Visualizing Scholarly Content

Martin Fenner

August 9, 2014

One topic I will cover this Sunday in a presentation on Open Scholarship Tools at *Wikimania 2014* together with Ian Mulvany is visualization.

Data visualization is all about *telling stories with data*, something that is of course not only important for scholarly content, but for example increasingly common in journalism. This is a big and complex topic, but I hope the following will get you started.

Learn the Basics

Work on visualization of scientific data should start with a good understanding of the best practices and pitfalls of data visualization in general, as well as the specific aspects of visualizing scientific data. The following resources have helped me get started - please suggest more in the comments:

- Visualize this. A book from Nathan Yau published in 2011. Very helpful in understanding the different ways data can be visualized (e.g. when to use a treemap or what is a chloropleth map), and an introduction to some tools using practical examples. Nathan's FlowingData blog is also a great resource.
- D3 Gallery. Lots of examples generated using Mike Bostock's d3.js visualization library. A great inspiration for data visualization on the web, even if you use a different visualization tool.
- ggplot2. Not only a very popular visualization library for the R language by Hadley Wickham, but also an implementation of Leland Wilkison's Grammar of Graphics. The ggplot2 book describes this powerful concept (p. 14):

  > In brief, the grammar tells us that a statistical graphic is a mapping from data to aesthetic attributes (colour, shape, size) of geometric objects (points, lines, bars). The plot may also contain statistical transformations of the data and is drawn on a specific coordinate system. Faceting can be used to generate the same plot for different subsets of the dataset. It is the combination of these independent components that make up a graphic.

Learn to use at least one visualization tool

There are many great tools available, pick one and learn it well. Some options include:

- **Excel**. Probably the most popular tool for data visualization. Commercial, with open source alternatives such as Libre Office.
- **R**. Software for statistical computing and analysis. Open source. RStudio is a powerful user interface for R and a good way to get started.
- **d3.js**. A visualization library for Javascript. Open source.
- **Prism**. A popular visualization tool among scientists. Commercial.
- **Datawrapper**. An open source tool and hosted service for data visualization.

I do most visualizations in either R or d3.js. Both are open source tools with a large community and a rich set of libraries, examples and documentation, and both take a systematic approach to data visualization (see grammar of graphics above).

Learn data analysis

Unless your interest is more in information design - see Information is beautiful for some great examples - data visualization is tightly coupled with data analysis. You need to know at least the basics of data analysis to do proper data visualizations, e.g. how to handle wrongly formatted data (e.g. text in a number column), missing values and outliers. The most time-consuming step in my experience is data transformation, i.e. bringing data into the format that you want for the analysis and visualization.

R, Python and the relatively new Julia are popular languages for data analysis available as open source. There are many packages for these languages that help with common data analysis problems. One additional advantage of using a proper language over a set of tools cobbled together is that it is easy to automatically recreate a visualization with a new set of data - convenient when you need to analyze and visualize an ongoing experiment that repeatedly produces new data.

Use a vector file format

Too many scientific data are still visualized using bitmap graphic formats such as `tiff`, `jpg` and `png`. These formats are not appropriate for charts and only make sense for images. They don't scale to the screen resolution, and it is very hard to impossible to reuse or even modify them. Use vector graphic formats such as `svg` or `pdf` instead. `svg` is my preferred format because in contrast to `pdf` it can be embedded into a larger HTML document, and R and d3.js (my preferred visualization tools) can generate this format. Inkscape is an open source SVG editor, and the commercial **Adobe Illustrator** can be used to manually polish graphics in `svg` or `pdf` format, e.g. for journal publication.

Get inspired by great visualizations

At the end of the day data visualization is all about telling a story with data. Unfortunately the current state of affairs for scientific visualizations is very different. In my opinion most graphs and figures used in publications don't provide the data underlying the visualization (Datawrapper is a great example how this can be done), focus too much on detail rather than the overall message, don't take advantage of the different chart types available, and are sometimes even misleading. And I'm not even talking about the fact that figures in scholarly papers are almost never interactive. It rarely happens that I read a paper and get excited by looking at a figure - if I do it is usually because the underlying data are so compelling that even the simplest visualization will convey the right message.

We should become more creative with visualizing data in scholalry documents, and one important step towards that goal is publishers accepting more reasonable file formats in manuscript submissions - instead of just `tiff` and `eps` (PLOS), or `tiff`, `eps` and `pdf` (Science), and often with a 10 MB file site limit.