchased in these areas were almost guaranteed to circulate. This helped refine our book approval plan, and also provided guidelines for purchasing that increased the likelihood of a positive return on the investment made in a book.

The **NIST** data, covering over 12,500 titles that had been checked out during the study period, was grouped into subject categories using call numbers. Once combined into these broad subjects, a formula that weighted number of items and number of circulations was used to score each subject. The process was time-intensive, and making regular updates was prohibitive.

I was still contemplating this problem when I started my position as Technical Services Librarian at the **University of Alaska Southeast**. The **NIST** study had required many days working with an Excel spreadsheet, primarily because of the large volume of data and the manual review of subject categories. To address the issue of data volume, I developed the concept of the circulation snapshot. Rather than collect information for every book that had circulated over a given period of time, I would try collecting data for one day only. For the circulation snapshot, I picked one day during the busiest part of the academic semester, and reported out from the ILS all the items that had a status of "checked out" on that day. Items were selected for the report based on having a current location of "checked out" in our ILS; the actual date of the check-out transaction was not a factor. The snapshot approach collected a smaller sample than my previous reports, and was therefore easier to work with but more likely to yield unrepresentative results. For instance, the snapshot data was more likely to be skewed by anomalous activity from one or two patrons during the snapshot window. However, I reasoned that if conducted over several semesters, the snapshots might reveal patterns that expressed general truths about the interests and needs of faculty and students at the university.

I began gathering circulation snapshots in the Fall of 2008. I picked a day near the end of the semester when many students were working on research papers and circulation at the library was at its highest point. I ran a report from our ILS that output every item currently checked out, listing for each its title, call number, and total circulations recorded in the ILS. I started running the snapshot report at least once a year, and collected the reports together to analyze once a few years worth of data had accumulated. I was still facing a tedious and lengthy analysis process.

Enter the Tag Cloud. By 2008, tag clouds had become features on many Websites, especially blogs and social networking sites like Flickr. A tag cloud, also known as a word cloud, is a "method to visualize textual data, where the importance of each word in the text is highlighted by its font size, and/or color."[2] In tag clouds, the importance and size of a word is often based solely on its frequency, but it may also be weighted by other factors. I realized the tag cloud might be a tool for basic text analysis that could speed up processing of circulation snapshots. Several tag cloud generators were available on the Web. Each could take large amounts of data and process it in seconds. They required no special programs or training. So I decided to try to use these services to analyze my circulation snapshot data.

The advantages of using tag clouds are speed and visual impact. Hours of time working with a spreadsheet is saved, and the result is a cool graphic. However, there are some trade-offs. Running a quick and dirty analysis requires uniformly truncating the call numbers of circulated items at the first full stop in the call number. Our library uses Library of Congress Classification (LCC), so for example, "QC981.8.C5I47 1998" becomes QC981 before being entered into the tag cloud. The truncation is done easily in Excel using the "Text to Columns" command on the Data tab. However, LC Classification is not designed with a consistent hierarchical numerical notation system, so the results of truncation in LCC vary. Sometimes, many different subjects will be contained within a single number through use of subject cuttering or decimal subdivision after the base number. Sometimes a single subject will be divided within a range of different base numbers. When I did a more manual analysis, I took these situations into account from the outset. When using the faster method with tag clouds, I could only refer back to the ILS reports after the cloud was generated to identify the popular subtopics within a broad base number.

My research has found that there are four tag cloud generating services that work well for circulation data analysis: **Wordle** (*www.wordle.net*), **TagCrowd** (*www.tagcrowd.com*), **Tagxedo** (*http://www.tagxedo.com*), and **Tagul** (*www.tagul.com*). Each one has different strong points that suit different types of analysis. All are free, although Tagul does require creating a user account. Here are my five favorite methods for call number and also title word analysis.

1. Basic Call Number Analysis using Wordle. Wordle is one of the tools I like best for basic call number analysis. This analysis only requires a list of the call numbers for all items that have circulated within whatever period of time you choose, or for all items currently checked out if taking a snapshot. Using Excel's "Text to Columns" command with a period as the delimiter, you can parse the LCC stems (e.g., E99, PS3566, QC981) into a single column separated from the remainder of the call number. Using the basic Wordle Create page, paste the list of LCC stems from the spreadsheet into the Web form and hit Go to generate your tag cloud. Wordle treats each LCC stem as a word. After the cloud has generated, you can choose the maximum number of "words" to show under the Wordle Layout menu, thus allowing you to show the top 25, or top 50 or whatever number you choose, call numbers based on frequency. Here is a tag cloud showing the top 25 call numbers at my library from the Spring 2012 circulation snapshot, created with Wordle.

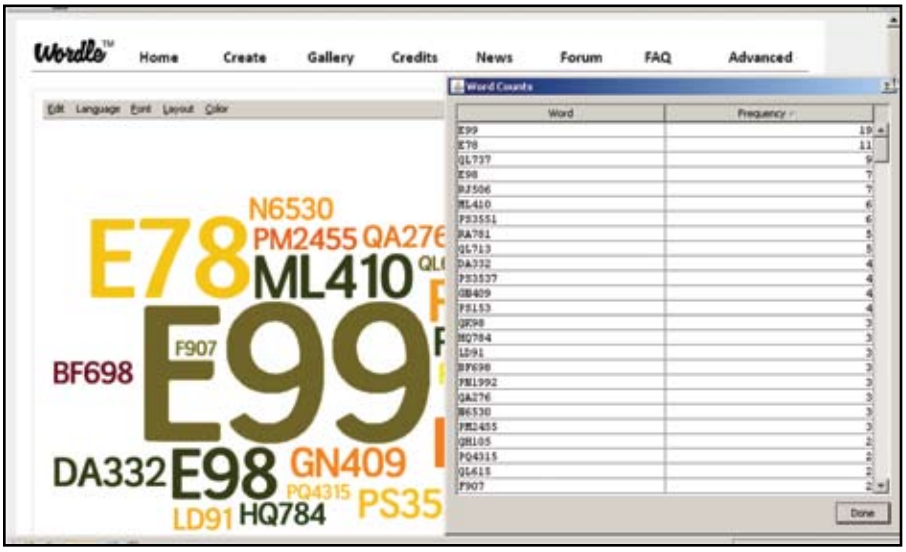
In this example, I used Wordle's options to keep all words horizontal and to prefer

continued on page 00

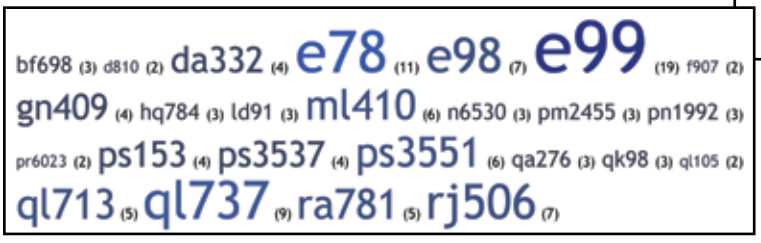
alphabetical order, both available in the Wordle Layout menu. Colors and fonts are easily customizable in Wordle, as they are also in Tagul and Tagxedo.

I have only tried these methods with LCC call numbers, however most of the techniques should work similarly for Dewey Classification. With Dewey numbers in Wordle, be sure to uncheck the "Remove Numbers" option on the Language menu.

One reason I like Wordle is it has a function that displays numerical word counts in a table. You can find the "Show Word Counts" option under the Language menu after you've created your cloud. Although the Java Applet that displays the word counts does not currently allow copying to your computer clipboard (as explained on the Wordle FAQ page), you can capture the table using a screenshot. Here is an example of how Wordle shows word counts in a pop-up Java Applet window. The table is sortable by word or frequency.



2. Basic Call Number Analysis using TagCrowd. TagCrowd is another tag cloud tool that can handle basic call number analysis. Creating a cloud in TagCrowd requires the same type of list as Wordle, and again you simply have to paste your list of call numbers stems into the tool and hit Visualize to create your cloud. TagCrowd also has the option to upload a file rather than pasting. One appealing feature of TagCrowd is that it defaults to presenting the "words" in alphabetical order. You can also choose to show the word count next to each word, using the Options menu. As with Wordle, you can choose to show the maximum number of words for the cloud, allowing you to generate top 25, top 50, or top 100, as you like. Font and color customization in TagCrowd is possible but requires editing HTML code and Cascading Style Sheets (CSS). Here is a tag cloud showing the top 25 call numbers at my library from the Spring 2012 circulation snapshot, created with TagCrowd, with the option to show word frequencies turned on and with the default font and color scheme.



3. Weighted Call Number Analysis using Wordle. If your ILS reporting capabilities include the option to add the total number of circulations in each item's history to your report, you can generate a tag cloud with Wordle that will weight each call number by the total number of circulations over time. In the resulting tag cloud, tag size reflects total circulations of items going back as far historically as the data in your ILS. This introduces a longitudinal dimension to even a basic circulation snapshot capturing checkouts on a single day. The weighted analysis requires a spreadsheet report with a column for the call number stem and a column for the number of total circulations for each item. You can create a weighted tag cloud using the Advanced tab on the main Wordle menu. Wordle will accept a weighted list in the format *word: number*, where the "word" is the call number stem and number is the number of circulations recorded for the item. For example:

E99:21
PS3566:12
QC25:3

To arrive at this format, insert a column filled with the colon character between your call number stem column and your total circulations column, then use Excel's Concatenate function to combine the three columns into one. Paste this list into the weighted word input box on the Wordle Advanced page and hit Go. Important: after your cloud has generated, to get multiple instances of the same call number combined into one with weights added together, turn on the "Make All Words Upper Case" option on the Language menu. Unfortunately, the "Show Word Counts" feature is not available with tag clouds created on the Wordle Advanced page.

4. LCC Top Level Class Analysis using Tagxedo. Tagxedo is a tag cloud generator that is particularly handy for looking at the top level LCC class break-down of your circulated items (e.g., E, PS, QC, etc.). It's handy because Tagxedo's default behavior is to strip all numbers off the call number, leaving only the introductory letters. You can use the same list of truncated call numbers for circulated items as you use for basic analysis in Wordle or TagCrowd. When you load this list into Tagxedo, you get back a cloud that only reflects the frequency of the top level LCC classes. Here is a tag cloud showing the top 50 LCC call numbers classes at my library from the Spring 2012 circulation snapshot, created with Tagxedo.

This example shows that E, QL, PS, and PR classes are the most popular broad areas in our collection (American History, Zoology, American Literature, and English Literature, respectively). Tagxedo offers a high degree of customization options, including the choice of over a hundred different shapes for your cloud. This example was created using the classic horizontal cloud option.

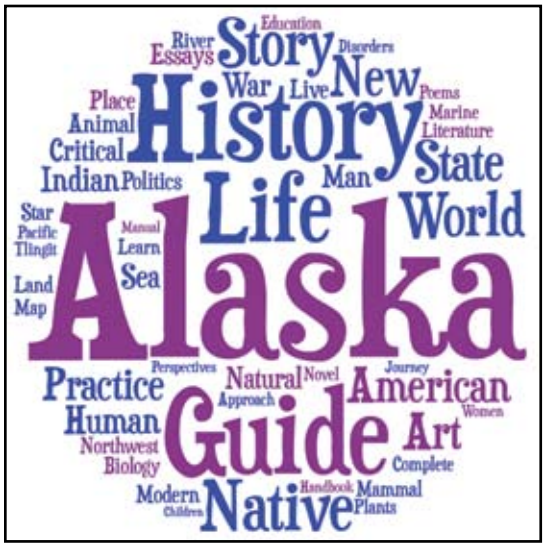
5. Tag clouds from title words using Tagul. Another interesting analysis is made by generating a tag cloud using the words in the titles of circulated items. I prefer Tagul for this type of cloud because of its

continued on page 00

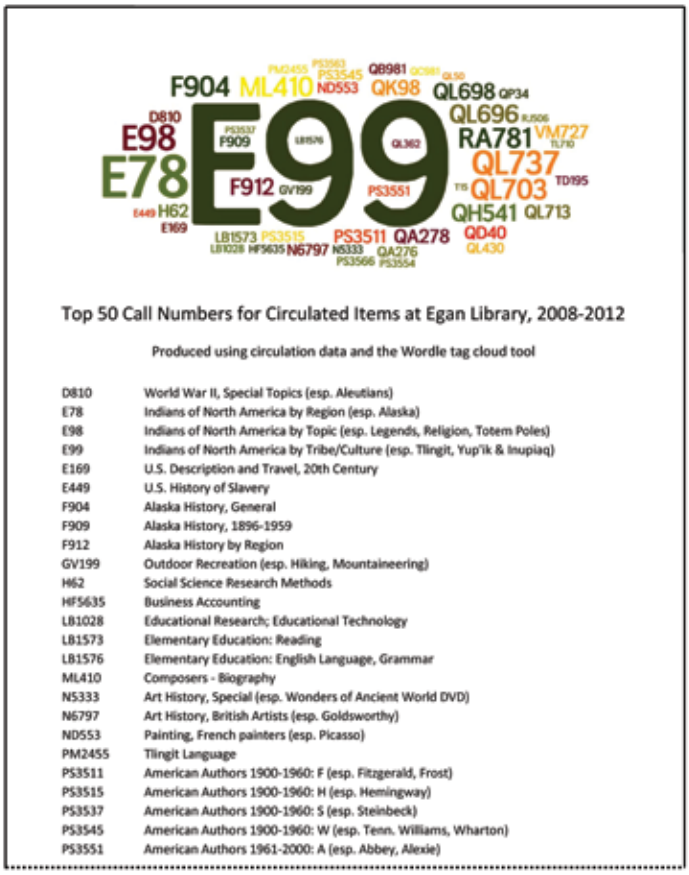
language customization features and overall great looking results. To get the data for this tag cloud, I used the same report from the ILS, which included the MARC field 245 (Title Statement) of each item currently circulating. Ideally for this report, you only want to include field 245 subfields $a and $b with title words, and exclude statement of responsibility data in subfield $c. If possible in your ILS, limit the output of your report to 245 subfields $a and $b only. If not possible, use the "Text to Columns" command in Excel to split the 245 field into two columns using the forward slash character as the delimiter. Most catalog records created after the adoption of AACR2 in 1978 will use the forward slash to separate title information from the statement of responsibility.

While any tag cloud generator will handle a list of title words, what I like about Tagul for title word clouds is its very transparent and customizable handling of stopwords and its ability to group similar words. Stopwords are words that are ignored by the program and not included in the cloud. I found Wordle did not make their list of stopwords readily available and TagCrowd's stoplist for English excluded words that I would not have chosen such as "men" (but not "women"). Tagul provides easily reviewed lists of stopwords for multiple languages, and allows you to remove or add custom words to the stoplist. If your list of titles includes words that are common but not very descriptive, such as "selected," it's easy to exclude them with Tagul and can even be done after you've first produced the cloud. Tagul also uses a word stemming algorithm that combines words with the same root together as a single tag. For example, Tagul combines "Alaska" and "Alaska's" into the same "Alaska" tag.

Here is a tag cloud created from the titles of circulated books at my library from the Spring 2012 snapshot, generated with Tagul.



This cloud beautifully expresses what is true and special about our university: there is a keen focus on our environment, Alaska, and its natural and cultural resources. This type of cloud is an effective graphic to show on your library's Facebook or Web page. It's a way to show that the library is tuned in to its users, and that the collection reflects the community. If your library participates in **ALA's** Library Snapshot Day, try making a cloud from the titles of books checked out on that day. We have done this for our Snapshot Day and posted the cloud alongside photographs taken in the library.



Using tag clouds to inform acquisitions. With the very first circulation snapshot cloud I made, I found having the visual image of what was popular very useful to me in collection development. Of course it was not the basis for every selection, but it did make certain titles stand out as ones that were almost certain to be used by our patrons. After I had four years of snapshots collected, I combined the weighted call number data for all years into a cumulative report. Using the Wordle Advanced page, I generated a new cloud with a weighted call number analysis, and from the result produced a list of the top 50 call numbers for circulated items from 2008 to 2012. I added a textual description of the subject alongside the call number and distributed the cloud and the list to everyone involved in collection development. I kept the list of call numbers in classification order so it would be easy for selectors to find their areas. When a base call number was the type that covered multiple topics through subject cuttering, I went back to my ILS reports, pulled out the most prominent subtopics, and listed those in parentheses after the main subject description. Rather than provide a numerical ranking, I let the cloud represent how each subject fared in relation to the others. See a portion of that list in the screen shot shown above.

The list shows that many topics of high interest are those with special meaning and importance here in Southeast Alaska: Tlingit language, Alaskan history, and the cultures of Alaska Native peoples. Some of the popular areas reflect important programs at our school, such as Elementary Education, and Outdoor Education. Others match up with popular courses, such as **Hemingway**, and World War II, both topics of special seminars taught here. One surprise was the popularity of composers' biographies. Our music program is relatively small yet this area was consistently high-use across all years of study. Whether this pattern represents the circulation habits of one or two avid music fans or a general interest in our community, I can't say. However, I did take note and recently added several new biographies to the Music section. They did quickly find readers.

I have also used this list of Top Fifty Call Numbers to set up notification reports with our book vendor. With our vendor, **YBP Library Services**, I can set up notification reports based on call number classification. I now receive a regular report listing newly published works in many of these areas.

More Tag Cloud Options. Now that I have figured out how to use tag cloud generators as simple and quick text analyzers, I have ideas for other projects. I've already run the same type of call number analysis on our **ebrary** eBook collection, using title, call number and usage data pulled from the **ebrary** administration site. Using the Wordle Advanced page, I input a list of the LC call numbers of accessed eBooks, this time weighted by user sessions. You can also input just the list of LC call numbers (without user sessions) if you want to consider only how many eBooks in a subject are used without regard to how many sessions for each eBook.

Tag clouds might also be made using the subject headings of circulated items. Both

continued on page 00

Biz of Acq
from page 00

Wordle and Tagul have a function to keep words together as phrases by substituting a tilde (~) for the space between the words. A list of subject headings from checked-out items, formatted to be kept as phrases, would create a cloud that reveals their relative frequency. Or, a list of author names from the MARC field 100, formatted as phrases and weighted by total circulations, could be turned into a tag cloud illuminating the most popular authors in the collection.

Conclusion. Tag clouds based on circulation data may be geeky, but they are also relatively quick and easy, fun, and informative. Like any statistical tool, they have limitations. As I noted when I did my first report on popular call numbers at **NIST**, circulation analysis only tells you something about the collection you already have, and nothing about what you don't. It doesn't say anything about subjects people are looking for but can't find in your library. Therefore, the tools of traditional collection development must still be utilized, including looking at interlibrary loan requests, patron suggestions and book reviews, and, in academic libraries, reviewing the curriculum each year.

Endnotes

1. **Hassler, H. C.** (2005). *What Are Our Customers Reading? An Analysis of the Most Frequently Used Subjects of the NIST Research Library Book Collection Based on Circulation (*NISTIR 7205). Gaithersburg, Md.: National Institute of Standards and Technology. Retrieved from: *http://www.nist.gov/nvl/upload/TopCallNumbers.pdf*.

2. **Levene, M.** (2010). *Introduction to Search Engines and Web Navigation* (2nd edition). Hoboken, N.J.: Wiley. p. 386.