https://doi.org/10.2352/issn.2168-3204.2021.1.0.8

Addressing the Challenges of Interoperability and Cultural Heritage Data

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

One of the ongoing challenges for effective utilization of heritage science data is the lack of access to well-organized and accessible extant data sets and the need to structure data in formats that allow interrogation and integration of related data. This need for data fusion expands to both subjective and objective measurements and descriptors, as well as a long-overdue need for established guidelines for metadata and shared terminologies, or more critically, ontologies. Research into this area has shown the need for Knowledge Organization Systems (KOS) that bridge and integrate multiple ontologies that address specific needs – for example the Getty Vocabularies for cultural heritage terms, the Linked Art model for a simplified core CIDOC-CRM, as well as the OBO Foundry and other scientific ontologies for measurements and heritage science terminology.[1]

Background

The intent of the Andrew W. Mellon funded research project "Assessing the Physical Condition of the National Collection" (ANC), [2] is to compare the physical, chemical and optical characteristics of 500 "identical" books from five large research libraries in distinct regions of the United States. The data will be used by collection care, preservation specialists and librarians 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 informing institutions about the potential risk of loss through the time period 1840-1940 – when mass production of paper began using acidic wood pulp. Interrogating the data will allow us all to fill gaps in our knowledge and guide the community by answering questions on how this time period of paper-based materials naturally age, 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 need for active interrogation of integrated data sets for this research necessitated an online platform that could be shared with partners, and that was robust enough to incorporate and integrate diverse scientific instrumental techniques such as size exclusion chromatography (SEC), tensile testing, pH acidity testing, Fourier Transform Infrared Spectroscopy (FTIR), Fiber Optic Reflectance Spectroscopy (FORS), various spot tests and ad hoc measurements (such as X-Ray Fluorescence, XRF), more traditional cataloging data and cross references, and the results of a visual assessment process that would need to adapt and grow as we learnt more from the project. To do this effectively we wanted to ensure we followed FAIR (findable, accessible, interoperable, reusable) scientific data principles [3], as well as LOUD (linked, open usable data) [4]. These requirements are critical for ensuring reliably accessible, reusable, and shareable data points and process descriptions.

Integrating and Interrogating the Data

The challenge was accessing and starting to review the ever-increasing data set to look for trends and markers, and assess the condition of the paper-based collection. This needed to start from the "visual" and somewhat more subjective assessment data, and our problem was finding a means to establish "semi-quantitative" condition assessment tools for librarians and other cultural heritage professionals working with at risk paper-based materials. Getting to this point required actively interrogating trends, and potential correlations, between objective and subjective data. Not only were we undertaking objective scientific analyses, we also developed a "visual assessment" with standardized descriptors and terminology, and were capturing data from both sides for a combined analytics approach. We quickly discovered that there were no agreed standards between heritage institutions for describing condition, one of the original drivers for this project.

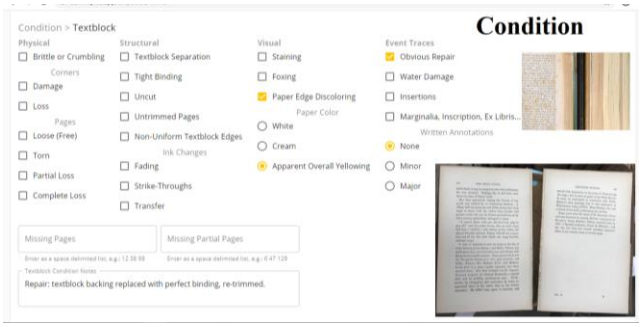

Figure 1. *"Visual Terminology" Condition Descriptors*

The language barrier has been a focus of some recent discussions. We realize that, while, as heritage scientists, we are comfortable using terms such as degradation and deterioration, these can be seen as value-infused terms that may be viewed as implying judgement, as opposed to a useful categorization. We have started looking at different ways to categorize degrees of deterioration, using terms that reflect that a book's condition is

"typical" for its age, while other volumes that may be more at risk than that norm are "a-typical".

This discussion links back to the Collections Demography project [5] (CollDem) from University College London (UCL) and partners, where the research project examined the challenges of preserving large heritage collections. The approach with CollDem was to view large collections and their management in the context of four aspects: collection use, material properties, environmental considerations and resource requirements. Our current research project focuses on the first three aspects, but we realize that every institution must be given the tools as well as the control of their collections and hence the decision making. The terminology in Collections Demography referred to collections as a population, and the accession/deaccession relating to "fitness for purpose". We consider that this new direction we are taking with terminology empowers heritage institutions to utilize and approach concepts of "levels of risk" in a more thoughtful and nuanced way, allowing also for the age of the collection item, and its inherent preponderance of age to be included in decision making. In the context of these four aspects, informed decisions are possible if evidence is available.

Heritage Science Data Methodology

As we collected and began to integrate the scientific analyses, we realized at a number of points that we needed to expand and revise our approach. For example, to order to interrogate the spectroscopic data in multiple ways, we needed to include both the raw data, as well as derived data renderings, and then include *query* tools that would start to capture and assess potential trends across the entire data set and within a variety of subsets. We also realized that in order to start to manage and engage with this ever-expanding data, and its complexity, we needed to be able to visualize the data in quick but thoughtful ways, that would allow for trends and connections between different scientific analyses to be immediately viewed and followed up on or discarded, and additionally; the ability to segregate data into subsets to look across/between/among institutions; and the usage (circulation) inherent material properties, and environmental aspects of the data.

We considered and rejected using artificial intelligence and machine learning models (AI, ML) as we were focused on extracting trends and exploring connections between the physical, chemical and optical properties in a series of research projects, and then evaluating how these might be predicted and linked to more quantitative subjective assessment criteria. Models may be derived from what we find, but we believed that, for instance, ML training sets would assume too much from too little at the outset. Our research would help identify what such a model might look like in the first place.

The Query tool allowed us to generate both 2D and 3D plots on an as needed basis from arbitrarily-selected measurement types. In Figure 2 below, we have mapped from CIELAB (L*, a*, b*) color space [6]; b* (yellowness of the paper) with the molecular weight of the cellulose, where lower molecular weight indicated a greater degree of deterioration and a less optimal condition for the textblock. The 3D plot adds the date of publication assigned to the z axis. The highlighted data point in the second screen capture shows an 1844 book with a b* of just over 9 and an Mw of 220 kDa. The outliers are editions and facsimiles from the 20th and 21st centuries.

The project's scope is for the condition of books published between 1840 and 1940, since mass production of books began over this time period, including methods that resulted in books and paper we now recognize as acidic. Since the project is still in progress, it will be interesting to see how the space in the large gap between those newer books and the older ones fills out as more books are analyzed.

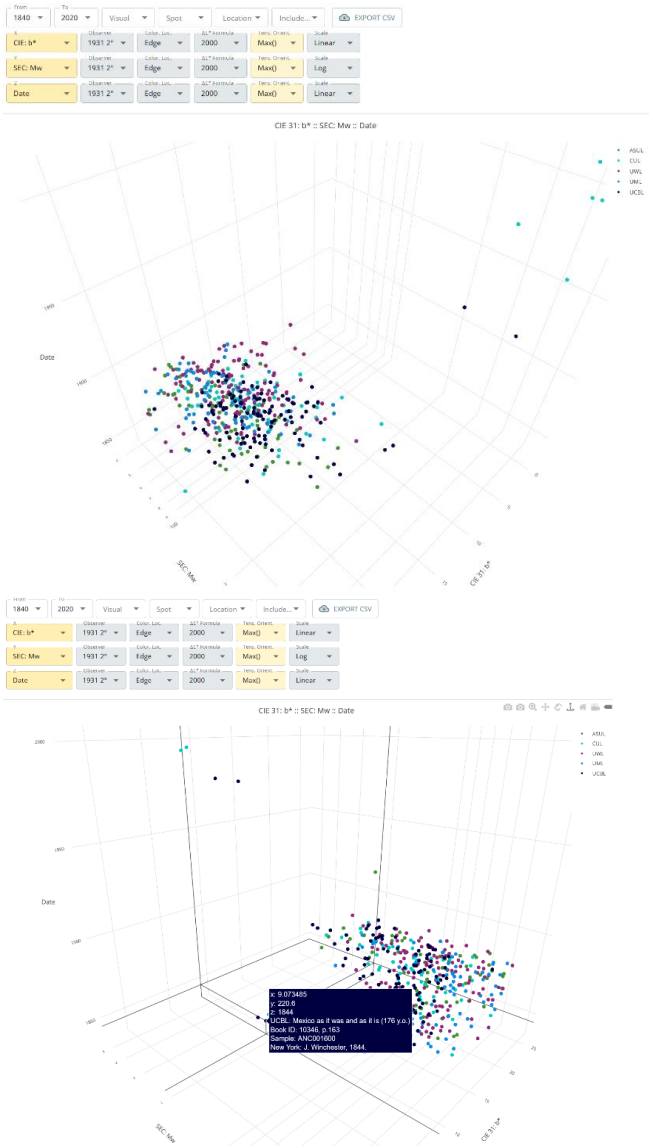

Figure 2. *Two views of the b* (yellowness) and Mw (paper molecular weight) with date of publication assigned to the z axis. The highlighted data point in the second screen capture shows an 1844 book with a b* of just over 9 and an Mw of 220 kDa.*

Considerations for Access and Reusing Data in other Formats and Contexts

Our approach to engaging with data sets required stepping back from accepted (and more static) methods of accessing data and creating an infrastructure that allowed expansion, re-interrogation, and active data sets that could also be combined to look for esoteric connections and markers of condition. There have been many moving parts that were challenging, including evaluating existing open-source software that already does some aspects of what we want. After evaluating a significant number of commercial as well as open-source software, we needed to find ways to interconnect the various software and the unique data characteristics and features to work together.

For data collection, we needed a more flexible solution than a traditional relational database (RDBMS) could offer, or would be efficient to work with given the frequently changing nature of the data types stored, and the kinds of relationships we wanted to explore between them. 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 CouchDB's "stored views" and REST API, which in turn was accessed via a React [8] application developed internally from a combination of our own and existing open-source packages. The tradeoff for this flexibility with documents and internal data relationships was needing to take on the responsibility for creating our own indices and relationship logic.

Results

Our decision to develop an expandable data storage system that connected the benefits of a select number of the most optimal tools and technologies we found, allowed us the flexibility we needed, while adding accountability: data processing decisions had to be made by us explicitly, and in a way that is clearly documented within our database's views and our front end code, rather than embedded in opaque third party solutions. In this way we also side-stepped the need to maintain a server-side application *and* a web-based client, which was crucial to deploying frequent updates in a timely fashion with limited resources, along with the occasional complete overhaul. The data was always cleaned and stored very close to its original 'raw' state, allowing us to simply add new series of transformations when we wanted to test new analytical techniques. Or, for instance, when we needed to add entirely new data types. These included the International Standards Research (ISR) papers, fully characterized reference paper samples for the 100-Year Paper Natural Aging Program [9], and other reference papers from Preservation Research and Testing Division's (PRTD) sample collection, all of which were subjected to the exact same micro procedures as the samples from the project's books.

Of particular benefit with this approach were the myriad ways we could refactor FORS data, both the spectra and the colorimetry data (see figure 3), to assess changes in textblock paper "color" and investigate whether this objective visual component could now be used and linked to other chemical and physical destructive test methods. Plotting CIE 1931 (x,y) and CIE 1976 (u′,v′) colorimetry data on appropriate chromaticity color-fields (see the two top plots in figure 3) involved a particularly enjoyable trek back through some mid-1990s NetPBM source code in C, and writing an equivalent in JavaScript that browsers could calculate and "paint" directly without relying on any third party libraries [10].

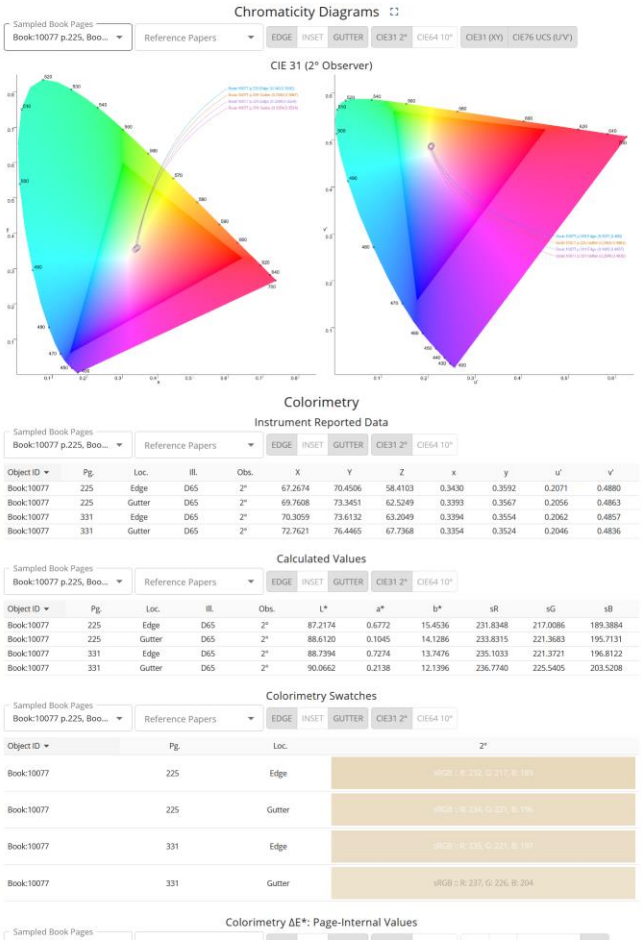

Figure 3. *Some examples of the colorimetry data we could parse and compare across measurements on the same page and between different books, including the "same book" from different institutions. Again, all data is transformed from near "raw" in real time for each application.*

The other side of this same coin, was that we smoothed the way for reliably sharing the data collected with others and for consumption by other types of application. See figure 4, for a fragment of a single analytical data point, in this case an FTIR measurement. Derived data, such as Kubelka-Munk function results, Savitzky-Golay smoothing and derivatives, and so on, are calculated in real time by the browser-based application itself, while the raw data itself is available in simple JSON format, potentially (assuming authentication requirements are met) accessible directly from the database itself by its ID. No drivers are required, everything happens via standard HTTPS requests. The views present a more complicated scenario than requesting a single data point, but the underlying principles are exactly the same for customized aggregates as they are for individual records.

```
{
  "_id": "analysis:0000009119",
  "_rev": "2-34e2cb10dd587afb65c8020d2bea0634",
  "sampleId": "sample:0000016788",
  "objectId": "book:10754",
  "barcode": "",
  "isInSitu": true,
  "type": "ftir",
  "procedure": "SOP",
  "analysedOn": "2020-11-27T19:33:42.000Z",
  "analysedByPersonId": "Forsberg_A",
  "lastModifiedOn": "2021-02-08T05:00:00.000Z",
  "lastModifiedByPersonId": "Forsberg_A",
  "notes": "",
  "data": [
   {
     "type": "raw",
     "src": {
       "header": "",
       "createdOn": "2020-11-27T19:33:42.000Z",
       "filename": "ANC_book10754_FTIR2.1.dpt"
     },
     "spectrum": {
       "values": [
         [
           "7996.2010724239",
           "1.1396551132"
         ],
         [
           "7994.1396645682",
           "1.1337426901"
         ],
```


Figure 4. *The first few lines for an example FTIR measurement as it appears in our CouchDB database.*

Figure 4, above, gives an indication of what we mean by 'raw' – for instance, the floating point measurements are stored as strings so as to require explicit casting to a precision decimal or floating point (as desired), and avoid intermediary platforms converting the value when received and passed on. Strings preserve exactly what the instruments' data files stored. JSON [11], the document format used, is not ideal for every scenario, but it has become something of a de facto lingua franca for data interchange. IIIF and LinkedArt [12] use the JSON Linked Data extension, JSON-LD, as their target format for related reasons – it is lightweight, and it presents barely any barrier for adoption across modern platforms and languages, and even not-so-modern ones in a pinch.

Just as the approach above helped us respond in a timely fashion to the use-case changes commonly found in active research projects, the same "store raw, transform on demand" approach has allowed us to evaluate, and re-evaluate, the seemingly ever-changing landscape of Linked Open Data. It would be premature, for instance, to represent our data internally with the Linked Art model, since that model is in active development by the community [13]. Instead, just as data is filtered and transformed as needed for Principal Component Analysis (PCA, [14]), we filter and transform the same to LOD models in JSON-LD. This has meant we can repurpose our data in diverse ways, including the dynamic generation of IIIF manifests and annotations for visualizations, and publishing data representations using bridged Linked Art and SemanticScience Integrated Ontology (SIO) models [15].

The dramatic increase in remote work this past year has helped strengthen our case for our Center for Heritage Analytical Reference Materials – Digital (CHARM-D, [16]) – a far more ambitious endeavor to develop a comparable framework for all of PRTD's scientific and cultural heritage data. That project's scale has required an entirely different set of technical tools, even though the principles, approach, and desired end result are the same.

Conclusions

As we have outlined above, the ongoing challenges for effective utilization of heritage data is the lack of access and creation of well-organized and accessible new and extant data sets, the need to structure these data in accessible, reusable and sustainable formats that allow interrogation and integration of related data, and the terminologies used to describe and engage with these datasets. To effectively bridge humanities and heritage science fields, the descriptors need to use plain and nonvalue-infused language for the interpretation, especially if the data will be developed into knowledge systems used for decision making at various levels in heritage institutions. Effective data fusion should include both subjective and objective measurements and descriptors, as well as a long-overdue need for established guidelines for ontologies.

Creating interoperable data infrastructures to reuse, access and integrate disparate heritage datasets allows for a more effective data analytical approach and the ability to extract more information for the preservation of cultural heritage collections. Further, creating tools that allow researchers to ask new questions of extant data, and integrate with new datasets, such as temporal and sensor data, including the effects of climate change, starts to open the possibilities for more effective collaboration and extraction of new knowledge from existing data. Coordinating a relatively simple accepted shared set of ontologies and terminology to describe data, instruments, techniques etc., and interoperable metadata will expand the connections between disciplines to link and reuse data.

References

[1] Getty Vocabularies: http://vocab.getty.edu/
Linked Art: https://linked.art/
CIDOC-CRM: http://www.cidoc-crm.org/
OBO Foundry: http://www.obofoundry.org/

[2] ANC public-facing site: https://nationalbookcollection.org

[3] M. D. Wilkinson, M. Dumontier et al., "The FAIR Guiding Principles for scientific data management and stewardship" Scientific Data volume 3, Article number: 160018 (2016), https://www.nature.com/articles/sdata201618

[4] FAIR: https://www.go-fair.org/fair-principles/
LOUD: https://linked.art/loud/

[5] Collections Demography: https://www.ucl.ac.uk/bartlett/heritage/research/projects/project-archive/collections-demography-dynamic-evolution-populations-objects

[6] CIELAB: https://en.wikipedia.org/wiki/CIELAB_color_space

[7] CouchDB: https://couchdb.apache.org/
Views: https://docs.couchdb.org/en/latest/ddocs/views/
REST API: https://docs.couchdb.org/en/latest/api/

[8] React: https://reactjs.org/

[9] 100-Year Paper Natural Aging Program: https://www.loc.gov/preservation/scientists/projects/100-yr_nat_aging.html

[10] Specifically, NetPBM's ppmcie.c package by John Walker: http://netpbm.sourceforge.net/doc/ppmcie.html

[11] JSON: https://www.json.org/
JSON-LD: https://json-ld.org/
[12] IIIF: International Image Interoperability Framework, https://iiif.io/
LinkedArt: https://linked.art/
[13] See Linked Art's Github commit log:
https://github.com/linked-art/linked.art/commits/master
[14] https://en.wikipedia.org/wiki/Principal_component_analysis
[15] SIO: https://semanticscience.org/
[16] Formerly CLASS: https://loc.gov/preservation/scientists/
projects/class.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.

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.