Effectively Visualizing Library Data

*Correspondence concerning this column should be addressed to **Eric Phetteplace**, Emerging Technologies Librarian, Chesapeake College, 1000 College Circle, Wye Mills, MD 26179; e-mail: ephetteplace@chesapeake.edu.*

Data visualization has become a hot topic over the last few years. This popularity can be seen everywhere. The *New York Times* revolutionized itself by creating gorgeous, interactive visualizations of everything from political campaigns to nutrition.[1] There is a suddenly vibrant data visualization blog scene, lead by stellar resources FlowingData and Information is Beautiful.[2] More and more software programs have appeared that make creating compelling visualizations easier and easier, but also more and more data are collected and analyzed digitally. Our new means of collecting data necessitates new means of representation to communicate the message of massive and messy sets of data. The immense success of computing in the second half of the twentieth century has brought about a correspondingly immense challenge: how do we deal with all of this data? Design thinking stepped up to the challenge by allying with statistics to create the half-art, half-science of data visualization.

Libraries have not been left behind; everywhere the rise of new forms of data presentation is visible. The Seattle Public Library features a live-updating dashboard showing recent circulations and intricate networks of keywords.[3] The Harvard Library Lab is working on a powerful tool to view collection size and circulation by subject heading.[4] The Indianapolis Museum of Art has a web dashboard that boldly presents their changing quantities of artwork, memberships, visitors, and of course Facebook fans.[5] Brown University Libraries are working on a similar idea that presents live data such as checkouts in embeddable "widget" form.[6] The North Carolina State University Libraries have an ambitious data visualization project that aims to visualize the usage of reference services, course tools, computer workstations, and group study rooms.[7] These are all great strides being made by innovative libraries. Hopefully, as more examples appear and these frontrunners release their techniques as open-source code, the barrier to entry will diminish for all libraries. In the years to come, public data dashboards on library websites may be as common as catalog search boxes. It is not the aim of this column, however, to enumerate the most beautiful graphs of library data on the web but to demystify the practice of data visualization so that we can begin to create our own.

THE PURPOSE OF VISUALIZATION

Data visualization has a clear purpose: to aid in our understanding of data. Visualizations help us recognize otherwise obscure trends. Many visualizations offer means of

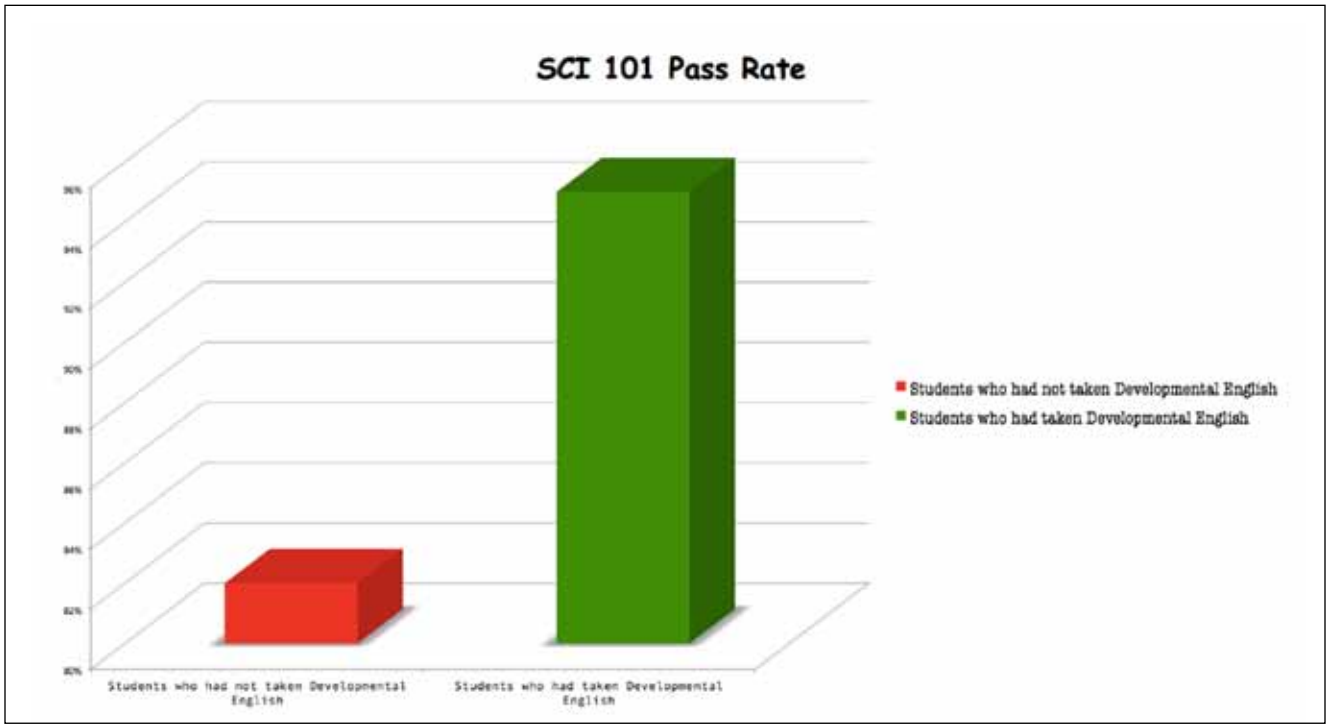

Figure 1. This chart borrows from one presented to faculty at my college. Can you spot its design flaws?

interaction, they do not merely report a single, unassailable truth, but give the end-user an opportunity to explore the data and reach possibly unanticipated conclusions. They help us simplify the interpretation of an intimidating universe of data, intimidating both in terms of the sheer size of datasets, which can consist of millions of data points but also in terms of the *dimensions* of the data. A dataset might have only a dozen distinct points, but each point could in turn have thousands of aspects to it, a veritable world unto itself. Think of library branches: even large public library systems may have a single digit number of branches, but each of those branches has distinct sets of opening hours, geographic coordinates, items in the collection, staff members, and services. Distilling those differences into a meaningful representation is a challenge.

Data visualization approaches the challenges of size and complexity by fusing the art of design with the logic of statistics. Success hinges on both; a brilliant statistician can produce incoherent spreadsheets, while a talented designer can create misleading visuals. Above all, visualization strives to *accurately* represent data; mere artistry descends into "chart junk" and delusion.

Figure 1 is representative of everything data visualization seeks to overcome. It is not simply that the figure is a regular bar chart—bar charts are fine, sometimes they are quite appropriate—but that it is deeply flawed on both aesthetic and logical levels. Aesthetically, the legend on the right contributes no additional information while wasting a significant amount of space. The 3D nature of the graph, far from adding a visual enhancement, makes it difficult to spot exactly where each bar ends: is the former one 81 or 82 percent, and is the latter 94 or 95? There are four different typefaces employed, none to great effect, and the vertical axis' label text is far too small. Finally, the red-green color scheme will not be apparent to people with certain forms of colorblindness.

Far worse are the logical flaws of the chart in figure 1: it uses an artificially high base value of 80 percent to make the two values seem much further apart than they are. While the second bar is actually only 13 percent greater than the first, it appears several times as tall. What's more, so much is left out of the picture: were these two groups of students from the same population? What were their differing demographics? What was the sample size? If only a handful of students were in the second group, then the difference between the two could easily fall within the margin of error. There are probably many more variables available in this data set and perhaps some send the opposite message; what if the average GPA of students in the first group is higher? Overall, the chart is presented not merely in a poor manner, but in a positively deceptive one.

FUNDAMENTALS OF VISUALIZATION

The first step of effectively presenting data has nothing to do with presentation, it is collecting the right data. As Edward Tufte, the so-called godfather of data visualization, says, "If the statistics are boring, you've got the wrong numbers."[8] In

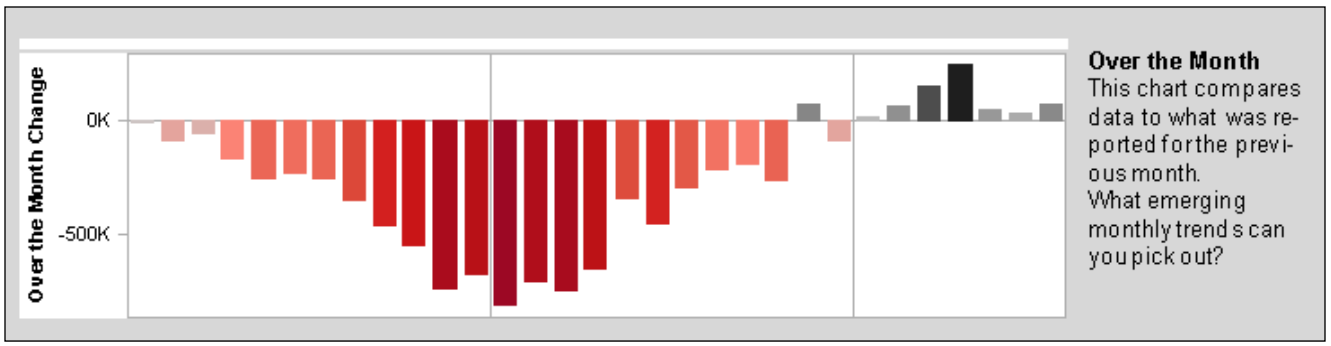

Figure 2. This simple graph from Tableau Public of job growth during recession uses color to accentuate positive and negative values, but note the matching shading as well: it is almost as effective in gray scale. www.tableausoftware.com/public/gallery/job-growth-recession

general, libraries tend to gather the same types of information: how many reference questions are asked, how many books are circulated, how many people enter a building through a particular gateway. These figures are collected because they are easy and obvious, but not necessarily because they are, by themselves, informative. Attempting to create a visualization with inadequate data will often be revelatory in and of itself; if your data has few dimensions, if it can be represented by two columns in a spreadsheet, if it's exceedingly predictable, then you will struggle to find an expressive way of representing it. Indeed, little can be done with reference question totals except to graph their quantity over time, or state results in a plain sentence: "We answered 2,000 questions last year." If that is the extent of the data you have, then that also will be the extent to which it can be visualized. Bounteous design cannot overcome data drought.

We must brainstorm potential collection strategies. Ask questions first and then determine what is needed to adequately answer them. Look at what you already have; can disparate data sets (your collection, your reference interactions, your patrons) be cross-compared in a new and useful way? What's more, expand your definition of what data are: with the powerful tools now available, raw text is more of a data set than ever before. Open-ended survey questions can be investigated in aggregate, rather than through the cherry picking of specific responses. If you cannot find usable data in your current collection strategy, then it is worth reconsidering why you are amassing abstract facts. At the very least, the fine-tuned organization of a library catalog is a wonderful starting place, ripe with potential analyses. Another common but unbelievably rich set of information is website analytics, which are easily collected through free tools such as Google Analytics or Piwik.[9]

Once data has been collected, understanding the basic types of visualizations allows you to select the most appropriate ones. Chances are you already know most of these: line graphs for temporal data, maps for geographic data, bar graphs for simple comparisons. But there are many more options, each suited to particular forms of data. A treemap is a rectangle filled with smaller rectangles that represent the data points; the area of each rectangle can represent a quantitative aspect while coloring in the rectangles can show categories or another quantitative aspect. The rectangles also can be grouped into two or more sets to show a further layer of categorization. See Information is Beautiful's "Billion Dollar-o-Gram" for a poignant example; this treemap helps contextualize unreal quantities of money while also displaying basic categorization via coloring.[10] A network graph shows a series of nodes linked by lines: nodes that share similar relations appear physically near to one another, but the size of a node (usually represented by a circle) also can display another quantitative aspect. Finally, a tree chart can show hierarchical organizations, often allowing users to explore the hierarchy but drilling down specific paths. Wikipedia's article for "chart" has a good section that displays many different types of charts with example images included.[11]

Sometimes the obvious tool is the right one, but it also can be built on. Can you overlay your map with a second type of visual? Can bars and a line on different scales coexist on a single chart? Experiments with crossbreeding visualization types can yield fruitful results or dangerous gibberish.

Color is a perfect example of a useful but dangerous enhancement. On the one hand, it illuminates a third data dimension without requiring too much cognitive effort on the viewer's part. On the other hand, it has more than its share of caveats: color-blindnesses render certain contrasts moot and viewers can be unintentionally deceived.[12] As Edward Tufte states, "The first principle in bringing color to information: *Above all, do no harm*."[13] One important tip: always try to employ texture contrasts in addition to coloring so that the visualization is just as effective in gray scale. You also can test the efficacy of coloring by removing any legends, then presenting the visualization to someone unfamiliar with the underlying data set. Can they intuit which colors represent higher quantities and which represent lower ones? If not, then perhaps another color scheme is called for.

Finally, many modern visualizations distinguish themselves from the static pie charts of the past with their interactive elements. Data points can be highlighted, muted, sorted, traversed, and viewed from multiple angles. Interactivity is

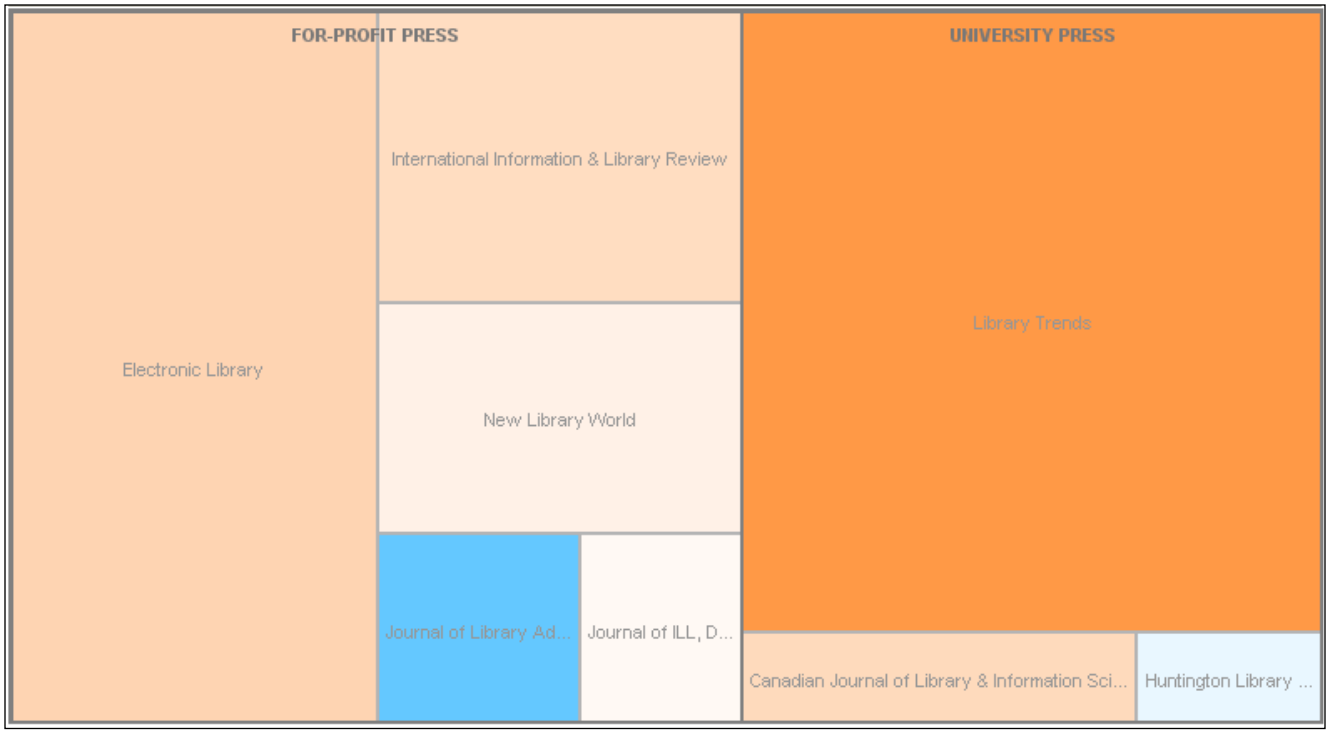

Figure 3. A treemap of a few LIS journals showing their publisher type, number of articles published in 2009, and ratio of citations received to articles published. www-958.ibm.com/software/data/cognos/manyeyes/visualizations/uiuc-lis-journals-sample-treemap

vital because it adds a discovery layer to the representation: suddenly, the creator of a visualization does not have a monopoly over its meaning because the viewer can unveil patterns on their own. One implication of this is that more and more visualizations live on the web, where a diverse range of viewers can access, explore, and remix them. Traditional desktop software produces only static charts that cannot be manipulated, but the tools discussed in the next section can produce visuals that live on the web and take advantage of end user interaction.

TOOLS OF THE TRADE

There are many, many lists of data visualization tools out there and this column will not try to compete with them. If you want a comprehensive and well-curated list of your options, DataVisualization.ch has a "carefully selected list of recommended tools" most of which are free to use and run in the web browser.[14]

It is important to know what you are looking for in a tool. Are you comfortable with code, so a programming language like R or Processing would not overwhelm you?[15] Or do you want a graphical user interface with fewer options but a gentle learning curve? Beyond how a tool is configured and used, it must be able to produce the right type of chart: if you want to make a steam graph or a force-directed layout for a network diagram, then Microsoft Excel is not going to be sufficient. With those questions in mind, you can peruse the options and test out a few. Do not stop until you have found the right one. The tools are rapidly becoming the easiest part of data visualization; it is the preceding sections of this column that will be a struggle. Accumulating meaningful data and imagining a way to visualize it are tasks that only get easier with repeated attempts. New visualization tools, on the other hand, seem to appear every week.

There are a couple standout visualization packages that do not require specialized knowledge. Tableau Public is a free desktop application that covers all the most common visualizations extremely well. The recession job growth chart in figure 2 comes from Tableau Public's Gallery, which gives a good sense of what it is capable of. All of its visualizations offer interactive filters and web exports, a massive advantage over standard spreadsheet software. Unfortunately, Tableau Public only runs on the Windows operating system.

IBM's Many Eyes is a Java-based web application for publicly sharing data and visualizations. There are many data sets already available, making it ideal for practicing and getting a feel for the possibilities of more sophisticated forms such as network diagrams and sentence trees. The text visualizations in particular stand out as potentially enlightening: try throwing survey responses or a favorite text into them to see if you can reveal an interesting motif. Many Eyes' chief disadvantages are that everything posted is public—so one must be careful of exposing potentially sensitive data—and that it is unlikely to improve beyond its current state because

its creators (Fernanda Viégas and Martin Wattenberg) no longer work at IBM.

I created the example treemap in figure 3 using Many Eyes. It shows a number of data aspects in one compact display. Data points are sorted into two categories (For-Profit Press on the left, University Press on the right), the size of the rectangles corresponds to the number of documents published, and the coloring shows ratio of citations received to documents published (blue is low, orange is high). What is more, the live treemap on the website provides users with the ability to rearrange how the data are presented, swapping 2008 figures for 2009, substituting SCImago Journal Rank for either citations received or documents published, or zooming in on a particular publisher type. One could also add other categorizations, such as open versus closed access. However, this image also shows some of the pitfalls of Many Eyes, as one cannot control the font size or color scheme. The only way to have complete control over presentation is by using programming languages and design software.

KEEP IN MIND

Data visualization is meant to illuminate, not obfuscate. Even as you add more layers and reveal additional dimensions, the data's structure and message should become easier to interpret. If a design is only adding more and more complexity then it is going against the very purpose of representation. If anything, the first step in data visualization is not adding extra dimensions with color or providing for interactivity with a set of data filters for viewers to choose from. Instead, it is avoiding all the mistakes that make for misleading infographics. I highly recommend Darrell Huff's classic *How to Lie with Statistics* for a thorough but accessible overview of common mistakes in data presentation.[16] For a piece more specific to libraries, try Ray Lyons' superb blog post "Beauty Is As Beauty Does," which debunks much of the glamor of 3D effects and flashy infographics in favor of accurate representation of data.[17]

It bears repeating: the first step to good data visualization is good data. Most of the thought and effort should go into to collecting and analyzing data; playing with visuals until you find a compelling option is the reward for your due diligence.

References

1. Small Labs, "New York Times Infographics," accessed July 22, 2012, www.smallmeans.com/new-york-times-infographics.
2. Nathan Yau, "FlowingData," accessed July 22, 2012, http://flowingdata.com; David McCandless, "Information is Beautiful: Ideas, Issues, Knowledge, Data—Visualized!" accessed July 22, 2012, www.informationisbeautiful.net.
3. "Making Visible the Invisible, 2005–2014," accessed July 22, 2012, www.mat.ucsb.edu/~g.legrady/glWeb/Projects/spl/spl.html.
4. "Harvard Library Explorer," accessed July 22, 2012, http://librarylab.law.harvard.edu/toolkit.
5. IMA Indianapolis Museum of Art, "Dashboard | Indianapolis Museum of Art," accessed July 22, 2012, http://dashboard.imamuseum.org.
6. Brown University Library, "dashboard_beta :: dashboard information," accessed July 22, 2012, http://library.brown.edu/dashboard/info.
7. NCSU Libraries, "Visualizing Library Data," accessed July 22, 2012, www.lib.ncsu.edu/dli/projects/dataviz.
8. Edward Tufte, *Envisioning Information* (Cheshire, CT: Graphics Press, 2006), 34.
9. Piwik # Open Source Web Analytics, "Piwik—Web Analytics—Open Source," accessed July 22, 2012, http://piwik.org.
10. David McCandless, "The Billion Dollar-o-Gram: 2009 Figures," accessed July 22, 2012, www.informationisbeautiful.net/visualizations/the-billion-dollar-o-gram-2009.
11. Wikipedia: The Free Encyclopedia, "Chart," accessed July 22, 2012, http://en.wikipedia.org/wiki/Chart#Types_of_charts.
12. Drew Skau, "Dear NASA, No More Rainbow Color Scales, Please," http://blog.visual.ly/rainbow-color-scales (accessed July 23, 2012).
13. Tufte, *Envisioning Information*, 92–93. Italics in original.
14. DataVisualization.ch, "DataVisualization.ch Selected Tools," accessed July 22, 2012, http://selection.datavisualization.ch.
15. The R Project for Statistical Computing, "What is R?," accessed July 22, 2012, www.r-project.org; Processing, "Overview: A Short Introduction to the Processing Software and Projects from the Community," accessed July 22, 2012, http://processing.org/about.
16. Darrell Huff, *How to Lie with Statistics* (New York: Norton, 1993).
17. Ray Lyons, "Beauty is as Beauty Does," accessed July 22, 2012, https://libperform.wordpress.com/2011/10/28/beauty-is-as-beauty-does.