Seeing a Path to Understanding: Visualizing SIS Data

1/29/2015
Ariel Deardorff, NLM Associate Fellow
National Library of Medicine

Project Sponsors:
Colette Hochstein, Biomedical Information Services Branch, SIS
Florence Chang, Chief, Biomedical Files Implementation Branch, SIS
Jeanne Goshorn, Chief, Biomedical Information Services Branch, SIS

Table of Contents

Abstract

Objective

The goal of this project was to explore data visualization technology in relation to datasets housed at the Specialized Information Services (SIS) division of the National Library of Medicine. The specific objective was to create one or more pilot SIS data visualization products.

Methods

An informal needs analysis was conducted to learn what the organization wanted to visualize and to inform selection of a visualization tool. Data were then extracted, processed and cleaned to be ready for use in the visualization tool. Two data dashboards were designed which then underwent user feedback and significant redesign.

Results

The visualization tool Tableau was used to create two data dashboards illustrating chemical records from the Hazardous Substance Databank (HSDB). The visualizations were designed to answer questions about records currency and importance based on the dates that they were last updated and reviewed, and the number of times the chemical had been cited in PubMed.

Conclusion

The visualizations allowed SIS to easily see which HSDB records haven't been updated in some time and helped them decide which records to choose for review by a scientific panel, a process that had previously been more time consuming and labor intensive. It is recommended that the data dashboards be maintained and updated so that they continue to be useful tools for decision making.

Introduction

Specialized Information Services

Specialized Information Services (SIS) is a division of the National Library of Medicine that focuses on a variety of information areas including environmental health and toxicology, chemical and drug information, HIV/AIDS, health outreach, and disaster information management. SIS maintains databases such as ToxNet (a database of hazardous chemicals, toxic releases, and environmental health) targeted population health sites like Arctic Health and Women's Health, disaster and emergency response tools like the Chemical Hazards Emergency Medical Management (CHEMM), drug and dietary tools like the Drug Information Portal, and HIV/AIDs resources like AIDSinfo (a site for federally approved treatment guidelines, clinical trials, drug and vaccine information).

Project Goals and Objectives

The goal of this project was to experiment with data visualization as a way of presenting and learning from data. More specifically, SIS was interested in methods of visualizing big data with the idea that "innovative visualization of this data will help SIS and its users to view, understand, and discover otherwise difficult to recognize relationships, patterns, and trends."

The specific objective of this project was to create a pilot data visualization tool using SIS data. The data in question consisted of chemical records from the Hazardous Substance Databank (HSDB) and the goal of the visualization was to help members of the SIS team understand what records needed to be updated based on a variety of factors including the date they were last updated, the date they were last reviewed by a scientific panel, and the number of citations the chemical had in PubMed. If the pilot proved successful SIS would investigate ways to maintain the visualization tool and explore further means of incorporating visualization technology into their workflow.

Methodology

Needs Analysis

The project began with an informal needs analysis of SIS team members to learn who would be using the visualization and what kinds of information they wanted to visualize. This initial meeting as well as subsequent discussions provided the basic guiding design principle for the visualization.

Choosing a Tool

While the SIS team was leaning towards using the visualization tool Tableau, they first wanted to ensure that there weren't other visualization tools more suited to their needs. To check for competing tools the Associate performed a quick environmental scan of popular visualization products and software (see Appendix A for a complete list). While several visualization tools were uncovered many of them were designed for geographic data, which was not the data type in question and would therefore not be useful for this pilot project. In addition, many of the tools required advanced JavaScript or other programming skills which didn't seem very user friendly and would make it much harder for SIS team members to adopt the tool. Because of these factors it was decided that Tableau was the best tool for this pilot project.

Tableau visualization software is designed to work with a variety of data formats and visualization styles including choropleth (density maps), bar graphs, bubble graphs, pie charts, and more. In addition, Tableau has a user-friendly drag-and-drop interface which makes it easy to create visualizations. Tableau also has a very strong user community and user forums, which makes it easy to ask questions and find tips. Finally, while an enterprise version of Tableau is available for purchase, there is also a free version (Tableau Public) that seemed well-suited to a pilot project and was therefore chosen to be the visualization tool.

Data Cleanup

Every visualization project requires a certain amount of data cleaning and normalization. Because this project used the free version of Tableau the data needed to be uploaded in a CSV or Excel file (the fee-based version allows users to connect directly to a variety of databases). This meant that data needed to be downloaded from various databases (HSDB, PubMed), combined and normalized in an Excel spreadsheet, and uploaded into Tableau. This phase of the project required considerable collaboration and discussion with the SIS data team in order to pull all the data and get it into the proper format.

Creating the Visualization

Once the data were in the appropriate format it was time to design the visualization. To begin with, several ideas were sketched out on paper to process how the various tools (color, size, and format) could be used to show different data points and aspects of the data. Once the rough sketches were completed it was possible to start building the visualizations in Tableau. This required creating several different charts and graphs, deciding which filters and tools to apply, and then combining everything into user-friendly dashboards.

User Feedback and Redesign

To ensure that the visualization dashboards were useful and usable, the designs were sent to the SIS team to gather feedback. Comments and input from the team helped to reformat the design of the visualization to one that was more focused on answering specific questions and prompting particular analyses.

Results

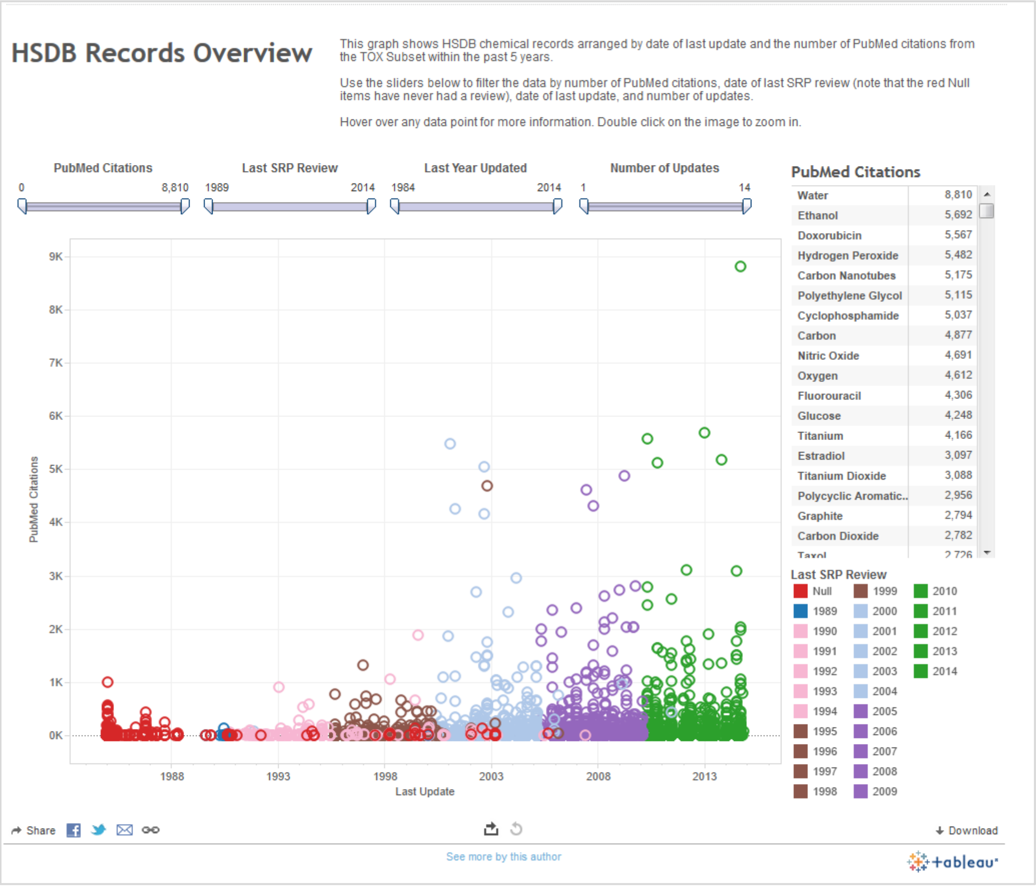
This project resulted in two data dashboards showing two ways of visualizing chemical records from the HSDB. The dashboards were designed to answer the following questions:

- Which chemicals have never been reviewed by a Scientific Review Panel (SRP)?
- Which chemicals have many PubMed citations but haven't been updated recently?
- Which chemicals have many PubMed citations but haven't been reviewed by an SRP recently?
- Which chemicals haven't been updated or reviewed recently?

Dashboard 1: Overview Approach

The first dashboard "HSDB Records Overview" (Figure 1) attempts to combine all of the different measures in one visualization. In this version, each circle represents one chemical record. The chemicals are arranged on the graph based on the number of their PubMed citations and the date they were last updated. The colors represent the years when the records were last reviewed by an SRP (the legend is found on the right). The filters at the top allow users to change the view to limit to records with a certain number of PubMed citations (if, for example, you only wanted to see the highly cited chemicals), as well as chemicals reviewed or updated before or after a certain date. The filters update both the scatter plot and the chart to the right, which simply lists all the chemicals in order of their PubMed citations and serves as another way to quickly see the data.

The strength of this dashboard is that it allows users to quickly see all the different factors at once. The weakness is that it can be a little confusing to compare everything in one chart and it can be more difficult to confidently apply filters.



HSDB Records Overview

This graph shows HSDB chemical records arranged by date of last update and the number of PubMed citations from the TOX Subset within the past 5 years.

Use the sliders below to filter the data by number of PubMed citations, date of last SRP review (note that the red Null items have never had a review), date of last update, and number of updates.

Hover over any data point for more information. Double click on the image to zoom in.

PubMed Citations

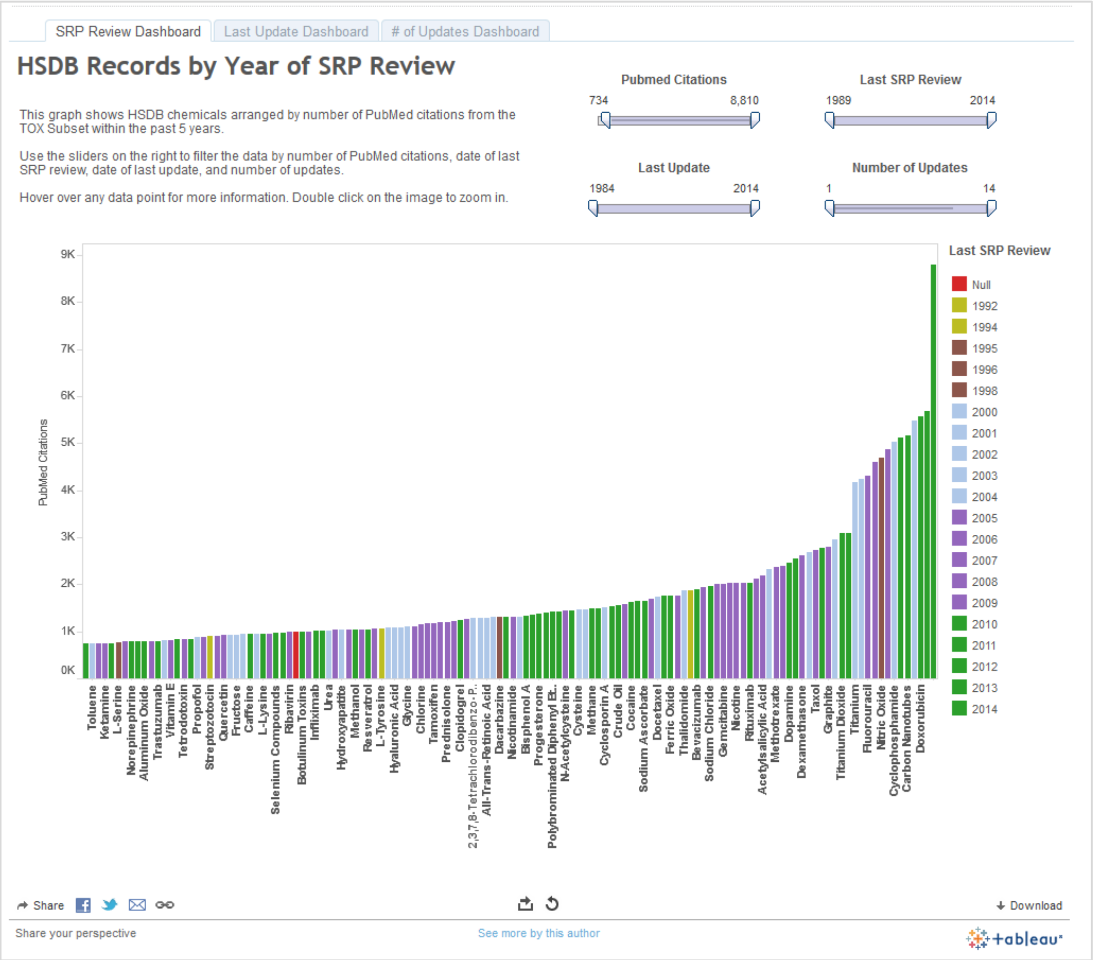
Water	8,810
Ethanol	5,692
Doxorubicin	5,567
Hydrogen Peroxide	5,482
Carbon Nanotubes	5,175
Polyethylene Glycol	5,115
Cyclophosphamide	5,037
Carbon	4,877
Nitric Oxide	4,691
Oxygen	4,612
Fluorouracil	4,306
Glucose	4,248
Titanium	4,166
Estradiol	3,097
Titanium Dioxide	3,088
Polycyclic Aromatic..	2,956
Graphite	2,794
Carbon Dioxide	2,782
Taxol	2,726

Figure 1: Overview Approach

Dashboard 2: Tabular Approach

The second dashboard "HSDB Records by Year of SRP Review" (Figure 2), was composed of three tabs, each of which took a slightly different view of the records. Each tab in the second dashboard shows every chemical record as a bar with the length of the bar representing the number of PubMed citations. The color of the bar represents different factors depending on the particular tab. In the first tab (shown below) the color represents the year the chemical had its last SRP review; in the second tab (not shown), the view is exactly the same except the color represents the year the chemical was last updated; and in the third tab (also not shown), the color represents the number of updates. The four filters work exactly the same on all the tabs and allow users to filter down the results to see records from a certain date, or with a certain number of citations.

This tabular view was designed to enable a more in-depth analysis of a particular aspect of the records such as date of last SRP review. The strength of this dashboard is that it makes it easier to quickly get an overview of one attribute of the data and to identify the extremes (the oldest, most popular, etc.). The weakness is that it is no longer possible to compare multiple factors such as number of PubMed citations, date of last update, and date of last review.



HSDB Records by Year of SRP Review

This graph shows HSDB chemicals arranged by number of PubMed citations from the TOX Subset within the past 5 years.

Use the sliders on the right to filter the data by number of PubMed citations, date of last SRP review, date of last update, and number of updates.

Hover over any data point for more information. Double click on the image to zoom in.

Pubmed Citations: 734 – 8,810

Last SRP Review: 1989 – 2014

Last Update: 1984 – 2014

Number of Updates: 1 – 14

Figure 2: Tabular Approach

Impact on SIS

The two Tableau data dashboards were well received by the SIS team. The fact that the data could now be examined in a visual manner made it much easier to spot patterns as well as errors in the data. The dashboards also allowed the SIS team to easily see which chemical records were in need of updating and reviewing. Whereas SIS team members had previously needed to download information from multiple sources and compare them manually – a labor and time intensive process - they could now use the dashboards to easily compare multiple factors at the click of a mouse. Data visualization, therefore, made it simpler and quicker to make informed decisions.

Discussion

New Experiences

One of the benefits of working on this project was the chance to work with a wide variety of people including programmers at SIS. Because this project involved drawing data from two different NLM databases, programmers were absolutely essential. The visualizations would never have been created without the data extraction scripts and skills of Florence Chang and her team of programmers at SIS.

Another interesting aspect of this project was that it allowed the Associate time to really delve into Tableau and take on the role of designer. While the Associate had been involved in several Tableau-related projects before, she had not had the chance to design for such a targeted audience so it was rewarding to create dashboards that were designed for a particular audience to meet a particular use case. Furthermore it was beneficial to have the time to explore more advanced features of Tableau and create more robust and useful visualization tools.

Challenges

While this project was very successful there were some challenges. Initially it was somewhat difficult to download Tableau Public as it is a cloud-based tool. While cloud-based technology is very prevalent in the consumer world (think Google docs, Dropbox, etc.) it has been embraced more slowly by government agencies like Health and Human Services. It was therefore necessary to present a business case for the tool to convince the NLM computer team to allow it to be used. Eventually permission was granted to use Tableau Public as part of a pilot project. While everything worked out in the end this setback was a good reminder to check one's assumptions when creating a project plan. In planning future projects it is worth ensuring that all the tools, products, and resources necessary for success can actually be procured.

Recommendations

Summary of Recommendations

As the data dashboards were well received by SIS the Associate recommended that they be maintained and updated so that they can continue to be useful in decision making. Moreover, given that Tableau was thought to be a useful tool it is recommended that SIS attempt to purchase a more robust version or continue with the current free version. Finally, as information visualization proved to be a useful means of analyzing and presenting information, the Associate recommended that SIS continue to invest in visualization technology and training and seek out other users at NLM to form discussion groups.

For a detailed list of recommendations please see Appendix B.

Acknowledgments

This project would not have been possible without the help of project sponsors Colette Hochstein and Florence Chang, project consultant Jeanne Goshorn, and SIS team members Jennifer Dong and Ziying Sherwin. Special thanks also to Kathel Dunn, Wanda Whitney, and Maureen Madden for their feedback and support.

Appendix A: List of Visualization Tools Analyzed

Name	Type	URL
General		
R	Bar, scatter, area,	http://www.r-project.org/
Tableau	Choropleth, bar graph, bubble, pie chart, keyword table	http://www.tableausoftware.com/public/community
Excel	Donut chart, bar graph, box and whisker, pie chart, line graph, area, scatter, surface, bubble, radar	
Plotly	Line and scatter plots, bar charts, box plots, bubble, contour, area, choropleth, histogram, polar charts, time series, surface,	https://plot.ly/python/
Bokeh	Line and scatter plots, bar charts, box plots, bubble, contour, area, choropleth, histogram, polar charts, time series, surface,	http://bokeh.pydata.org/
Sci2	Network, choropleth, bar graph, circular network	https://sci2.cns.iu.edu/user/index.php
Qlik	Choropleth, bar, map, box plot, pie chart, area chart	http://www.qlik.com/
GIS/Mapping		
GeoDA	GIS/Mapping	http://geodacenter.asu.edu/
GeoCommons	GIS/Mapping	http://geocommons.com/
GRASS GIS	GIS/mapping	http://grass.osgeo.org/
Mango Map	GIS/mapping	https://mangomap.com/
ArcGIS Desktop	GIS/Mapping	https://www.arcgis.com/features/
Mapbox	GIS/mapping	https://www.mapbox.com/
Network		
Gephi	Network	http://gephi.github.io/
Cytoscape	Network	http://www.cytoscape.org/
IN-SPIRE	Terrain and network	http://in-spire.pnnl.gov/
Word Cloud		
Wordle	Word cloud	http://www.wordle.net/
Tagxedo	Word Cloud, fancy	http://www.tagxedo.com/

Appendix B: Detailed Recommendations

1. Maintain and Update the HSDB Dashboards
As the HSDB data dashboards proved to be useful in decision making they should be maintained and updated. Because assembling the data requires a certain amount of processing and cleaning, updating should probably be done on a quarterly basis.

Maintaining the dashboards will entail:

- Keeping the dashboards stored in an account in Tableau Public
- Sharing the URL links with anyone who wants to access the dashboards

Updating the dashboards will require:

- Updating the underlying data spreadsheets with new citation counts, update dates, review dates, and number of updates
- Updating the link between the data and the spreadsheets in Tableau Public
- Saving any changes to the dashboards back to the cloud

2. Improve the HSDB Dashboards
Given more time or expertise there are certain changes that would improve the data dashboards including:

- The ability to easily switch between PubMed citations from the last 5 years and citations from the last 10 years
- A way to list “Null” in the legend as “Never” to reduce potential confusion

3. Investigate further uses for Tableau at SIS
It is recommended that SIS explore further uses for Tableau and other visualization tools. This could involve similar dashboards for other databases such as ChemID or the Drug Information Portal. In addition, SIS might wish to explore visualization as a tool for the presentation of information in reports or publicity. Targets for this other form of visualization might include information about outreach efforts or grant recipients.

4. Seek out other Tableau users
One of the benefits of Tableau is its strong user support groups and it is therefore recommended that SIS seek out support for training and visualization ideas. It is very likely that there is already a user group at NIH and as more areas of the library investigate Tableau it might be worth starting a Tableau user group at NLM. A strong user community at NLM will help make information visualization an essential part of communication and data-driven decision making.