https://doi.org/10.2352/issn.2168-3204.2020.1.0.17

Access to Collections: Challenges of Physical and Digital – Assessing Digitization Decisions

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

Abstract

Access to collections is expanded through digitization, but are we saving the "best" volumes, which volumes are the best, and how do we make that decision? Capturing "real" collection data to objectively make and support those decisions is part of Library of Congress (LC) research. Current data suggests that most cultural heritage institutions have digitized less than 10% of their collections, so preservation of the print record is critical for long-term access to this knowledge. This is especially true for 19th and 20th century paper-based materials, where mass production methods resulted in less stable paper. Moving from subjective to objective based data for retention and withdrawal decisions is critical for the robustness of the print corpus and the future of digital collections.

Background

One of the challenges with assuring digital access to heritage collections is acknowledging that most of our collections are still in print-only form: more than 90% of most special collections have no digital surrogate and for general collections from 1840 to pre-Google, the results are also low. To advance our presence in the digital realm we need to understand more about the condition of special collections and what can be digitized, and depending on the condition, how effective decisions could be made for prioritizing digitization.

A national research initiative funded by the Andrew W. Mellon Foundation, "Assessing the Physical Condition of the National Collection" [1], is undertaking the task of providing data to objectively assess the condition of books held in the United States by performing an in-depth scientific analysis on a representative sample. The project is testing the "same" 500 volumes from six research library collections for the time period 1840-1940, since this period aligns with the advent of mass wood-pulp paper production and the introduction of acidic paper content. By doing so, we can determine which is the "best volume" for long-term preservation, and which criteria can add value to digitization. Proportional stratified sampling with time periods (decades) as strata allows a representative statistical study to be undertaken.

There have been previous studies overseas looking to assess the impact of various storage conditions using simple laboratory analyses such as pH, mainly to look at the impact of storage parameters on book condition [2]. For this study, given the challenges of tracking environment when collections move, and the much greater use of HVAC systems in the United States, we have yet to assess this component, given that the inherent properties of the text block seem to be having a greater impact. Additionally, we have imposed critical measures to statistically assess the impact of researchers and assure any results obtained remove any bias.

Historically, by 1830 there were 60 paper-making mills in operation in the United States. Papermaking moved from rag-based materials to a search for alternative fibers, becoming more critical during the civil war when demand accelerated costs. Steinway noticed sheet music from Germany was printed on paper from wood-pulp. He helped set up the first wood-pulp mill in 1867 [3].

Retention Decision-making

The motivation for this research was that many institutions are currently making withdrawal and retention decisions based upon subjective and incomplete information – it is difficult to ascertain which is the "best" book to save within a collection, and considerably more so across collaborating shared print institutions. The data will help ensure that large-scale withdrawal of materials does not compromise the overall soundness of cultural heritage collections, informing the shared print, preservation and digitization communities.

A number of shared print and future of the print record initiatives have noted the need for objective data to assist with decision-making. The project's ultimate goal is to fill gaps in our knowledge to guide the community as it develops a national print archiving effort as part of shared print initiatives, by answering questions about which materials are at risk, as well as allowing institutions to be able to predict with a strong probability of accuracy good quality and poor quality copies of books. The information will allow for pragmatic decisions about digitization and serving collections, and contribute toward advancing the widespread use of FAIR data principles (findable, accessible, interoperable and reusable). Until there is a focus on data being reusable, digitized collections will continue to remain only potentially available.

Collection Condition Challenges

The challenge that quickly became self-evident was "*what does identical mean?*" This was closely followed up by: can institutions trust the cataloging and condition information they rely on to make retention and withdrawal decisions? Current efforts in shared print management rely almost exclusively on the content of books to determine retention, without considering the physical condition of the volumes in question. For lack of better data, books that share an identical catalog description – same date of publication, same edition, and so identical content – are treated as equivalent duplicates. On the basis of assumptions such as these, we risk withdrawing books that are physically distinct due to their different original paper types and compositions, usage, storage, stack locations, and environments. Inaccurate catalog records risk the

withdrawal of irreplaceable books. Cataloging discrepancies are ubiquitous with some errors being more problematic than others. It has been estimated that even a small rate of errors can lead to serious challenges. A set of 1 million records that is 95% accurate would include 50,000 errors [4]. Informal conversations with colleagues suggest up to 3% of collection items are "not on shelf" and that decisions are being made based upon the catalog stating these volumes are present.

The research project addresses the challenges of the prevailing selection method and its potentially compromising, if unintended, consequences. Following upon a number of shared print and future of the print record initiatives arguing the need for objective data to assist with decision-making, this project will provide a data-based methodology to objectively assess the condition of the books held in the United States by performing an in-depth scientific analysis on a representative sample.

Research Methodology

To solve these challenges, we undertook research to compare the physical, chemical and optical characteristics of a selection of library materials across six large research libraries in distinct regions of the United States. The data generated will be used by shared print practitioners to determine the current physical state of items held nationally with the intent of identifying those materials that are in good condition, where they can be found, and inform institutions about the potential risk of loss of the printed corpus held within the country.

Our research starts with a formalized "*visual assessment,*" derived from practices currently undertaken in libraries, to better understand the challenges and specific issues with subjective assessment of condition, and to ultimately move away from this method. The first step is to simply check whether the book received is the same as the one listed in the master catalog of 500 titles. This involves confirming the title, author, volume number, dimensions, publication date and location. Using the same online form, we then keep track of the results of a double fold test conducted on the page to be sampled (whenever possible, all measurements are taken from the same page of each institution's copy), thorough descriptions of the book's binding and text block, and a condition assessment of both the binding and text block.

To better link and correlate the subjective and objective assessments, we collated a visual terminology to standardize how people are visually assessing condition and what seems to best relate to how people are making subjective decisions for whether to retain or withdraw a volume.

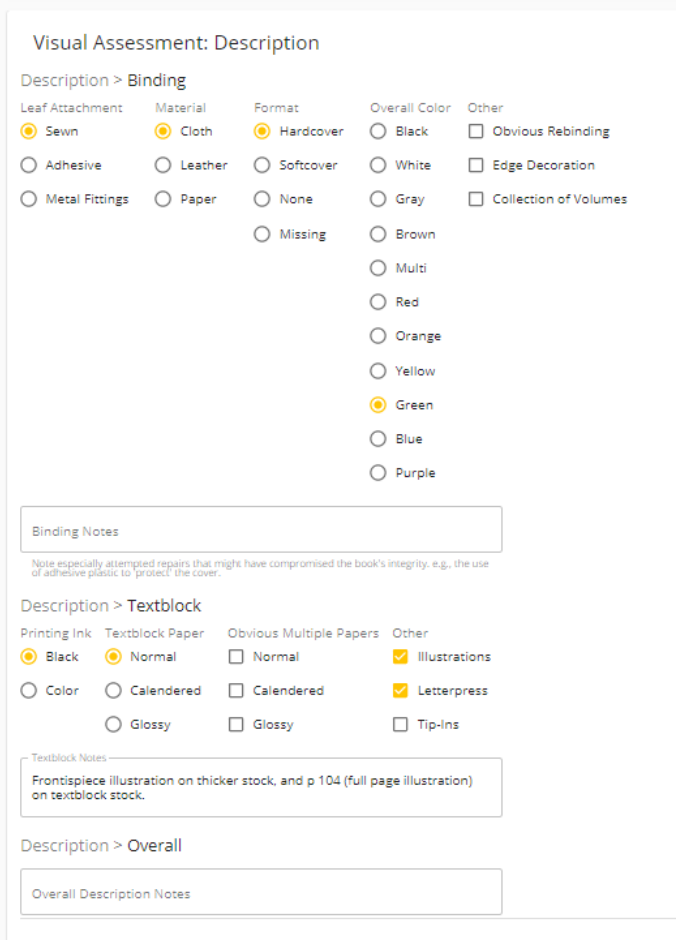

Figure 1. Visual Assessment Descriptive Information

Linking visual assessment to objective test methods has led to an advanced and reusable data infrastructure for re-interrogation of the data to follow trends, allow for correlations with the visual assessment, and the development of more objective and accurate on-site or stack collection assessment tools. Trends, correlations, deviations, and outliers can be identified within and across the subjective and analytical data collected for books, instances of the "same" book, institutions' collections, decades of publication, paper types, known reference papers, and so on. We are currently running Principal Component Analysis (PCA) analyses for infrared (IR), colorimetry data derived from FORS, some key visual assessment data, and the results from the invasive techniques outlined below.

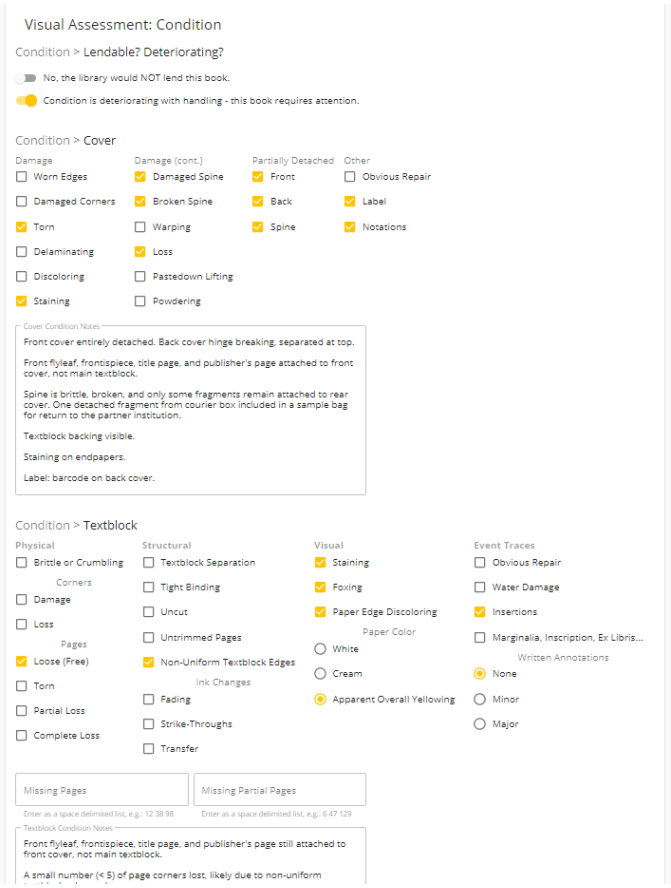

Figure 2. Visual Assessment Condition Category Separation

Condition assessments are organized so as to separate irrecoverable physical damage (loss, tearing, brittleness and crumbling) from damage to the structural integrity of the book (loose or missing covers, tight binding, text block separation, etc.), visual damage (staining, foxing, edge discoloration), and traces of events that might have damaged the book, but are not necessarily inherent to the paper or binding itself (e.g., water damage).

With our laboratory analyses, all efforts have been made to minimize the physical impact on the sample books by taking the smallest test sample possible that would still yield the necessary information for us to determine condition. The analytical evaluations selected for this project are outlined below. A strip of paper 10mm x 140mm is removed from a page edge, and this is all that is required for all of the destructive and non-invasive tests.

Fiber Optic Reflectance Spectroscopy (FORS) is used for non-invasive measurements from multiple standardized locations on the sample page of the ultraviolet and visible spectral regions. From these we gather data on paper composition and its impact on the paper's current condition, along with colorimetry data to quantify both color and color change (between the page edge, an inset location, and gutter) as evidence of the state of degradation. External Reflectance Fourier Transform Infrared (ER-FTIR) spectroscopy is a further non-invasive method we are employing to characterize differences between paper types as well as specific changes at the molecular level that occur with the degradation of materials.

Micro Size Exclusion Chromatography (SEC), pH, and tensile testing are the three primary invasive analyses conducted for this project, consuming the barest minimum of sample material possible with current technology. From these, and against reference papers with known characteristics, we gather, respectively, data for: the molecular weight and thus degree of degradation from undamaged paper cellulose; the acidity of the paper, since acid is a well-known catalyst for cellulose degradation; and accurate, quantified, measurements of the paper's strength. We have also conducted X-Ray Fluorescence (XRF) spectroscopy, and a barrage of standardized spot tests (for Aluminum, lignin, protein, rosin, and starch), for some institutions' collections in order to ascertain whether those tests would reveal useful information, and to confirm what we were seeing in the primary test suite.

Data from the EAST Boston Consortium suggests a 5% loss in condition from collection use (twenty additional checkouts). Using molecular weight data to measure reduction in strength can provide information about when a collection item is no longer fit for purpose [5]. In other research projects such as Collections Demography, "fit for purpose" relates to the usability of an item – the interaction between the inherent properties of the paper and the use of the item – how the institution assesses the ability to serve the collection item to users.

Results

The research is still in progress but to date our results have been startling. Even the simple catalog checks have been instructive – first, did we even receive a physical book from the partner institution that should correspond with the book requested? In some cases, we know the title is not currently available for them to send us. In others, it is clear the OCLC holding records are incorrect, that this institution simply does not hold that title.

Not infrequently, we receive a later edition, or a volume other than the first. In a few cases we have received a first edition, as requested, but the title was published in a different country. Three "books" do not even exist – they are instead sections in a miscellany. We have received 20th and 21st century facsimiles created by libraries and third parties, and these books are cataloged, stored, as if they were the 19th century originals. Other books, legitimate first editions from the same publisher in the same country, have arrived in different original formats (sizes, paper thicknesses, uncut or publisher-trimmed), different original bindings, and different original text edge decorations. In short, and as many suspected, OCLC's records are clearly not as accurate as one would hope.

Once partner institutions have had a chance to send books that they could not with their first shipments, we will have more firm data on cataloging accuracy from this proportional representative sample of library holdings for our first period, 1840 to 1900.

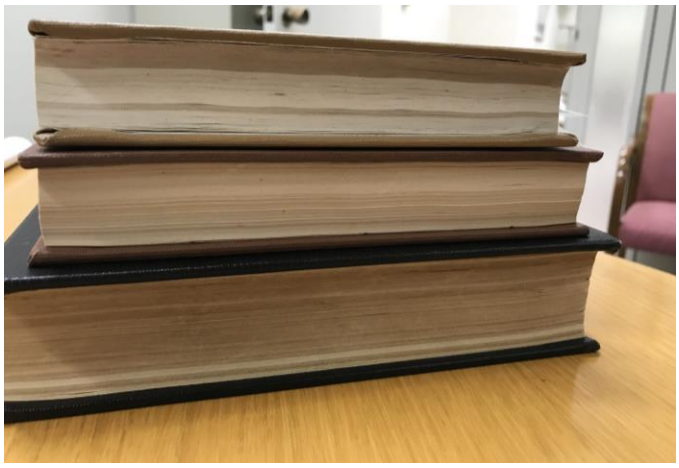

Figure 3. Multiple paper types within a single volume: text block edges.

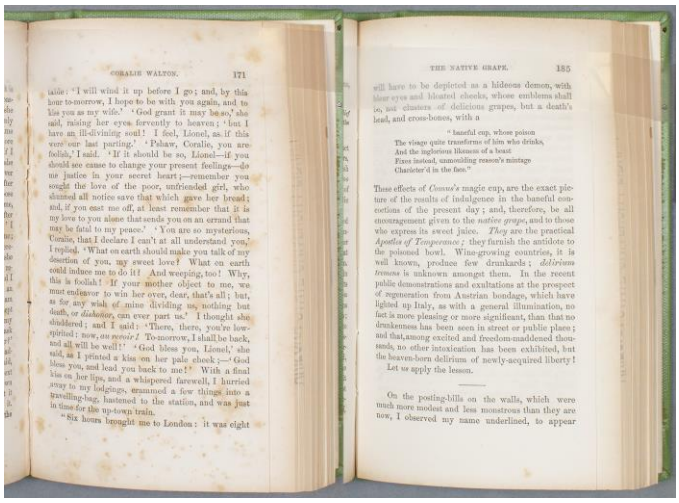

Figure 4. Multiple paper types within a single volume: "Leaves from an Actor's Notebook" (1860), pp. 171 and 185.

The complex of issues related to "sameness" is not confined to OCLC inaccuracies. At least for the mid-through-late-nineteenth century, publishers exercised no small amount freedom in their practices, even for a single edition. Several of these differences for the "same" book have played their part over time as factors in the book's current condition. Even identifying multiple paper types in our visual assessment, originally intended to facilitate record keeping for multiple samples from books with a mix of text block paper and many plates, has drawn attention to printing practices resulting in more than one type of paper becoming part of the text block itself.

These paper types have different properties and hence different preservation challenges. We are still investigating whether this is consistent for popular press books, or possibly a standard practice for printers, the result perhaps of stock availability from mills. There are suggestions of a "paper cartel" that we are still investigating. In the 1890s, Wayland claimed that because he had to pay the extortion of all of the trusts, he had to run his press in a capitalist fashion, and that this included the paper he had to buy for printing [6].

Research results have shown that the double fold is inconsistent and those institutions making withdrawal decisions based upon double fold may be withdrawing volumes that can still be circulated and digitized. Replicate testing at LC in the Preservation Research and Testing Division confirmed that multiple replicate test samples provided a wide variety of answers, confirming the inherent subjectivity and inaccuracy of this test [7].

The formalized book descriptions have helped identify some of the above-mentioned variations within and across editions of the same book, but also to identify full and partial rebinding efforts, previous owners and donations, paper types, plates, inks, printing methods, and so on. The longer-term effects of waves of institutional rebinding practices can often be seen in the current condition of books – of particular note, enthusiastic overstitching, and to a lesser extent, retrimming books to very nearly their printed text area, have severely curtailed the usability of some books. Overstitched books with paper that has since become (or was already becoming) brittle have suffered the worst.

Figure 5. Comparison of Tensile Stress at Maximum Load for samples in horizontal versus vertical orientations.

Early analytical findings have presented many fascinating avenues for further research, some promising to extend well beyond the scope of the present project. In some cases, initial data is interestingly counter to common beliefs regarding printing, paper-making, and binding. For instance, in figure 5, we see that it is just as likely for books to have been bound with the paper's machine direction oriented horizontally as it is for it to have been bound with a vertical orientation. The tool in the figure has been developed to aid with quickly grouping, filtering, and then comparing for any discrete datapoints assigned to x and y axes, from the full dataset as it grows. Other complementary tools are designed to aid with

comparing books, physical samples' data, and reference papers across any and all of the analytical and subjective data collected. See, for instance, figure 6, where Chromaticity diagrams are generated for visualizing color data from multiple points on the same page, different pages, different books, and standard reference papers.

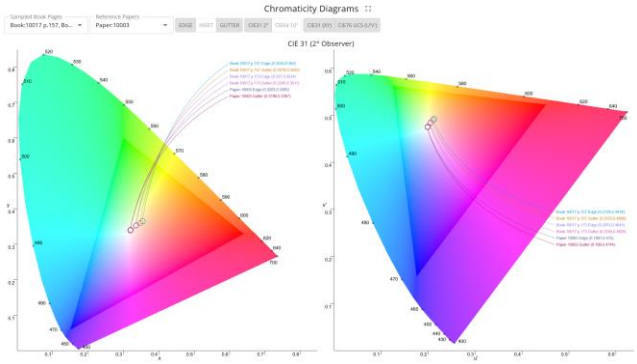

Figure 6. Dynamically generated CIE1931 (xy, left) and CIE1976 (u'v', right) Chromaticity diagrams for edge and gutter measurements taken from two pages in the same book, ISR5 measurements added for comparison.

One of the major benefits of the structure of this research project is the creation of active, reusable data, in a manner that allows constant re-interrogation of the data, and linkages to be made between the visual assessment – closely replicating the current methods of assessment of collections – and the objective laboratory analyses. Further, the correlation between invasive and non-invasive assessments of the paper condition is leading towards a methodology for the capacity to utilize non-invasive measures for quick assessment tools. A key goal of this research is the creation of more accurate "in the stack" tests that are simple, reproducible and useful for quick condition and retention decision-making.

Conclusions

The implications from this research are that better-informed decisions will support better digitization planning as cultural heritage institutions identify at-risk materials before it is too late, and can then digitize to preserve the knowledge and content within. There are significant challenges with not being able to quickly and accurately make these assessments for digitizing at-risk collections

Identical is not identical with collections having greater variability than expected for the same book, and evidence to date raising concerns about cataloging accuracy. The development of simple tools will assist collections' care in more accurately determining those papers that are most at-risk. The creation of a corpus of data from a representative sample of the national collection using objective tests to correlate physical, chemical and optical properties allows us to move from destructive to non-invasive analytics, and to create simple, yet reliable and accurate, stack tools for preserving our collections.

Building a data infrastructure that allows the capture of active datasets follows FAIR data principles and enables the reuse of data both within this research project, and for integration into future heritage collection assessments. This focus and approach will greatly benefit and assist the preservation of heritage collections throughout the nation, and internationally.

References

[1] https://nationalbookcollection.org/
[2] https://www.ucl.ac.uk/bartlett/heritage/research/projects/-projectarchive/identical-books
[3] Lyman Horace weeks, "A history of paper-manufacturing in the United States, 1690-1916", The Lockwood Trade Journal Company, Verlag (1916).
[4] Zachary Maiorana, Ian Bogus, Mary Miller, Jacob Nadal, Katie Risseeuw, and Jennifer Hain Teper, "Everything Not Saved Will Be Lost: Preservation in the Age of Shared Print and Withdrawal Projects", White Paper, (2019).
[5] https://blc.org/east-project
[6] Elliott Shore, "Talkin' Socialism: J.A. Wayland and the Role of the Press in American Radicalism, 1890-1912", University Press of Kansas, (1988).
[7] Internal Preservation Research and Testing Division Report, Aug (2016).

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.