https://doi.org/10.2352/issn.2168-3204.2023.20.1.33

Preservation Equity: Decision Making and Data

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

Abstract

One of the continued challenges for preservation resources is the demand for objective data to make informed retention and withdrawal decisions. Discussions within the shared print and print repository communities have circled around the integral question pertaining to the selection process of books to be incorporated into a national shared print system, namely the minimum number of copies such a system must maintain. The challenge has been knowing the condition of those volumes that are withdrawn or retained, since all decisions have been based solely upon a shared catalog where partners do not have data to know the condition of others' volumes. This conundrum led to a national research initiative funded by the Mellon Foundation "Assessing the Physical Condition of the National Collection" to create a baseline of understanding of the actual condition of the national collection in research libraries and collections. The project undertook an extensive assessment capturing data from 500 "identical" volumes each from 5 different research libraries and analyzing the dataset to answer the following questions: What is the general condition of library collections in the 1840-1940 period? Can the condition of collections be predicted by catalog or physical parameters? What collection assessment tools help determine a book's life expectancy? Filling the gaps in knowledge for understanding the physicality of our collections is helping us identify at-risk collections and explain the high percentage of dissimilar "same" volumes due to the impact of paper composition. Predictive modelling and simple assessment tools allow more accurate prediction of good and poor-quality copies of books, as well as what is typical and atypical for specific decades.

Introduction

Currently in the United States, there is a national system for managing academic library collections across multiple institutions, which minimizes redundancy and the associated costs of traditional single institution collection management. Individual libraries, through local consortia and regional networks, work more frequently collaboratively to provide access to printed materials, making decisions about what to keep and what to discard as an interdependent collective of academic resource providers. Previously, local decisions governed by traditional principles of local ownership, are increasingly made within a wider context of responsibility. These consortia, collaborations, and regional partnerships are generally referred to as 'shared print' projects. *These efforts in shared print management have relied almost exclusively on the content of books to determine retention, without considering the physical condition of the volumes in question*. Without this knowledge, books that share an identical catalog description—same date of publication, same edition, with identical content—are usually treated as equivalent duplicates, which can lead to the retention of books that are physically quite distinct because of usage, storage, stack location, and other differentiating factors. The most egregious result of this content-only selection process is to retain duplicate books for shared print facilities that are materially near the end of their lifespan and will quickly deteriorate, defeating the very purpose of retaining those volumes that were to ensure the longevity of the collective. A number of shared print and future of the print record initiatives noted the need for objective data to assist with decision-making [1,2].

As part of the "Assessing the Physical Condition of the National Collection" [3], we created an extensive database of condition data for 2500 library volumes from five large research libraries in the United States over the 1840-1940 time period when printed books were moving from the more stable rag papers to acidic wood pulp papers as the paper manufacturing industry expanded. This preservation research evolved from the challenges faced by many institutions, where they were making withdrawal and retention decisions based upon subjective and incomplete information. The complex dataset includes cataloguing, descriptive, visual condition assessment and scientific (physical, chemical and optical) data from at least one paper type in every volume, with data collected from a representative statistically stratified random sample of the national collection. The complexity of this data for assessing paper-based library collections includes historical cataloguing challenges, subjectivity of visual condition assessments, and improving the capacity to describe visually the physical and structural information that impacts condition. The final issue is linking these multiple data points with the scientific physical, chemical and optical datasets, translating information to knowledge. The intertwining of economic and societal impacts on the changes in paper production and publishing, along with use and environmental factors, led to increased complexity, with the need to integrate multiple data types. The research data demanded an intensive data analytics approach to enable extracting the major variables and inherent paper properties that put specific components of our collections at higher risk.

Challenges

The first three data problems we tackled: 1) what was the condition of these supposedly "identical" books; 2) how to move from the existing lack of data and attempt to link subjective (if it existed) assessment with more objective scientific assessment, and; 3) how to fuse data from the multiple test methods to find the most efficient and best predictors for condition. The immediate problem that exploded the project scope: within the first shipment from each of the partners of the "same" volumes, based on the catalog information, these "same" volumes were of different sizes and thicknesses, had multiple paper types within one volume and the catalog records defied easy interpretation.

Therefore, even before we started working on the scientific condition assessment data, we had two new challenges, how to determine what was a cataloguing error and defined "not the same", and then how to sort through and coordinate the catalog

data. We needed to create an online platform to allow ease of comparison for catalog data to help with these initial unexpected issues. Once the descriptive and defining catalog metadata had been captured, the next step was to include capacity in the platform for visual descriptive data. The visual images characterize specific differences between these "identical" volumes often dramatically, and served to provide an enduring access to how similar or different the catalog data replicated the actual title, dates, publishing location etc. captured from the front page, spine, etc. (see figure1). Along with other statistical diagnostic tools, the next step was to incorporate and start to analyze the scientific dataset (five physical, chemical and optical analyses per volume) for trends. These potential patterns and relationships would then help us start to see connections or variability within the huge dataset that would point to condition data for decision-making.

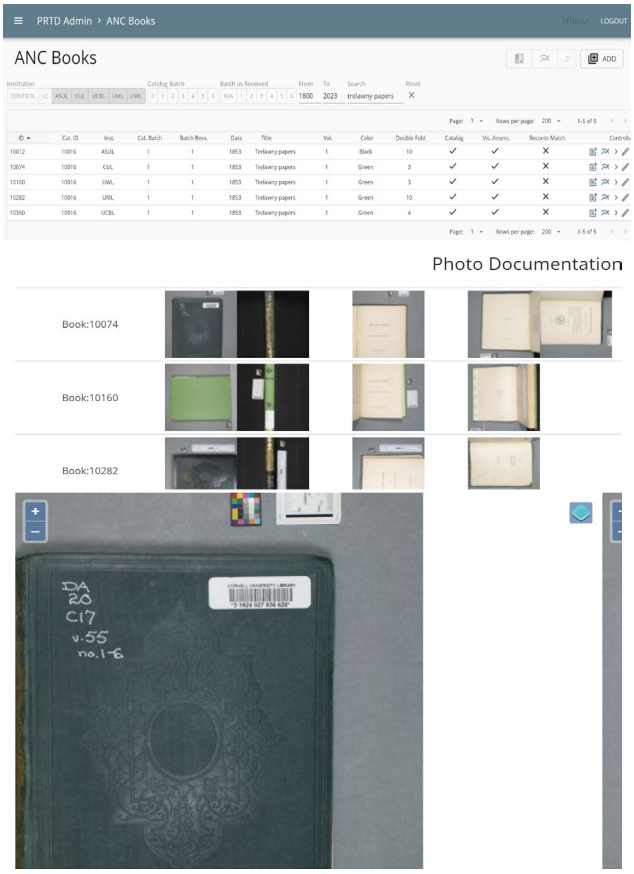

Figure 1. *Catalog and Photo-documentation for "Identical" Volumes*

Research Approach

To start to address the first cataloging challenge, the online platform was created with a Catalog" comparison tool. The dataset began with a "visual assessment" that captured a snapshot of each of the supposedly identical 500 volumes. This data collection was based on a IIIF [4] photo-documentation platform that enabled interaction with the catalog and visual assessment. The images were important: once the volume returned to the partner institutions, we would no longer have access to the catalog data that was in the physical volume -- a critical component to comparing to the online catalog information. From this database, was built a Compare tool that began to seamlessly integrate more complex continuous data, reports on individual volumes, and allowing visual comparison of images and catalog data between "identical" volume sets. The data terminology that described the "visual assessment" were based on linked open data (LOD) heritage and scientific terms from a range of ontologies. A selection of linked ontologies – a "bridged" Knowledge Organization System (KOS) [5] that we created since no one ontology included all the terms needed. To standardize the visual assessment criteria, we constructed a "visual terminology" implemented with IIIF images to reduce subjectivity in the visual assessment data capture. As illustrated in figure 2 for every term that described condition, two or more images that visually explained how these terms were captured.

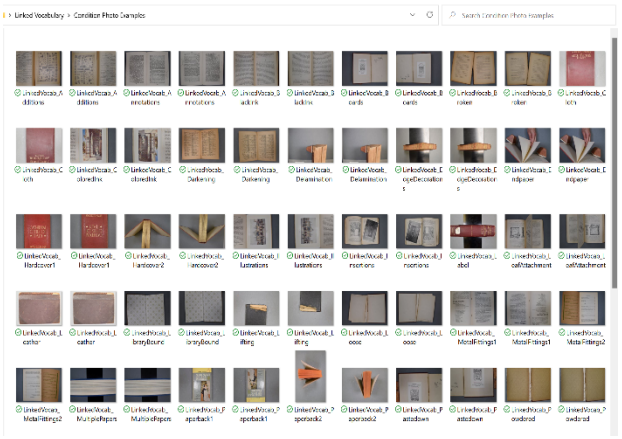

Figure 2. *Visual Terminology of Condition Assessment Terms*

To understand the "condition" of these identical volumes, the online platform was expanded to incorporate and integrate diverse scientific instrumental techniques, both destructive and non-invasive; such as size exclusion chromatography (SEC), tensile testing, pH acidity testing, Fourier Transform Infrared Spectroscopy (FTIR), and Fiber Optic Reflectance Spectroscopy (FORS). Partners had allowed us to remove a 3/8" (9.5mm) strip from a page without impacting any text or annotations and test methods has been miniaturized to ensure statistically significant data from very small samples. All the data was extracted raw to ensure we created FAIR (findable, accessible, interoperable, reusable) scientific data principles [6]. Too many research programs do not allow for reusable data, and we knew that with the large amount of data being collected, and the critical need for effective decision-making, that the dataset needed to be future-proofed. Within the platform we developed an expandable data storage and querying model that took advantage of key technologies in the Apache Software Foundation's CouchDB [7], including the latter's "stored views" and REST API. To quickly interrogate this expanding dataset, we created a Query tool that generated 2D and 3D plots to quickly search for potential correlations between condition data components and use these for additional chemical statistical analyses.

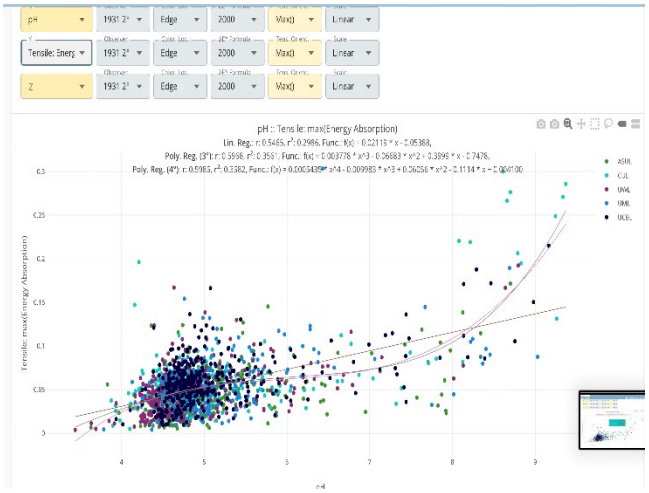

Figure 3. *2D Plot of pH vs Tensile Strength to Query Relationships*

Results

The Query tool enabled us to quickly review links and disconnects between the groupings of the condition and characterization of paper type. To date, one of the useful predictors of condition relate to the paper type – is it rag (generally more stable) or the more acidic paper pulp. Utilizing fully characterized paper reference samples from the Center for Heritage Analytical Reference Materials (CHARM) [8], we could compare project test data to data for these benchmark samples. We have used these data sets with their Fourier Transform Infrared Spectroscopy (FTIR) measurements alongside chemometrics [9] and statistical modelling with Principal Component Analysis (PCA) [10] to develop additional analyses, including a pulp predictor model to allow separation of paper compositions.

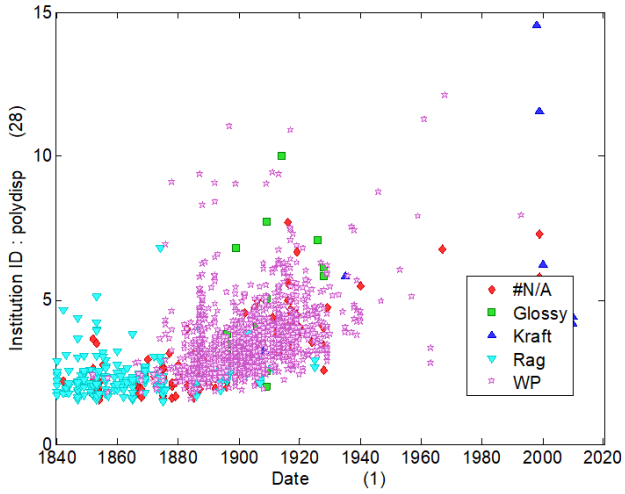

Figure 4. *Pulp Prediction, Analyses by Grouping of Paper Type.*

In figure 4, as the time period moves from 1840-1880, the bulk of the volumes transition to wood pulp (WP)-based papers, but we also discovered that there were a number of volumes whose paper that did not fit the model (red diamonds). These anomalies (as well as those in the pre-1880 period), related to the extremely experimental period of paper production at this time: paper manufacturers experimenting with different processing techniques and combinations of paper pulp.

Table 1. *Nineteenth Century Paper Production Developments*

Date	Paper Production Event
1837	*The Panic of 1837 in which over 600 banks fail overnight. Many paper mills could not make their margin calls on loans and go into receivership. In the aftermath, new material including manila rope, rag-bale ropes, hemp sails, canvas sheets, rope, yarn, and burlap. Manila paper first produced at the West Groton Mill in CT, with other mills soon taking up the trend heling to relieve the strain on the rag market.*
1830s-1860s	*The quantity of patents filed related to preparing wood pulp and straw for paper production by both chemical and mechanical means greatly increases, from approximately 2 a year to more than 30 a year. Bleaching also becomes common practice to produce white paper.*
1851	*Hertfordshire, England: Hugh Burgess and Charles Watt are successful in making pulp from wood by chemical process (boiling wood in caustic alkali at a high temperature, with possible substitution/addition of chlorine or hypochlorites for the caustic alkali). A patent follows in 1852.*
1854	*Philadelphia, PA: Burgess moves to the US and secures a patent with Morris L. Keen, who had been working on a mechanical process for deriving pulp from wood. Continuing their experimentation, Warren Mill eventually transitions solely to wood pulp paper.*
1854	*Hamilton, OH: Piece felts needed for the cylinder-mold machine are made, rather than imported, for paper mills in the US by a woolen mill owned by Asa Shuler. This is an example of the first instance of US made replacement parts (covers, wires, screens, continuous felts, etc.) for the new machines, which soon becomes the standard.*
1863	*Royersford, PA: American Wood Paper Company organized and became the leading manufacturer of soda-pulp and paper.*
1860s	*London, Paris, US: Henry Voelter constructs a machine and invented a process for grinding wood into pulp. This process quickly traveled to the US, and in 1866, two of the new pulp-grinding machines were imported. The patent was bought for the US in 1869.*
1870s	*Sweden, US: After experimentation failed in the US in the decade preceding, the sulfite process came into practical use, in which sulfurous acid is used to dissolve intracellular matter of wood leaving fibers to be turned into pulp. The process spread to London and to the US in 1884. This led to a need for overcoming mechanical difficulties, in which digesters were designed and used to reduce the cost of repairs.*
Late 1800s	*During the latter half of the 19th century as a result of the great demand for books and news preceding and following the American Civil War, a series of price hikes in the manufacturing industry allowed for a vast expansion of the industry across the country, with 555 paper mills reported in the US in 24 states. The northeast, mainly New York, Massachusetts, Pennsylvania and Connecticut, had the largest concentration of mills. While many longstanding paper mills were able to expand and produce paper manufacturing empires with bigger mills, better machinery and improved methods of manufacture, many new mills were also built.*
Post 1910	*As a result of the above increase in paper mills, after 1910, the US was able to initiate a consistent exportation of paper pulp and paper materials, with imports of paper beginning to decline.*

Figure 5 illustrates the way that properties can vary depending on the sub-population of the paper type, here showing the distribution for changes in tensile strength (stress at break) from 1840-1940, for rag and paper pulp compositions.

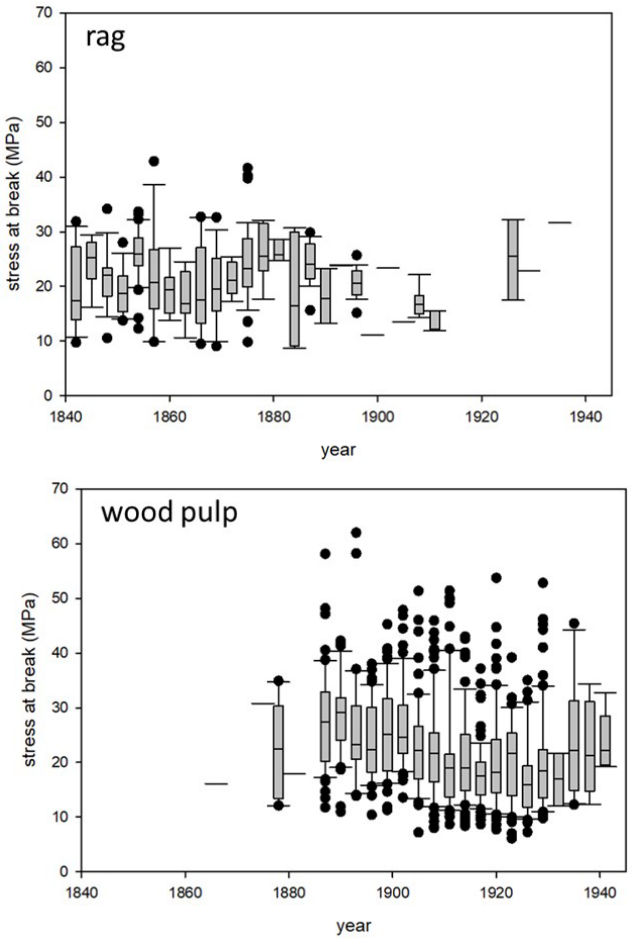

***Figure 5**. Comparison of Rag and Wood Pulp Stress at Break 3-year Periods*

It has been fascinating to see how much economic and societal data feed back into the scientific results that we have captured, and how this history of paper production has led to consideration of impacts and changes in paper composition from these factors, that impacted how to best predict "at-risk" time periods. While initially we had considered using decade intervals for analysis, data indicated that a 3-5 year time interval was better at capturing changes in paper production methods.

What was deemed "typical" from the data for those specific time periods was crucial in order to better separate and identify time periods that showed poorer quality volumes. As part of this examination, we also wanted to move towards describing volumes as either "typical" or "atypical" for that time period, enabling us to look at those that were above and below the grouping of the norm for that time period. Why were some in better/worse condition? This framing also empowers collection managers to consider their own appetite for risks related to what may be "typical" or not. The additional component in relation to "typical" is connecting these data to identify what production changes and/or unusual additions to the paper, are causal in imbuing these volumes with characteristics outside the typical for that time period. This has led us to multiple approaches of how to disseminate and share the data for determining time periods or volumes more at-risk, in ways that link to catalog information that can be easily accessed. For example, a reprint within 5-10 years is often considered the "same" given that the content has not changed. However, we have examples where we can observe a distinctive change in paper composition of books reprinted in 1918 when the original volume was published in 1911, as illustrated in figure 6.

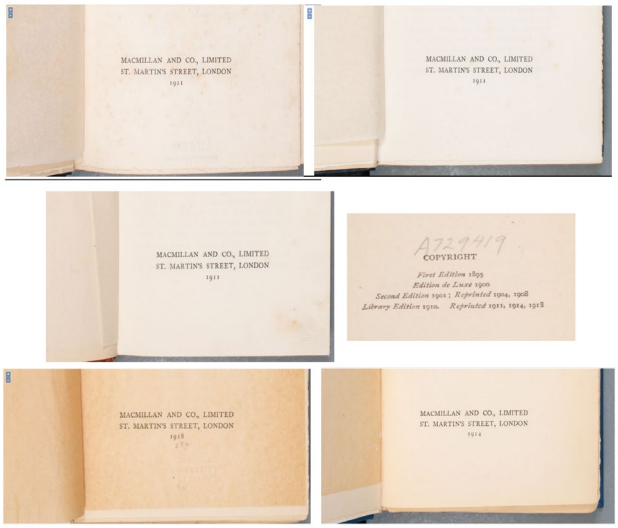

***Figure 6**. The "same" 5 Volumes: Upper 3 Images – Rag Paper; Lower 3 Volumes – Wood Pulp Paper Composition*

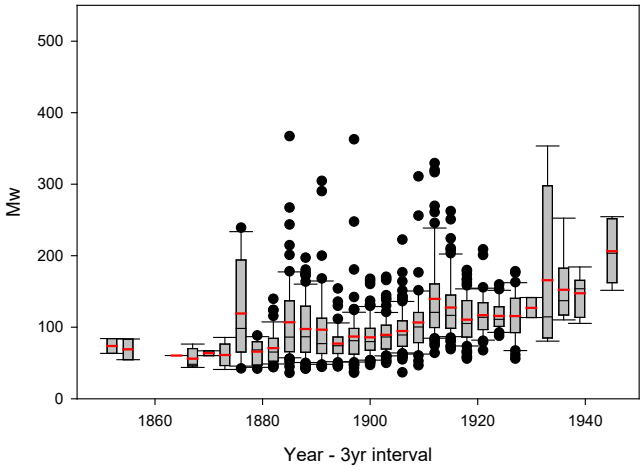

***Figure 7**. Identifying "At-Risk' Time Periods – Volumes Outside the Typical*

Figure 7 outlines how those "typical" result for volumes within the same 3-year time period can be clustered, while there are some outliers that have results indicating better, or less optimal physical and chemical properties.

Further data analytics reveal links between aspects of color space and condition that have led to the review of use of small portable colorimeters that would capture objective standardized color data: we could then link to a slider that divided color data into categories of risk.

One of the continuing challenges: how was the data from this first large statistical representative sample of the condition of

paper-based volumes from the higher-risk period of introduced acidic paper pulp, could be utilized for improved decision making, so that the condition of volumes was taken into account before possibly good quality volumes were withdrawn. Discussions with colleagues in collection care, and other organizational decision-making levels, indicated that lack of personnel, and resources would often out-weigh the capacity to undertake a condition assessment. This was especially true when large quantities of books were required to be reviewed to address space allocation issues. Figure 8 starts to visualize the complexity of the types of damage and how that narrative often needs to be extracted into smaller sections, while also recognizing the multi-dimensional aspect of the data set.

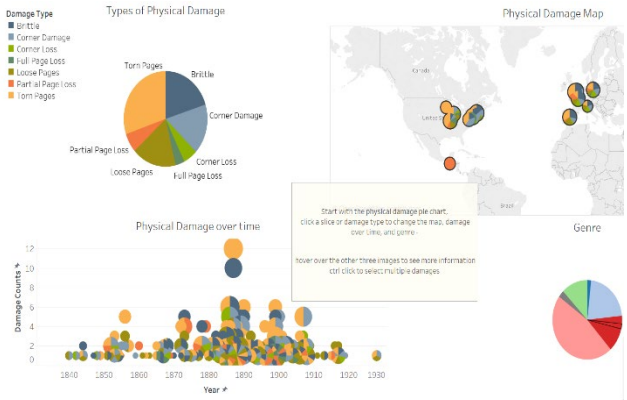

***Figure 8**. Translating Data to Create a Risk "Matrix" for Identifying "At-Risk' Materials*

Many have asked us why we did not use artificial or augmented intelligence (AI) on this large dataset. As the complexity of the above data shows, we did not want to fall into the trap of over-training data sets and following accepted assumptions. It has become apparent that heritage institutions need to address and review current research library collection assumptions. For example, the data quickly invalidated the use of the "double fold test" [11] to predict strength or condition. Our tensile data illuminated the fact that contrary to accepted library lore, it was almost as likely that the paper would be strongest in the vertical direction as it was in the horizontal. That is, books in this 1840-1940 time period were not necessarily bound to make best use of the papers' machine direction (figure 9).

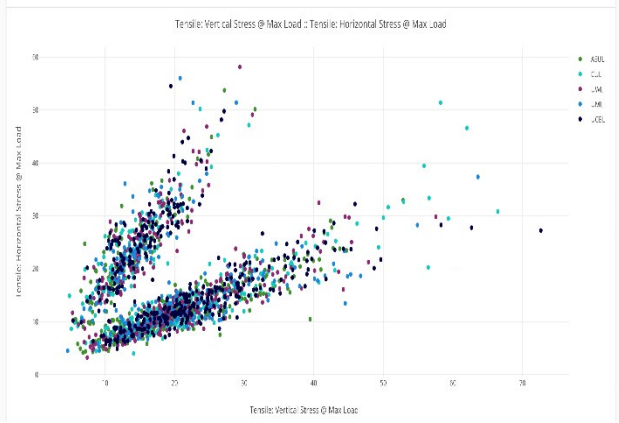

***Figure 9**. Horizontal versus Vertical Strength of Paper*

One of the other broad assumptions we had stated at the origin of the project, was that there would potentially be three major impacts on the collections being analyzed:

- Environment
- Use
- Inherent Paper Properties

What quickly became evident that there were multiple paper types in one book, and these were not just photo paper inserts, they were clear sections that demonstrated significant changes in visual condition (figure 10). These papers in the one volume had always been together since the book was published and bound, and so, these different paper types have been exposed to the same environmental conditions and use.

***Figure 10**. Example of Volume with Multiple Paper Types in one Volume*

To date, we are seeing up to 15% of earlier volumes that are showing visibly different paper types within one volume. The caveat is that we are observing this in the older volumes, and it is possible with some of the "younger" volumes, the change is not yet as visible, therefore this number may be higher.

***Figure 11**. Multiple Copies of Letters of John III, King of Portugal, 1521-1557' (1931)*

We have one example where we see four copies illustrating the same pattern in paper types within the volumes, whereas most other "same" volumes have a random difference. These are multiple copies of the Letters of John III, King of Portugal (1931), illustrated in figure 11. This shows that the printers swapped batches/paper stocks (knowingly or not) when moving between printing one set of plates and another. It seems reasonable to assume that at time of production papers were visually close enough to identical as to make no real difference. Here we see that over time, even the environment (across four distinct geographical

locations in the US) has played much less of a role than the impact of the original paper composition. While for preservation of large physical collections we need to consider and assess the impact of the material, the environment and usage, the inherent properties of the paper when the books were produced seems the be the most compelling aspect for predicting condition.

The other aspect of the data and trends we are seeing that while we had selected this random stratified statistical sample from shared "general collections" the data is directly relevant to unique and distinctive collections from this time period. The ability to predict paper types, and to explore the experimental paper production so we can provide this matrix and categories of catalog information that inform at-risk is critical data that needs to be disseminated, shared, and utilized. Further, as we are seeing such diversity in the composition and therefore longevity of the "same" volumes, that conclusions form this dataset need to be incorporated into collection care decisions.

Conclusions

The project data from the 500 "identical" volumes from 5 different research libraries has not shown any difference in data related to location of storage and more specifically environment. The large volume of books that contain multiple paper types, which have aged in different ways, is a clear indicator of the inherent property of the paper composition. This is the critical factor for determination of longevity of these volumes. As we share the "at-risk" components from the data, whether a specific decade and/or paper manufacturer, a printer known to use lower quality paper, or a period in time that related to an economic event, we want to assure that the collection assessment tools being developed allow collection care professionals to make informed decisions, and when possible, keep additional volumes, or good quality materials.

This project has allowed us to utilize a truly large data set to reveal new information about our physical collections. The gaps in knowledge about the physicality of our collections are being addressed so we identify at-risk collections and explain the high percentage of dis-similar "same" volumes due to the impact of paper composition. Predictive modelling and simple assessment tools are allowing more accurate prediction of good and poor-quality copies of books, as well as what is typical and atypical for specific decades.

Further, in relation to the diversity of paper composition and longevity of these materials, heritage organizations need to consider to begin using the data that relates to their collection needs at an institutional level. This is also a plea encouraging colleagues to consider the institution using data in a way that allows them to make those collection decisions for preservation of their unique and distinctive collections, which may differ from the needs of other institutions.

Creating interoperable data infrastructures to reuse and integrate heritage collections, enables the ability to extract more information for preservation. Creating tools that allow researchers to ask new questions of extant data, and integrate with new datasets, including climate, societal and economic data, opens possibilities for extraction of new knowledge from existing data.

References

[1] Robert Kieft, Future of the Print Record Report https://printrecord.mla.hcommons.org/files/2016/12/FPRWhitePaperDec2016.pdf
[2] Ithaka http://www.sr.ithaka.org/blog/the-future-of-the-print-record/
[3] ANC public-facing site: https://nationalbookcollection.org
[4] IIIF: International Image Interoperability Framework, https://iiif.io/
[5] Marcia Zeng – Knowledge Organization Systems (KOS) https://www.academia.edu/26672820/Knowledge_Organization_Systems_KOS_
[6] FAIR: https://www.go-fair.org/fair-principles/
[7] CouchDB: https://couchdb.apache.org/

[8] Heritage Sample Archives Initiative and CHARM https://www.iccrom.org/projects/heritage-samples-archives-initiative, Formerly CLASS: https://loc.gov/preservation/scientists/projects/class.html
[9] Chemometrics https://en.wikipedia.org/wiki/Chemometrics
[10] Principal Component Analysis https://en.wikipedia.org/wiki/Principal_component_analysis
[11] Explanation of the "Double Fold" Test https://dictionary.archivists.org/entry/double-fold.html

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, and a Board Member in 2020.

Dr. Davis is a Chemist in the Preservation Research and Testing Division at the Library of Congress. He is currently involved in work to analyze the Library's various paper and polymer collections, with the goal of correlating fundamental polymer properties to degradation processes in paper-based cultural objects. 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.