Seeing the Story: Designing accessible data visualization tools tailored for natural history collections.

Jessica Mailhot*

CU Boulder

ABSTRACT

Natural history collections (NHCs) are dynamic and active agents of discovery, serving as record-keepers of nature by collecting and preserving irreplaceable biological specimens. Researchers from a wide range of fields utilize these biological resources to address topics of global importance. NHC databases are crucial for specimen management, preservation, and accessibility to researchers. As these datasets continue to grow in depth and breadth, so too has the need for innovative and effective tools to analyze, explore and communicate them. Data visualization has rich potential for addressing NHCs' specific objectives, and while there are a suite of different projects experimenting with this application, there are still many barriers preventing data visualization from being widely accessible to any NHC. This project aims to address these barriers by weaving together training in data visualization and NHC management to design a widely-applicable dashboard template in Tableau Public, so that any NHC can download the template and connect it to their own dataset with the help of a simple tutorial.

Keywords: Museum, natural history collection, dashboard, user-centered design.

Index Terms: Human-centered computing ~ Scientific visualization, Human-centered computing ~ Visualization toolkits

1 INTRODUCTION

Natural history collections (NHCs) play an ever-evolving and irreplaceable role in enabling research of global importance. They contain millions of biological specimens, each with a unique set of associated data that act as a snapshot of the natural world at a specific time and place. These66 attributes include collection date, location, species identification, measurements, parasites, contaminants, overall health, and many more. Specimens are carefully persevered and cared for in perpetuity so that they will exist for generations to come. This makes NHCs unlike any other scientific resource, enabling research questions that span whole continents and centuries.

Most specimens were collected long before modern technologies were thought possible; DNA sequencing, isotope analysis, and CT scanning are just a few of the new ways that old specimens are gaining new relevance in modern day. As anthropogenic climate change continues to impact the biosphere, NHCs provide one of the only means to establish a baseline to document change over time. Applications grow in number and relevance with each passing year.

As NHCs continue to become more important and active into the future, it is imperative to find the most effective ways to ensure their prudent and supported management as well as their ability to be used by researchers and educators. Over the past few decades, NHCs have been committed to digitizing their specimen data into complex databases, a monumental task since historical records are voluminous and hand written. NHCs use their datasets in several ways: 1) to inform collection management decisions, 2) to allow researches around the globe to browse for and access specimens and their data, and 3) to quantify their impact and justification for the resources they require. As these datasets continue to grow in depth and breadth, so too has the need for innovative and effective tools to analyze, explore and communicate them.

Data visualization has rich potential for addressing NHCs' specific objectives and challenges, as it has already proven in other fields that do similar data-driven decision-making with complex datasets for many audiences. These tools, especially dashboards, have already begun proving their potential in NHCs for a variety of datasets and audiences. However, this frontier is still relatively new and there are a number of barriers preventing all NHCs and their staff from being able to utilize data visualization, including time and resources, IT experience, profession-specific training materials, and consistency between existing examples to follow. Not only does more work need to be done to perfect the collective portfolio of NHC data visualizations, but there is a significant lack of addressing accessibility, with most existing dashboards being built in isolation using computer code by larger institutions. Data visualization has proven to be powerful and easy to use in many other fields, while it has remained a new frontier in NHCs.

This project aims to directly address some of these barriers by designing a specimen data dashboard which can be easily used as a template for someone else to then download and populate with their own specimen dataset. This is just one

* jema9330@colorado.edu

component of a larger graduate project, which will entail designing a suite of dashboards for a variety of collections management data types and objectives. The products of this larger project will all be freely available to NHC staff on a website that also includes tutorials, a discussion board, a contact portal, and additional resources. This hub of custom-made tools and materials will make data visualization more accessible to all NHCs, big or small.

2 Related Work

Over the past few decades, data visualization has gradually found its way into museums around the world, as a tool for understanding and communicating museum data [1][9][16]. This pioneering movement has produced tools to facilitate acquiring funding, connecting with researchers, expanding outreach, amending policy, and contributing to annual reports [14][16]. Museums are a fitting arena full of potential for data visualization because of the complexity of the datasets, the variety of possible audiences and users, the growing demand for metrics and data-driven decision-making, and the ongoing braiding of digitized datasets in large aggregator platforms. Experimental projects in visualizing museum data include examining bias in a collection's history [8][9][15], illustrating specimen scope and status [3][16], identifying a collection's unique specialties [8], mapping storage areas [12], and applications in exhibit spaces [5][11][13]. Each project like these is a contributing step into this frontier which can inspire and inform the next.

Many museums and related organizations have been specifically experimenting and implementing dashboards, which are data visualization tools that are designed to illustrate a dataset from multiple angles simultaneously using multiple interactive visualization types all arranged together in one display. One example is designed by the Distributed System of Scientific Collections (DiSSCo) shown here in Figure 1, which illustrates relationships between collection size, collection type, and country. Dashboards provide at-a-glance monitoring and sophisticated visual querying [6]. They visualize a dataset in a number of interactive ways that in concert provide a more comprehensive view of its nature than a single static image [2]. NHC dashboards have been designed for objectives such as tracking the progress of digitization [2], monitoring internal operations such as staffing and budgets [22][32], sharing information between departments [4], and discovering key characteristics of collections [2]. They are being included in museums' annual reports and even made accessible to the public to proactively practice transparency about how well the museum is succeeding at following its mission [10][11].

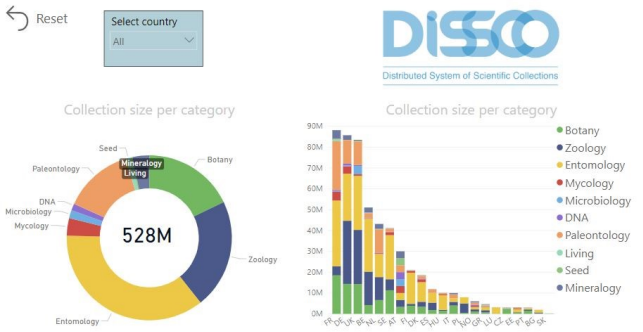

Figure 1: This DiSSCo NHC dashboard illustrates specimens by collection category, country and count using an interactive donut chart and bar chart. It also has a dropdown filter for country.

3 Project Description

3.1 Overview and Data Source

The final dashboard was constructed in Tableau Public because of its approachable user interface, abundance of existing resources, and that it is free to download and use. All of these qualities help address the accessibility barriers mentioned in the Introduction.

Because the dashboard is meant to act as a template, a placeholder dataset was used to populate it for guidance during the design process and to allow interested NHC staff to view and test the dashboard before deciding to download their own copy for their own data. The placeholder dataset chosen is a real-world NHC specimen dataset from the CU Boulder Museum of Natural History Herbarium "Specimen Database of Colorado Vascular Plants", which is publicly available for download via the Global Biodiversity Information Facility (GBIF) [7]. This type of dataset is ideal because the majority of NHCs with digital datasets publish them on GBIF, which uses a standard format and vocabulary. Having the dashboard designed to communicate with this standard dataset format will allow other collections already in GBIF to have a compatible dataset and not need to do extensive modifications to have it work properly.

The dataset was downloaded directly from GBIF on 06 December 2019. Each row is a specimen, with columns for each of the data fields associated with the specimens, such as eventDate, year, countryCode, and scientificName. All empty columns were deleted to reduce file size and clutter. This reduced dataset contains the most common data fields that would be minimally represented in most specimen datasets, which the dashboard was then designed to utilize. The goal is to make the minimum requirements of a connected dataset to be as universal and attainable as possible.

3.2 Dashboard Design

The overall objective of this dashboard is to illustrate a collection's specimens from multiple angles and allow for open-ended exploration as well as answering specific questions. The dashboard is made up of several components: Timeline, Treemap, Map, Legend, and Filters. (Browser version of the dashboard: https://public.tableau.com/profile/jessica.mailhot#!/vizhome/FinalProject_15756046674660/Dashboard1?publish=yes)

The Timeline illustrates the number of specimens collected across time and subcollection. The taxonomic category of class was chosen as the subcollection delimiter because it arranges the data into a manageable number of categories; using a taxonomic level above or below class would create too many or too few categories. Class is a categorical attribute that is encoded here by both color and orientation. Partial opacity, bright hues, black outlines, and balanced size ranges were chosen to counteract the density of the data along the timeline, allowing the user to more easily distinguish colors and individual circles. Time is an ordinal attribute that is also encoded as orientation along the x axis. Specimen count is an ordinal attribute that is encoded as circle size, with one circle for each year and class.

The Treemap illustrates the taxonomic diversity and distribution of the specimens. Here the same color scheme is used for classes. Specimen count is again encoded as size. Taxonomy is a hierarchical categorical attribute encoded as orientation, with umbrella-ed groups arranged together.

The Map illustrates the spatial distribution of the specimens based on from where they were collected. This particular collection is only from within Colorado, with the vast majority including county. Specimen count is encoded as a color ramp in the map, based on a green hue not used in the class pallet and semantically symbolic of vegetation.

There are filters available on the bottom right of the dashboard which effect all three of the visualizations. The filters for year allow the user to manipulate a slider-controlled range of dates or a picklist of specific years. The user can also filter by class from a multiple value picklist. An intuitive "Rest Filters" button reverts the entire dashboard of any filters enacted.

Besides the designated filters, the user can also manipulate the displays themselves. Hovering over datapoints triggers a tooltip box with additional details in all visualizations. Hovering also highlights elements in the Timeline and Treemap, with the Timeline also highlighting all circles for that year across all of the classes to allow for quick cross-category comparisons. Clicking on any element in any visualization then filters all visualizations to that included selection of data. This can be done multiple times to narrow down even further to specific criteria of interest. The "Reset Filters" button undoes these filtering actions as well.

4 Discussion

This project had two parallel goals: to make a dashboard that is easy to duplicate and that is easy to use. The duplicability is achieved by using Tableau as a platform, which is free to use, supported by a vast host of resources, and allows for people to directly download this dashboard to use as a template. It was designed to communicate with a standard dataset format which is used by the majority of NHC datasets. A tutorial will accompany this dashboard to walk the user though how to use Tableau, download and open this dashboard, ensure their dataset is compatible, connect it to the dashboard, and make customizations. This process allows for someone to visualize their specimen dataset and benefit from the power and applicability of data visualization technologies and techniques without the need to fully design and construct a custom-built dashboard from scratch.

The design decisions were also made to support an intuitive and impactful user experience. The complexity of the specimen dataset was distilled to the main data attributes of consequence: time, space, and taxonomy. By having those as the scaffold, simple interactions allow the user to intuitively filter down the entire dataset to specific specimen criteria of interest while understanding how they fit into context of the entire collection. This dashboard can hopefully be used for a variety of audiences and for a variety of applications, including but not limited to: communicating the collection's strengths to administrators, allowing a researcher to explore the resources available in a collection that are relevant to their investigation, facilitate decisions made by the collection managers and curators regarding its gaps and strengths, being embedded into a museum's public-facing website to spark awareness and curiosity for a museum's research collections, etc.

This dashboard is meant to be a work in progress, growing more effective with use and feedback from NHC staff. There are many potential issues that may surface as this dashboard is applied to collections of various sizes and types of specimens. There are also many issues to anticipate while preparing and connecting to the new datasets that need to be explicitly covered in the tutorial, as it is very likely that the users will not have any familiarity with Tableau beforehand. There is still so much to be explored with the applications of data visualization in NHCs and museums in general, and it is this project's goal to add momentum by bringing more people into the conversation.

References

[1] Block, F., Horn, M. S., Phillips, B. C., Diamond, J., Evans, E. M., & Shen, C. (2012). The DeepTree exhibit: Visualizing the tree of life to facilitate informal learning. IEEE Transactions on

Visualization and Computer Graphics, 18(12), 2789-2798. doi:10.1109/TVCG.2012.272

[2] Casino, A., Raes, N., Addink, W., & Woodburn, M. (2019). Collections digitization and assessment dashboard, a tool for supporting informed decisions. Biodiversity Information Science and Standards, 3 doi:10.3897/biss.3.37505

[3] Chang, Michelle T. (2003). Collection understanding. Master's thesis, Texas A&M University. Texas A&M University. Available electronically from http://hdl.handle.net/1969.1/69.

[4] Chapman, J., & Yakel, E. (2012). Data-driven management and interoperable metrics for special collections and archives user services. RBM: A Journal of Rare Books, Manuscripts, and Cultural Heritage, 13(2), 129-151. doi:10.5860/rbm.13.2.379

[5] Deal, L. (2015;2014;). Visualizing digital collections. Technical Services Quarterly, 32(1), 14-34. doi:10.1080/07317131.2015.972871

[6] Few, S. (2013). Information dashboard design: displaying data for at-a-glance monitoring (2nd ed.). Burlingame, CA: Analytics Press.

[7] GBIF.org (06 December 2019) GBIF Occurrence Download https://doi.org/10.15468/dl.tnmbnz

[8] Graham, M., Kennedy, J., & Downey, L. (2006). Visual comparison and exploration of natural history collections. Proceedings of the Working Conference on Advanced Visual Interfaces - AVI 06. doi: 10.1145/1133265.1133329

[9] Kräutli, F. (2016). Visualising cultural data: Exploring digital collections through timeline visualisations

[10] Kurihara, N. (2013). Utility of hair shafts from study skins for mitochondrial DNA analysis. Genetics and Molecular Research, 12(4), 5396–5404. doi: 10.4238/2013.november.11.1

[11] Liffick, M. (2008). MW2008. In The IMA Dashboard. Retrieved from https://www.museumsandtheweb.com/nominee/ima_dashboard.html

[12] Molineux, A., Zachos, L., Criswell, K. E., & Risen, T. (2012). GIS, The Key to Collections Management of a Large Research Archive. Collection Forum, 26(1-2), 60–69.

[13] Rao, S. (2017, April 14). When big data meets art appreciation. The Boston Globe. Retrieved from https://www.bostonglobe.com/lifestyle/2017/04/13/when-big-data-meets-art-appreciation/HqeuVGv9qdm2PGJAeYAuZK/story.html

[14] Ravenscroft, J., Liakata, M., Clare, A., & Duma, D. (2017). Measuring scientific impact beyond academia: An assessment of existing impact metrics and proposed improvements. PloS One, 12(3), e0173152. doi:10.1371/journal.pone.0173152

[15] Shirey, V. (2018). Visualizing natural history collection data provides insight into collection development and bias. Biodiversity Data Journal, 6(6), e26741-8. doi:10.3897/BDJ.6.e26741

[16] Smith, V., Paul, D., Woodburn, M., Grant, S., Singer, R., & Love, K. (2018, August 14). Shining a New Light on the World's Collections. Retrieved from https://www.idigbio.org/content/shining-new-light-world's-collections.