https://doi.org/10.2352/issn.2168-3204.2021.1.0.18

Visualizing for Different Audiences: Various Views

Fenella G. France; Library of Congress; Washington, District of Columbia, U.S.A
Andrew Forsberg; Library of Congress; Washington, District of Columbia, U.S.A
Andrew Davis; Library of Congress; Washington, District of Columbia, U.S.A
Hadley Johnson; Library of Congress; Washington, District of Columbia, U.S.A

Abstract

The past year of off-site telework allowed Preservation Research and Testing Division (PRTD) staff to do a deep dive into serious considerations about data analytics and data visualizations. Much of this work related to utilizing the tools we had available, linking the visualizations to data analytics for cultural heritage and heritage science research projects, with a strong focus on how best to adapt visualizations for specific audiences, ranging from scientific colleagues, conservation and collection care, interested public, and personnel wanting to use the information for a range of decision-making functions. Some of the factors we assessed related to the amount of information or data presented, whether to present minimal data with hover-over functionality to encourage exploration or allow different views for different audiences, what was "too much" data, what programs people were familiar with and the types of presentations, graphs, scatterplots, bar-charts, interactives etc. Significant discussions and reworking of visualizations answered some questions, while exposing many more.

Background

The preservation research projects in PRTD at the Library of Congress (LC) generate a lot of data that creates new knowledge about our collections. In alignment with our strategic goals for greater accessibility to collections and research created at LC, we have been developing a data visualization module focused on the best ways to share that data for a diverse range of audiences. This has been brought even more to the fore through the Andrew W. Mellon funded "Assessing the Physical Condition of the National Collection" (ANC) project [1] where we are being asked to share complex data with many factors involved to multiple audiences with a range of needs for how they will use the data. Time necessitates that we cannot completely rework these visualizations depending on the audience, so the question has been how best to create a multifaceted approach with in-house bespoke solutions based on Open Source software as well as commercially available tools, that resonate with our partners and collaborators, researchers, scholars and the general public. A further complication is the most challenging question we are often asked: "who is your audience?" This can and likely will change depending on who we are working with.

Challenges

The problem, apart from limitations of "fitting data" to what is allowed in many off-the-shelf (OTS) software packages, continues to be trying to understand the commonalities and differences between viewers, their likes and dislikes. This has exposed a fascinating conundrum as we balance the types of graphs we are used to using as scientists with how to present the same data in a user-friendly manner to educate, increase interest, and assist with decision-making.

The intended audiences for this research data visualization include, but are not limited to: associate librarians, collections care personnel, head librarians, conservators, heritage scientists, and administrative and preservation managers. While there is some commonality stated among these users, potential usage for the data generally included being able to guide decision-making for preservation, as well as identifying what volumes to retain or withdraw based on the assumption that other "better condition" same volumes were available for loan from partner institutions. There were also comments from our meetings and discussions about using this data to determine where research libraries should be focusing and justifying fiscal resources for new acquisitions. Further, once we have generalizations about subject headings, publication location or decade, we may be asked if this can then be used to survey internal collection data to review whether individual collections are "at-risk" and need attention, should be digitized immediately before further loss, or are simply not cause for concern.

The data sets we used were created and extracted from the ongoing ANC research project, and continue to be updated as we analyze more volumes. We began collecting this data early in the second half of 2019. In the lead up to actual data collection, significant work went into refining measurement protocols, the file export structures from instruments, and file naming and formats in order to allow for ease of interaction and extraction into multiple data visualization programs, approaches, and programming packages. Tidy data is of course a huge issue with extant data sets [2] and one of the data challenges we had from the beginning of the project was how to extract and correct the OCLC data sets from each of the five research libraries, since rechecking the catalog information against the actual dates of publication, volumes and editions revealed huge inaccuracies between the existing catalog information and the physical objects.

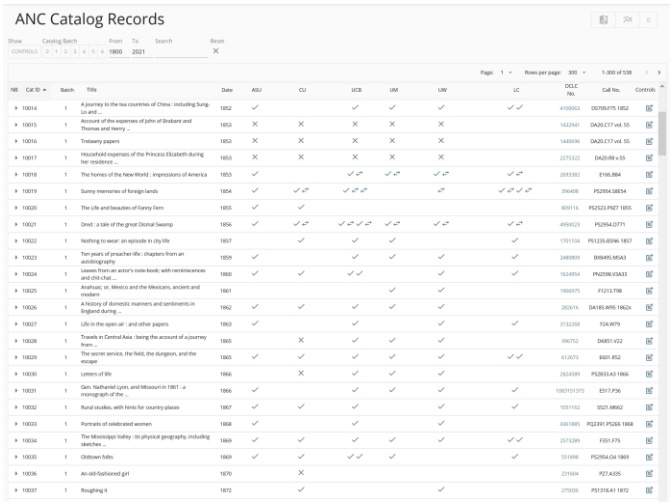

Figure 1. *Illustration of inconsistent catalog data. Ticks indicate the book matches with what OCLC claims the institution has; crosses, that the date published is incorrect, or that there is some other significant cataloging inconstancy, such as the edition was published in a different city and country; left and right arrows together, that the edition is correct, but a volume other than the first was received. In some cases, only a second or later volume was available, despite the records on OCLC.*

Data Visualization Methodologies

We investigated how best to represent the data sets for various audiences by combining a range of visualization methods, starting with the creation of online Query tools for ANC to assess data trends and look for correlations through 2D and 3D plots. This effort was augmented with chemometric tools, primarily Principal Component Analysis (PCA) and other multivariate statistics [3]. These allowed groupings of complex data and seemed to be more intuitive for heritage science colleagues. Some colleagues felt the scatterplots and complexity of the data was difficult to interpret and required an intimate understanding of the chemometric tools and approach, especially with respect to PCA scores and what the groupings meant. This approach for visualizing the data had been used to bring the extremely large number of variables we had in the data down to a more intelligible human scale, and for to more easily find variables that mapped together or had no discernible relationship.

We investigated multivariate approaches as well as other ways of visualizing the data to take this further, since being able to find and illustrate subtle trends in data was important for developing new, simple, onsite tools for collection care staff to quickly assess large collections and identify the "at-risk" sections. Figure 2 illustrates the visual aspect of a PCA plot. In this case, simple paper yellowness (using colorimetric b*, measured independently of the PCA model) seems to illustrate a connection to poorer condition of the paper-based collection materials, with pH and strength properties separated along principal component axes PC1 and PC2.

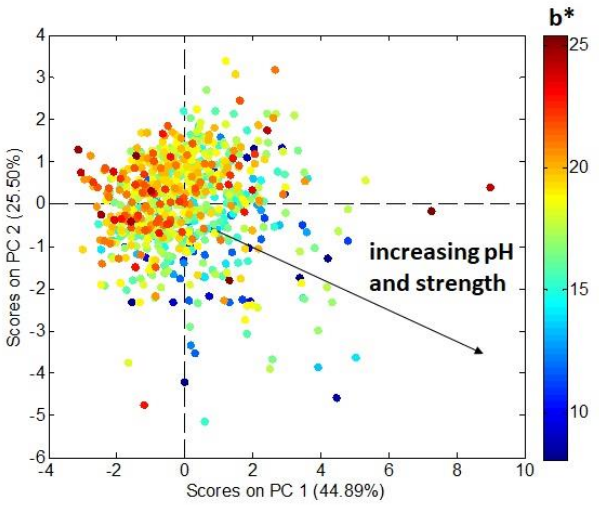

Figure 2. *PCA scores plot from test samples' physical and chemical condition, colored by paper b* (yellowness). Overlaid arrow indicates the general direction of increasing paper strength and condition (by pH and tensile strength) relative to the principal component axes PC1 and PC2.*

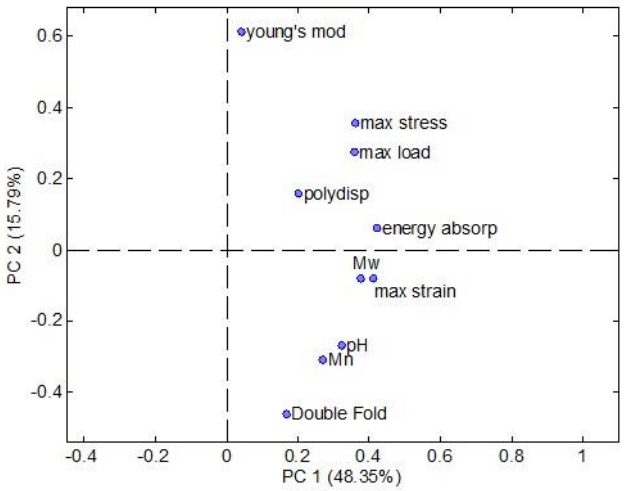

Figure 3. *Loadings plot from PCA of ANC samples, including most chemical and physical laboratory test variables. Variables close to each other in principal component space are indicative of likely correlation or redundancies between those variables for describing variations within the data.*

In addition, the wide range of test variables and instrumental outputs can be quickly checked for correlations or redundancies using a PCA loadings plot, shown in Figure 3. Variables with similar loadings in principal component space are likely correlated and may also be redundant, since they serve to describe the same variations in test condition. Using Fig. 3 as an example, pH and number-average molecular weight (Mn) appear highly correlated, which makes sense from the standpoint of cellulosic degradation, but less intuitive correlation are apparent too, such as that between weight-average molecular weight (Mw) and maximum tensile strain (max strain). Without this ability to quickly determine which variables

were the most critical for mapping to condition, we could spend literally days painstakingly trying to look for correlations that may not exist or may differ between specific periods for books in the 1840-1940 timeline.

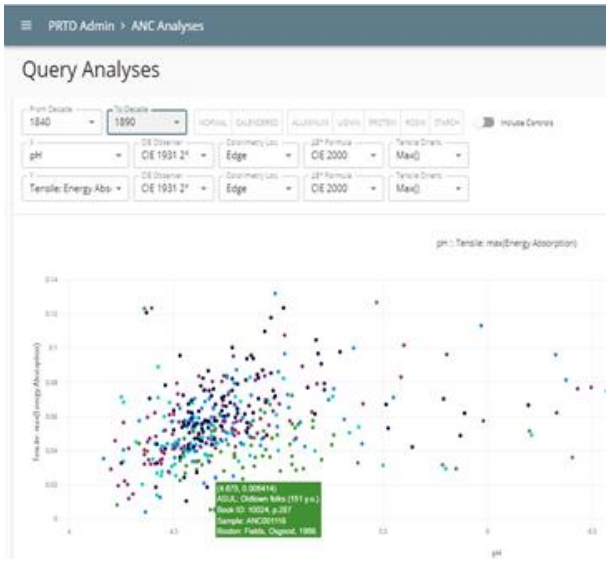

Figure 4. *Using the Query Tool to Quickly View Data Relationships*

Since teleworking allowed a greater capacity for staff to work through learning and expanding data visualization with new software tools, we assigned a PRTD staff member to work with Tableau (potentially available at LC in the near future) to create alternate ways of visualizing the data in a more interactive way [4]. This also aligned with a wide range of investigative questions from partners, as well as our own interests. For instance, how might the location, publisher and/or genre or subject category of the book volume relate to the condition? And, did publishers use lower quality textblock paper for popular fiction, and higher quality for reference books? We wanted to find ways to move from more traditional scientific plotting approaches to offer interactive representations of subjective data, allowing users to engage with, discover, and choose the specific portion of the data they were interested in. For example, they could hover over a city to see what types of damage seemed more prevalent from that location or publisher, and they could zoom in to just look at, say, New York, or expand out to view a world map where different-sized circles indicated the number of books from each city.

Feedback from viewers was highly variable. People who were more familiar with the data had very specific types of data representation they wanted to see. Others, seeing the visualization for the first time (for some it was even their first time seeing Tableau data representations) were alternately confused by the amount of data or simply fascinated with the different ways they could engage with it.

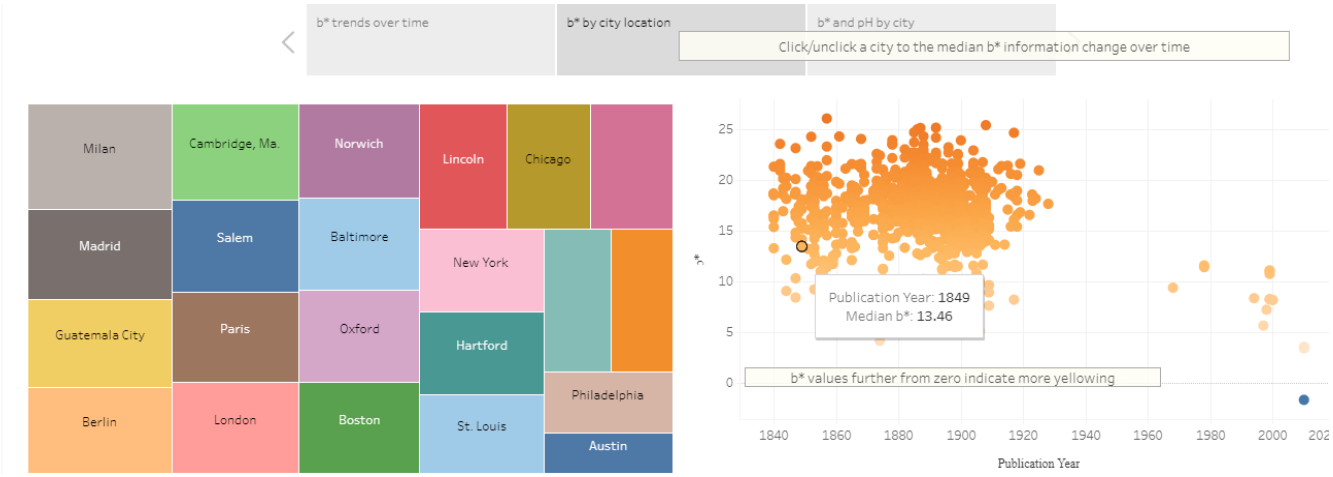

Figure 5. *Visualizing b* (yellowness) versus publication year and location, to predict potential damage.*

Results

We found that the combination of options, depending on the audience, was helpful for reviewing our data (see figures 4 through 6). The downside was that as soon as we created one visualization, we thought up a new version or alternate that could be more effective. As we added levels of complexity, the benefits of Tableau's tabs became more and more evident: we could alternate quickly between views with more or less information. The interactive components also allowed users to zoom in, hover over, and focus on the aspect that most interested them. Again, we had some viewers wanting less, others wanting more information, making how we utilize Tableau very much an ongoing "work in progress" in conjunction with our other Data Visualization Projects with IIIF.

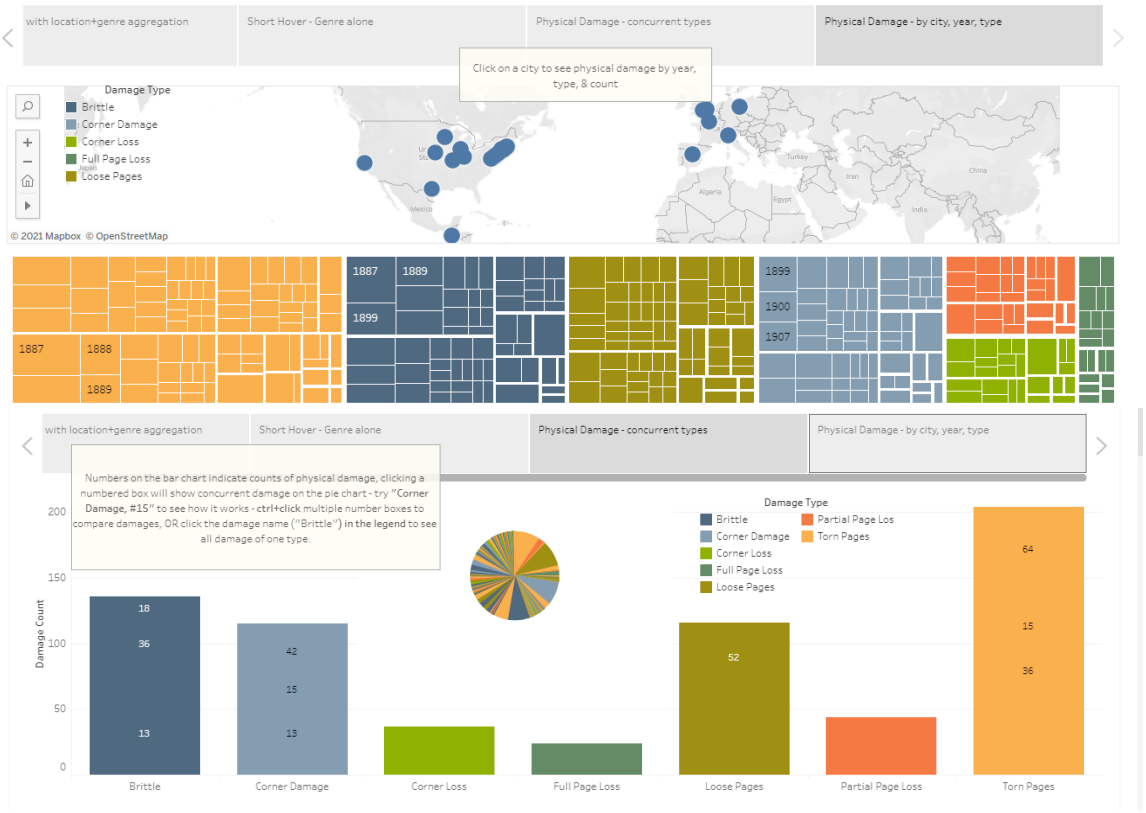

Figure 6. *Visualizing data by Genre, Year and Type of Damage.*

Conclusions

Our approach to visualizing data was to both utilize the tools we had available and to create new interactive tools that reflected the needs of an expanding and increasingly complex data set. The work involved linking the visualizations to data analytics for cultural heritage and heritage science research projects, with a strong focus on how best to adapt visualizations for specific audiences, who ranging from scientific colleagues, conservators, preservation professionals in collection care, and management personnel wanting to use the date for diverse uses of decision-making. Connecting data analytics and visualizations was an excellent way to explore the data generated by many projects, as well as evaluate the best options for sharing, educating and creating interest in those data visualizations. We undertook a multifaceted approach to creating new tools, using deep dives into chemometrics, as well as exploring OTS and online accessible software. We delved into the challenges of multiple ways to present heritage science data for multiple audiences, as well as its potential to be used as a preservation assessment tool and for decision-making. We quickly met the conundrum of "you can't please all of the people all of the time" and had an extremely useful series of ongoing discussions, reworking data views and engaging with explorations into what worked and what did not. We look forward to discussing this and hearing more lively, engaging, and diverse viewpoints.

References

[1] ANC public-facing site: https://nationalbookcollection.org
[2] Tidy Data: https://vita.had.co.nz/papers/tidy-data.pdf
[3] https://en.wikipedia.org/wiki/Chemometrics
[4] https://www.tableau.com/

Author Biographies

Dr. France, Chief of the Preservation Research and Testing Division at the Library of Congress, researches non-invasive techniques and integration and access between scientific and scholarly data. An international specialist on environmental deterioration to cultural objects, her focus is connecting mechanical, chemical and optical properties from the impact of environment and treatments. She maintains collaborations with colleagues from academic, cultural, forensic and federal institutions through her service on a number of international bodies. In February 2016 Dr. France was appointed as a CLIR Distinguished Presidential Fellow.

Dr. Forsberg, a Preservation Researcher in the Preservation Research and Testing Division at the Library of Congress, previously a CLIR/DLF/Mellon Postdoctoral Fellow in Data Curation for Medieval Studies, researches using internet-based technologies to improve data sharing and collaboration between the sciences and humanities in cultural heritage institutions. He has been a professional in the web development industry since the mid-1990s, and an academic researcher and lecturer in Medieval and Early Modern literature and literary theory.

Dr. Andrew Davis is a chemist in the Library of Congress's Preservation Research and Testing Division. He studies paper and polymeric materials such as cellulose, adhesives, and modern media, with the goal of applying fundamental polymer science to collection-related materials for better informing the preservation of Library collections. Andrew is also involved in work to better understand the role of light, oxygen, and material order to better enable public display of light-sensitive objects. Andrew received his PhD in Polymer Science and Engineering at UMass Amherst and has previously worked in the Central Research Laboratories of 3M.

Ms. Hadley Johnson is a Preservation Technician in the Preservation Research and Testing Division at the Library of Congress. She graduated in 2020 from George Mason University, adapting quickly to a "virtual" work environment and is responsible for all the Tableau data in this paper.