.

Future directions.

María Gimena del Rio Riande.

Collections as Data: State of the Field and Future Directions

Position Statements

April 25-26, 2023

Collections as Data: 50 Things Revisited: a personal journey

Sally Chambers, KBR, Ghent University, DARIAH, Impact Centre for Competence in Digitization[1]

My first encounter with 'Collections as Data' must have been around 2017. I was lucky enough to be tasked with organising a series of workshops to foster international exchange of knowledge and experience between the DARIAH community and digital Arts and Humanities projects in the United States and Australia, as part of the EU-funded project DESIR. One of those workshops was to take place at the Library of Congress. When starting to chat through ideas for the workshop with Abigail Potter she suggested that it may be interesting to think about a 'Collections as Data' workshop as part of the series[2] that the Library of Congress was organising. It was then the idea of Collections as Data: Digital Scholarship in the Arts and Humanities: An International DARIAH Exchange, which took place in October 2018, was born.

My second encounter of the 'Collections as Data' kind was this time thanks to Steven Claeyssens. Following a quick chat at the DH Benelux conference in Amsterdam in 2018, he asked me if I was going to the 'Labs event' at the British Library in September 2018. Thanks to Steven forwarding me the invitation, I was able to join for one day and witness the birth of the International GLAM Labs Community.

Almost 5 years later these two "chance encounters" have resulted in: 1) collaboratively authoring Open a GLAM Lab at a writing sprint in Qatar; 2) being successfully awarded our own 'Collections as Data' project at KBR, the Royal Library of Belgium. DATA-KBR-BE which aims to facilitate data-level access to KBR's digitised and born-digital collections for digital humanities research, which is now entering its fourth and final year; 3) in December 2022, the IMPACT Centre of Competence on Digitisation celebrated its 10th Anniversary with a workshop and writing sprint by creating a - soon to be published - collaborative white paper on *Sharing and Sustaining Digitisation Knowledge: an IMPACT White Paper*[3]; 4) having my PhD topic *"From digital collections to 'collections as data': addressing the challenges of using digital cultural heritage collections in (digital) humanities research"* approved by the Ghent University Faculty of Arts and Philosophy in January 2023 and 5) this year's DARIAH Annual Event 'Cultural Heritage Data as Humanities Research Data?'[4], where Thomas Padilla will be our keynote speaker, will explore the three themes of: 1) sustainable workflows for data management and curation; 2) imagining experimental data spaces for cultural heritage and 3) advancing digital methods for the analysis of cultural heritage.

---

[1] I am currently affiliated with: KBR, Royal Library of Belgium; Ghent Centre for Digital Humanities, Ghent University; Digital Research Infrastructure for the Arts and Humanities (DARIAH) and the Impact Centre of Competence in Digitisation.

[2] Collections as Data: Stewardship and Use Models to Enhance Access in September 2016, followed by Collections as Data: Impact in July 2017.

[3] Chambers, S., Candela, G., Vodopivec, I., Cuper, M., Parkoła, T., Antonacopoulos, A., Atanassova, R., Kaukonen, M., Gabriëls, N., Depuydt, K., de Does, J. Romein, C.A, et. al (2023). *Sharing and Sustaining Digitisation Knowledge: an IMPACT White Paper.* https://doi.org/10.5281/zenodo.7680489 (to be published shortly).

[4] See also: Tasovac, T., Chambers, S., & Tóth-Czifra, E. (2020). Cultural Heritage Data from a Humanities Research Perspective: A DARIAH Position Paper. https://hal.science/hal-02961317.

However, if there is one single 'collections as data thing' that has had the most impact for me, it would be the following: *"Within the next 1-6 months, what can you do within your own workplace to realistically and relatively easily improve/ move forward collections as data work?"*[5]. The 'Collections as Data' geeks among you, may recognise that this practical and down-to-earth question was posed at the starting point of 'Collections as Data: 50 things you can do'. No matter how large or small your cultural heritage institution is, *50 things* offers practical steps to start making 'Collections as Data' a reality. Even the first of the 50 things "*Know how optical character recognition (OCR) output is produced in your digitization workflows. What software is used? What formats are created? What levels of accuracy are produced? Where is it stored? Is it available for user download?*" was enough to get the ball rolling for me. If we revisit *50 Things* today they remain as important, challenging and practical as when they were published.

So thinking ahead from "50 Things" to a "Top 5 Things": where should we focus our efforts in the coming years?

1. **Digitisation Quality Assessment & Improvement**: a significant challenge for many cultural heritage institutions is how to assess the quality of their digitised collections and, when needed, improve it. For this, we need to develop a **digitisation quality assessment framework** that can be used in multi-institutional contexts, alongside solutions for quality improvement, post-correction and re-OCRing.

2. **Collections as Data Workflows**: building on experimental initiatives, such as Collections as Data and Cultural Heritage Data Labs, we need to continue to explore how to sustainably embed such activities into the day-to-day activities of cultural heritage institutions. The Collections as Data Checklist being developed by the International GLAM Labs community may be able to help us here. DARIAH will build further on this work in the context of the common European Data Space for Cultural heritage.

3. **Corpus building and dataset creation:** full-text access to cultural heritage data is an important first step. However, what is currently lacking are tools to help us create datasets, sustainably publish and cite them. For this, better understanding of what a corpus (or dataset) is, is essential (McGillivray et al., 2020). Corpus building or dataset creation needs to be supported by robust workflows as is the case when research using digitised historical newspapers (Oberbichler, S. et al., 2021). Perhaps the Datasheets for Datasets may also offer some solutions in this respect?

4. **Sharing and sustaining digitisation knowledge**: with more than 25 years of digitisation experience behind us, it is time to take stock of the current state-of-the-art of the digitisation of cultural heritage, understand what the key challenges are and how they can be addressed? Currently knowledge about cultural heritage digitisation remains fragmented. A key challenge in many

[5] https://collectionsasdata.github.io/fiftythings/

cultural heritage institutions is how to integrate state-of-the-art digitisation research into day-to-day digitisation practices and operations.

5. **Experimenting with data spaces and cultural heritage clouds**: the emergence of initiatives such as the common European Data Space for Cultural Heritage and Cultural Heritage Cloud are set to innovate access to and sharing of cultural heritage data. How can we build the necessary capacities in cultural heritage institutions to be able to proactively and inclusively contribute to such initiatives?

It's Like 10,000 Editions When All You Need Is a Book: Why Clustering Books, or "Works," Is Key to Unlocking Collections as Data

Melanie Walsh, University of Washington

"It's like ten thousand spoons when all you need is a knife"
-Alanis Morissette

Allow me to admit something a little blasphemous. As a literary scholar and data scientist, I care a lot about books, and I care a lot about book data. But I don't typically care about all the different editions and variations of a book. Typically, I care about "the work," in the Functional Requirements for Bibliographic Records (FRBR) sense—"a distinct artistic or intellectual creation by a person, group, or corporate body, which is identified by a name or title."[1] Although this interest might make me an outlier within the LIS community, I'm not alone. Many academic researchers, not to mention readers and industry professionals, share this cruder, more abstract interest in the work—in studying, searching, or recommending Jane Austen's *Pride and Prejudice*, not a specific edition or variation of Jane Austen's *Pride and Prejudice*. If we want to make collection data more useful for applied computational research and other applied use cases, then I argue that we need to pay more attention to "works" in this sense. We need to create more tools and resources for computationally clustering bibliographic records into works, and we need to prepare collections in ways that anticipate their subsequent use and analysis at the work level.

In this statement, I will share a concrete example that illustrates why scholars and other users might be interested in computationally engaging with collections at the work level, and why the current configuration of most collections makes pursuing this interest challenging if not totally prohibitive. Then I will briefly discuss some existing tools and methods for clustering records into works and suggest how these tools might be more fully integrated into the research and collections communities.

Why Researchers Care About Collections at the "Work" Level:
A Case Study with Library Circulation Data

What books did people turn to when the COVID-19 pandemic struck? Were people reading books about pandemics, like Emily St. John Mandel's fictional pandemic novel *Station Eleven* or Giovanni Boccaccio's 14th century plague tales *The Decameron*?

These are questions that I set out to answer in fall 2022 with the Seattle Public Library's publicly available circulation data (in my humble opinion, some of the most exciting library collection data that is available today). Each month, the SPL publishes aggregated data about how many times each book (or, more specifically, each item) in their collection is checked out—including how many times the book is checked out as a print book, e-book, audiobook, etc. This data is enormously promising for applied research, especially for research on readership, reception, and patron borrowing patterns. However, the SPL's circulation trends are not recorded in a way that makes it particularly easy to answer these kinds

of questions, because there is no easy way to cluster or merge disparate editions and variations of a book into a single "work."

Both titles and unique identifiers like ISBN numbers or OCLC numbers typically differ across different editions and variations of a book, so they cannot be used reliably to identify or analyze books at the work level. For example, in the SPL circulation data, the title of *Station Eleven* is variously recorded as "Station Eleven: A novel," "Station Eleven: A novel (unabridged)," "Station Eleven / Emily St. John Mandel.," and "Station Eleven [sound recording] / Emily St. John Mandel." If a researcher is interested in a single work like *Station Eleven*, it is feasible to cluster and merge titles in a targeted, semi-manually way, which is what I did for my analysis. But this is not feasible for large-scale analyses that consider more works. Last year, I had several conversations with David Christensen, the lead data analyst at the SPL, about the challenges that I was facing with discrepancies across titles, and he was able to add ISBN numbers to the SPL circulation records, which represents a useful step forward. But unfortunately, ISBN numbers are still unique for each edition and variation of a book, so this addition doesn't fully address the problem.

Importantly, computationally analyzing or engaging with books at the work level is an issue far beyond the SPL's circulation data. There are countless other research applications and use cases in which work level data is similarly essential. So how can we address this need?

How to Enable Researchers to Computationally Work with "Works"

In a dream world, I believe we would have authority records for works. Something like "VIAF for books" would solve most of the problems that I've outlined in this statement. Until then, to work with works computationally, we will likely need to cluster and connect identifiers, like ISBN or OCLC numbers, which correspond to the same work. There are some existing tools that help with this kind of clustering—as well as some unfortunately retired ones. In 2007, OCLC debuted the xISBN API, which allowed users to input an ISBN number and retrieve "related" ISBN numbers, but they announced the retirement of this tool in 2015. The social cataloguing website LibraryThing built their own tool, ThingISBN API, modeled after the xISBN API, which similarly accepts an ISBN number and returns related ISBN numbers. But rather than relying on OCLC information to determine related ISBNs, this tool relies on crowdsourced user determinations of which editions or variations are "related" to one another, making ThingISBN more reliable in some ways and less reliable in others. Finally, the Open Library also has an API for engaging with books at the work level. (For a detailed comparison of these tools, I recommend David Fiander's work [2].)

While I am only at the beginning stages of exploring these tools and the broader challenge of engaging with collections at the work level, I can say with confidence that we need better, more robust documentation, examples, and resources to support researchers in actually using these tools with collection data. While these tools are very promising, they cannot be fully embraced by researchers or other users without more support and guidance. But if that scaffolding is built, then I believe the potential of collections can be unlocked in broad and exciting new ways.

Works Cited

1. Bennett, Rick, Brian F. Lavoie, and Edward T. O'Neill. "The Concept of a Work in WorldCat: An Application of FRBR." *Library Collections, Acquisitions, & Technical Services* 27, no. 1 (March 1, 2003): 45–59. https://doi.org/10.1080/14649055.2003.10765895.

2. Fiander, David. "Comparing the LibraryThing, OCLC, and Open Library ISBN APIs." *The Code4Lib Journal*, no. 21 (July 15, 2013). https://journal.code4lib.org/articles/8715.

Against Archival Collections as Data

Michelle Caswell, PhD

Information Studies, UCLA
caswell@gseis.ucla.edu

When we transform collections of records into data, we lose context that is crucial for the ethical stewardship, interpretation, and reuse of archival records. The challenges that arise are questions of both scale and ethics.

I am a scholar of archival studies, a subfield of information studies interested in records as potential evidence of human activity that crosses space and time. A requirement for the ethical interpretation and preservation of this potential evidence is context; the context of who created a record, why, where, how and when. Without this context, we do not know what a record or collections of records could be *evidence of*. Content devoid of context is meaningless in archives.

The context of record creation is also known as provenance, the origin or source of a collection. It is the single most important principle in dominant Western archival theory. It is both how we physically organize archives (collections are grouped together by the entity that created or accumulated them, rather than by subject) and how we gain intellectual control over them (in knowing provenance, we can better interpret the meaning of collections).

Contextual information about provenance is conveyed through metadata and through descriptive tools like finding aids and bibliographic records. As archival theorists Verne Harris and Wendy Duff have asserted, how we convey this context is storytelling.[6] It is an art, not a science. It is always already, fundamentally, political. Archivists tell stories about the stories the records in their care tell. In so doing, we make descriptive choices. We leave out some details, highlight others. These omissions and embellishments, absences and stressors, compound and accrete to add new layers of silences and meanings to collections of records. The stories archivists tell are rich. They are not confined to controlled vocabularies and subject headings, though we may use those too. Our scope and content notes cannot fit into spreadsheets.

Our claims are specific. They are small and they are situated.

Datafication, turning these rich stories into data points, enables legibility across contexts. Yet when we transform collections into data, we flatten vibrant, messy stories. We dis-embed them from the context of their creation. We oversimplify them. We render them less-meaningful so that they can be legible and interoperable, so that they can talk to each other. The meaning we generate across

[6] Wendy Duff and Verne Harris, "Stories and Names: Archival Description as Narrating Records and Constructing Meaning." *Archival Science* 2 (2002): 275-276.

collections and across contexts may enable us to make larger-scale claims, but those claims are less reliable, less meaningful, less rooted in context.

We need more context, not less. We need process over product. We need slower archives, not faster.[7] We need complex and messy relationships (between ourselves, our records, their communities of origin) nourished over time, not seductively easy ways to make false but generalizable claims. And we need spaces for emotions—joy, anger, grief, relief—that cannot be robustly represented by data. Data points cannot fully convey the stories the records in our care can tell and the stories archivists can tell about them.

I'd take a story over a datapoint any day.

[7] Kimberly Christen and Jane Anderson, "Toward Slow Archives," *Archival Science* 19 (2019): 87-116.

Anticipating Use

Geoff Harder, University of Alberta

In 2011, my library was pleased to support a team of researchers from Virginia Tech and the University of Toronto - part of the *Digging into Data Challenge* (funded by NEH, NSF, SSHRC, Jisc). We were asked to supply copies of METS-ALTO tagged newspaper files from the *Peel's Prairie Provinces* collection, a trove of digitized textual and visual Western Canadiana. The research project was titled, "An Epidemiology Of Information: Data Mining The 1918 Influenza Pandemic." The researchers used data mining and interpretive analytics to "...understand how newspapers shaped public opinion and represented authoritative knowledge during this deadly pandemic" and "... to understand the spread of information and the flow of disease in other societies facing the threat of pandemics."[1] It was hard not to think of that project when society shuttered less than a decade later from COVID-19. Authoritative knowledge was again under debate, but this time magnified by social media and other channels of (mis)information. In many ways, the past was replaying itself - serving as a reminder to librarians and archivists that fact and fiction will be viewed through multiple lenses, including the computational, and the information and data generated and collected from this period will tell many different stories about our society.

I wish I could claim that our digitization and metadata teams saw a pandemic study coming - we didn't. However, we did anticipate that computational research techniques and methodologies were quickly evolving, and with some additional investment in descriptive metadata and markup of the newspaper collection, our time and effort might one day prove helpful to researchers. Making the data openly accessible also helped, even if our delivery mechanisms were primitive (sneakernet might have been the delivery mechanism).

So why the story? I confess to being nearer to the beginning than the end in my journey to understand collections as data, returning often to the basic question of what happens next. As libraries, we are tasked with making accessible and stewarding the entirety of our collections, and our collective collections, knowing use and purpose is difficult to predict. To understand pandemics, patterns of misinformation spread, or other challenges, we need to think differently about how to activate the whole of our collections as data for research and learning: this is a non-trivial change for many of our organizations, including mine, and one that implicates staffing, operations, and other aspects of our planning and priorities setting.

I welcome and approach the working meetings with questions such as the following:

1) The Santa Barbara Principles provide a helpful frame for viewing and planning collections as data, including the principle that "Collections as data designed for everyone serve no one." Know your user, know your communities. Intentionality behind decision and action is important when curating collections, particularly in light of limited supporting resources. There is often a risk that looking too much around the corner at what future researchers might want or do results in failures to respond and meet the needs of current-day users. As with my earlier story, we didn't

anticipate a study of the pandemic, but we did reasonably expect that investment in higher-level data might enable new kinds of research. Robert Grossman recently wrote, "It's not easy to predict what archived data will lead to great science."[2] How do institutional priorities get shaped with collection as data principles in mind? Where do we see progress being made? What can we learn from and build on? One thing seems clear: we need to listen to and respond to users, take some risks, and learn from our collective experience.

2) As an administrator, how best do we address the challenge of making collections as data affordably operational rather than… expensively experimental? We know that less is not more; less is less. Our organizations are continually needing to find efficiencies in every aspect of our planning. How does the library community more readily and efficiently ensure collections as data continues to progress in a way that demonstrates value and impact proportional to investment? *Collections as Data: Part to Whole* provides some inspiring examples. However, there is still a gap in many of our organizations when it comes to moving this work from bespoke projects and arrangements to operational practice. Are we ready for that step to happen?

3) Long-term stewardship of collections needs sustained effort. Collections as data might increase those stakes. What needs to change most in our organizations, and in our thinking, to best support collections as data in terms of lifecycle management? Trustworthiness, as outlined in the Santa Barbara principles, comes from doing, not saying. So what skills, attributes and characteristics do we need to reflect in our organizations, from mission to practice, to ensure the long-term utility and usability of collections as data? Persistence, authenticity, and a number of other considerations are being actively worked on and addressed but there is more to do.

4) Libraries and archives must consider and respect principles such as FAIR[3] and OCAP,[4] and resist the urge to let collections as data become part of an unchecked manufacturing process. Being thoughtful doesn't always go hand-in-hand with being efficient and fast (even ChatGPT agrees with me on this point). Library and archives communities, in partnership with other communities, need to be diligent in our effort, even if sometimes that means slowing things down to get things right.

I am very appreciative of the community-building being undertaken with the Collections as Data project. We owe many thanks to the team and event organizers for their efforts. Thank you for the opportunity to be part of the conversation.

1. Digging Into Data Challenge (2011).
2. https://diggingintodata.org/awards/2011/project/epidemiology-information-data-mining-1918-influenza-pandemic[2] Grossman, R.L. Ten lessons for data sharing with a data commons. Sci Data 10, 120 (2023). https://doi.org/10.1038/s41597-023-02029-x
3. FAIR Principles. https://www.go-fair.org/fair-principles/
4. The First Nations Principles of OCAP. https://fnigc.ca/ocap-training/

Archaeological collections and the ethical challenges in the use of digital repositories

Mercedes Okumura, Laboratory for Human Evolutionary Studies, University of Sao Paulo, Brazil

The first known archaeological collections were originated in association with the phenomenon of collecting, which began to gain importance during the Renaissance period. From the 15th century onwards, transoceanic expeditions and colonialist relations resulted in the material evidence of the existence of previously invisible human groups to Europe (Lopes 2009: 12). Such materialization would give rise to ethnographic and archaeological collections, later to be curated in museums. Such institutions, created mostly in Europe between the 17th and 19th centuries, forged a way of organizing, classifying, and subjugating the world. Nowadays, there is a movement towards shifting the protagonism of those in charge of collecting and organizing such materials in favor of a process of engaged decolonization of these collections, questioning the true ownership of artifacts, as well as the ways in which acquisition, care, and use (including digital repositories) can be considered minimally acceptable under current standards. Archaeological collections present specific challenges giving the wide range of materials (from stone tools to human remains) and the ethical concerns that the study of each category will raise.

Collections as Data: Part to Whole aimed to discuss how to implement collections as data and, at the same time, how to develop services that would further improve their scholarly use (Padilla et al., *s.d.*). We call attention to the need to involve more than academic scholars as users of such data and make it available for a broad public of stakeholders. In this light, archaeological collections can be considered as a special case, given that many of the artifacts can be related to historically marginalized groups, which are underrepresented in history, as the result of an absence of historical records and/or historical erasing.

The curation of archaeological collections, including their availability in digital repositories, implies the generation of interpretations about the artifacts and the human groups associated with them. Such interpretations present a great potential to influence the public's perception of past cultures and can rectify or reify concepts about these cultures. Even the selection of artifacts or themes to be part of the digital repositories, might reflect the institution’s agenda (Barker 2010), which is usually a hegemonic vision (Smith 1994). Many institutions around the world have been engaging with the local communities, establishing a dialogue that aims to stimulate more ethical curatorial practices. For example, the Museum of Anthropology at the University of British Columbia, started a pioneering dialogue with the Musqueam First Nation, who originally lived in the land in which the museum is located. Such partnership resulted in the curatorial collaboration of the permanent exhibition, as well as other important actions (Muntean et al. 2015).

Community engagement can only happen if the stakeholders are aware of the very existence of the collections, and if they are able to relate and engage to such materials, bringing underrepresented narratives to light. Digital access can be a very powerful tool in this process. We completely agree with the statements presented in the *Collections as Data: Part to Whole* grant call (p. 9): collections as data

should “seek to illuminate absences and/or underdeveloped parts of the historical record”, as well as “represent a diversity of content types, languages, and descriptive practices”. However, digital repositories of some archaeological materials, including (but not limited to) ancient human remains, pose specific ethical concerns and they need to be carefully designed, in order to avoid the misuse of such collections.

All three statements prescribed by the *Santa Barbara Statement on Collections as Data*[8] are of utmost importance when dealing with archaeological collections and they imply in problems that have been pointed out by several of the position statements presented by Bailey et al. (2019), like the lack of clear standards for publishing such data, the data formats (which can create the digital exclusion of some stakeholders), among others. Such technical complications, as well as the ethical issues raised in this position statement, represent the main challenges to be faced when one works with digital repositories of archaeological collections and the data that can be generated from them.

Bailey J et al. (2019). Position Statements. Always Already Computational: Collections as Data. Zenodo. https://doi.org/10.5281/zenodo.3066161

Barker A. Curating archeological artifacts. In: *Encyclopedia of Library and Information Sciences*. Third edition, pp. 1371-1379. 2010.

Lopes M. M. *O Brasil descobre a pesquisa científica*: os museus e as ciências naturais no século XIX. 2ª. edição, Brasília: Editora Hucitec, UnB, 2009.

Muntean R, Hennessy K, Antle A, Rowley S, Wilson J, Matkin B. ʔeləẇ̓k̓ʷ–Belongings: Tangible interactions with intangible heritage. *Journal of Science and Technology of the Arts*, 7(2):59-69, 2015.

Padilla T, Kettler H. S., Shorish Y, Varner S. *s.d*. “Collections as Data: Part to Whole". https://collectionsasdata.github.io/part2whole/

Smith L. Heritage management as postprocessual archaeology? *Antiquity*, 68: 300-309, 1994.

[8] “the work of the cohort would (1) support ethically grounded creation of collections as data that help collection creators avoid reinscribing bias in the cultural canon and harming marginalized communities, (2) provide access to collections as data in ways that align with the needs of multiple users, including users who have a range of technical expectations for types of access, and those who have ethical concerns about the scope of access, (3) develop organizational frameworks that support productive partnerships between libraries, departments, labs, research centers, university information technology, local cultural heritage organizations, university presses and more, and (4) demonstrate commitments to utilization of open source technologies that interoperate with a broader open scholarly communication infrastructure.”

Balance of Power: Finding Sustainable Models for Collections as Data Labor

Julia Corrin, Carnegie Mellon University

The “collections as data” umbrella is so large that it can be unwieldy to discuss. The realities of enabling solutions to support computational research vary widely between libraries, archives, museums, and large service providers like JSTOR. Because of this wide range of stakeholders, I want to clearly position myself –I am both a practicing archivist and a digital collections specialist at an institution with a long history of digitizing archival collections and a focus on applied learning. In this context, collections as data offer tantalizing opportunities: new user communities, expanded understanding of our collections, and institutional relevancy. But it also has major implications for why and how we do our work as archivists and curators, which is something with which we must engage more meaningfully.

Thus far, several early collections as data projects have placed the burden of labor on the institution. These projects bear the same hallmarks as many early digitization projects, which should be a cause for concern. The initial wave of digitization projects in the 1990s and 2000s focused on boutique solutions for single collections. Both users and curators have come to realize that this is not the most efficient or effective way of providing access. Most institutions now present multiple collections in a unified repository, allowing users to engage with a broader information landscape and simplifying curation and preservation for archivists.

So, as a practitioner, I am concerned that aspects of the Santa Barbara Statement–particularly the concept that “collections as data designed for everyone serve no one”–encourage this hyper-specific, boutique approach to collection work in ways that could have a negative impact both on the labor of archivists and on the quality of research done using these data sets.

First, the “build it and they will come” mentality risks developing resources that no one actually wants while diverting labor from projects with a demonstrated need. Many archivists are likely to remember conference presentations on projects and sites that, after an exhaustive review of the work involved, would admit, when pressed, to not having generated any users. While heavily mediated data sets derived from archival collections can serve as an example of the possibilities of data-driven archival research, they also divert labor from projects with known users. Archives have responsibilities to a wide variety of constituencies – academics, students, media, and genealogists, to name a few. Investing in the creation of a boutique data set inherently reduces support for these other user communities and reduces our capacity to address collection backlogs.

Second, archivist-created data sets foster a service provider/client mentality between archivists and researchers. In his position statement for the 2017 Collections as Data Conference, Tim Sherratt wisely noted that “if we think of machine-actionable data as a product or service delivered by institutions, then researchers are cast as clients or consumers.” This dichotomy between provider and client puts researchers and archivists in opposition, in which one always wants more than the other is able to provide. It also devalues the labor of the archivist and renders their work invisible. For the researcher,

the data set has sprung fully realized from nothingness. Providing ready-made data sets encourages scholars to set unrealistic expectations for the level of service libraries and archives should provide. It also renders the labor of archivists as "less than." The historian does not have the time or resources to convert analog data into a digital format, so the library should bear that burden and provide a level of support we would not consider reasonable for traditional forms of research.

Third, we risk fostering bad research by separating researchers from the process of generating data. The creation of any data set or any collection involves choices: big and small, obvious and hidden. In an archival context, these choices begin long before the idea of using a collection computationally is even considered, starting with the original decision of whether to collect something at all. More decisions are made when converting that analog information into a digital format and again when providing access. The digital surrogate of a collection that reaches a user may appear to be "complete" but is, in reality, only a portion of a larger whole. When researchers are not required to engage in the process, they may fail to interrogate these complications. As Tim Sherratt states in his position paper, "search interfaces lie, and data sets have holes."

In the hopes of making collections as data something that can be approached by a wide variety of archival institutions, I propose three areas of continued engagement:

1. **Work to identify the hallmarks of a data-friendly collection**: Are there types of information particularly well suited to digitization for this type of computational work? When assessing potential projects, are there elements archivists can include in their collection assessments?
2. **Position archivists and librarians as partners and participants in research**: Good collections as data work requires the expertise of researchers and practitioners. Let's work to raise archivists and librarians out of the acknowledgments section and credit them as partners and co-authors.
3. **Credit dataset development as scholarship**: Researchers are increasingly encouraged to deposit their data in institutional repositories, and many institutions are working to include those deposits as part of assessments related to tenure. Similar credit and encouragement should be given for scholars to resubmit their work - such as cleaned transcriptions - to the archives.

Works Cited

Sherratt, Tim. "The struggle for access." (2019). Position Statements --- Always Already Computational: Collections as Data. Zenodo. https://doi.org/10.5281/zenodo.3066161

Padilla, Thomas, Allen, Laurie, Frost, Hannah, Potvin, Sarah, Russey Roke, Elizabeth, & Varner, Stewart. (2019). Santa Barbara Statement on Collections as Data --- Always Already Computational: Collections as Data (Version 2). Zenodo. https://doi.org/10.5281/zenodo.3066209

Bridging academia's preservation and documentation of collections and communities at large

Gabriela Baeza Ventura, US Latino Digital Humanities Center, University of Houston

The divide between academia and the community has promoted and propagated data creation that is erroneous in its descriptions of cultural heritage, and the structures and processes for data creation, description and documentation impede communities of knowledge from describing their own history and culture in their own language. Fortunately, Collections as Data provides a valuable slew of tools (online training, speaker series, workshops, fellowships and principles) that spearhead the work of bridging the deep gap between academia and cultural heritage organizations. More work needs to be done with regards to language justice, precarity of Spanish-language materials, scalability, and community as an active participant in knowledge production.

The Recovering the US Hispanic Literary Heritage Program (Recovery) and the US Latino Digital Humanities Center at the University of Houston have been locating, preserving, and disseminating the written legacy of Latinos from the colonial period until 1980. The program and center have been working closely with community members at all levels (local, nationally and internationally, students–K-12, undergraduate and graduate) to create partnerships and provide training and mentorship. This work has made important strides in bridging the gap between academia and community.

In the process of creating the "Periodicals in the US-Mexico Border Region" project, in which Recovery will make available 200 newspapers through a portal, the preparation of the files and documentation included description in Spanish so the data is accessible to communities and archives in Mexico and Spanish-speakers in general (95% of the newspapers are in Spanish). The OCR process became a challenge for two reasons: 1) language and 2) condition of the newspapers. The latter clearly speaks about the fragility of the materials, but also documents the lack of care and resources allocated for the preservation of this history in the United States. Some of the struggles in the creation of surrogates and accessibility is the OCR process, considering that much of Recovery's archives are in Spanish and that the program's collection of newspapers (1,500) includes many that are in poor condition. As a result, we are in need of processes and tools that can preserve these precarious materials.

When Collections as Data processes are recommended for cultural heritage institutions, one must take into consideration the specific needs of those entities, keeping in mind the range of resources at their disposal (labor, budget, time constraints, technology, etc.). The processes should be scalable to the institution's resources; and members of these institutions–both staff and representatives of the communities they serve–should be active participants in the decision-making process. In this way, the dissemination of collections can become more widespread and include those who are locked out of the knowledge preserved in academia or whose knowledge rests outside of academia. An example of this is the simple spreadsheet created by USLDH to record DH projects developed by Latina/o scholars that became a tool for a community historian to research and record LULAC history. This community member transformed the spreadsheet into something useful for the community *and* academia. Another example

of bridging the gap between academia and the community is Recovery/USLDH internships funded by local organizations in Houston. These internships are aimed at students interested in acquiring skills in research, curatorial methods, digital curation, data creation, digital tools and project management. As the interns work on culturally relevant collections, they become empowered as knowledge producers, as well as community partners and liaisons. Community organizations, such as the local League of United Latin American Citizens (LULAC), are stakeholders in supporting the education of their future professionals and the preservation of marginalized histories. Moreover, these opportunities have immediate, visible impact on the field at large and the specificity of their local community.

As communities become more invested in research and preservation of their history, and in the mentoring of future scholars in the humanities, the question that lingers and preoccupies us is how can we create partnerships that ensure respectful processes and best practices that are truly inclusive of the creators of knowledge? How can we best acknowledge the role communities have in knowledge production, nation building and community activism? Community members, as protectors of this knowledge, have cultivated and documented data to fight against erasure of their communities. We recognize the responsibility of Recovery and USLDH scholars as intergenerational witnesses of Latina/o history.

Collections as Data: Building Research Capacity

Smiljana Antonijevic, Illinois Institute of Technology; University of Belgrade

The "Collections as Data: Part to Whole" initiative provided valuable results by correctly identifying the problem of "significant shortcomings within existing systems for supporting scholarly use of these [new kinds of] collections" (p. 3), and by addressing that problem through a set of projects. [1] The projects brought together librarians, faculty, students, and cultural heritage institutions to develop and implement collections as data along with tools, tutorials, models, and other services supporting scholarly use of those collections.

To continue building on these achievements, this work might be extended to support a strategy for three interconnected types of digital research capacity building in the humanities and social sciences (HSS)— 1) knowledge, 2) value, and 3) technical capacity. Doing so would address some of the systemic barriers to the best uses of collections as data. This strategy emerged in my ethnographic study (also supported by the Mellon Foundation) of 258 participants from 23 educational, research and funding institutions in the US and EU. [2] The strategy is now being put into practice through the research project "Distributed Archiving at the Institute for Philosophy and Social Theory (IFDT)" at the University of Belgrade, funded by the Filecoin Foundation for the Decentralized Web (FFDW). [3]

Knowledge capacity. A precondition for meaningful analytical use of digital research tools, methods, and collections is the ability to understand and employ those resources in a systematic, learned manner. Yet, both my ethnographic study and the reports from "Collections as Data" projects reveal a critical lack of knowledge capacity among HSS faculty and students. HSS scholars are often aware of methodological and epistemological benefits of digital research resources, but struggle to acquire skills that would enable them to reach beyond search and access level. An important impediment is the structure of academic rewards, which especially affects younger scholars who are advised to do "tenurable" work first, and then, if the time allows, to learn and do "digital stuff." Another important impediment is the currently prevailing organizational approach that places digital scholarship in libraries and other spaces outside central disciplinary workflows. HSS scholars point out that this approach focuses on technical skills rather than on field-specific research questions and needs, and stress that working with disciplinary colleagues inspires them to discover new tools and methods, and to apply them in their own work.

Value capacity. Disciplinary values and assumptions profoundly influence scholars' response to transformations, including the adoption of digital data, tools, and methods. HSS disciplinary perspectives and values— such as interpretability, reflexivity, contextualization, subjectivity, the (im)possibility of endpoint of inquiry— thus need to be carefully considered. Rather than digitizing materials and/or making them amenable for computational analysis with the hope that "data will drive" the research, developing collections as data should be driven with the epistemological and methodological questions of how a digital resource would be used in scholarly practice, what kind of research questions would it generate and help answer. A digital resource should be shaped by research practices, reflecting scholars'

ways of inquiry and working that closely link data with interpretation, with HSS values built and reflected in a digital collection.

Technical capacity. Studies and practice overwhelmingly show that digital research resources beneficial for HSS research should be interoperable, open source, sustainable, interconnected, easy to use, and simple to learn. Less often stressed, but equally important, is the need to be adjusted to HSS disciplinary needs. For example, one of the main challenges is to balance the need for precision, which enables scholars to compose queries that deliver precise meaning, and the need for ambiguity, uncertainty, and serendipity in HSS inquiries. To build tools adjusted to disciplinary needs, HSS scholars need to be actively involved in the development of those tools, which is still however rarely if ever the case. Digital scholarship is not merely supported by technology, but constituted through it, as digital research tools and resources shape both the object of inquiry and the way(s) of knowing it. Consequently, user involvement is not just a design or a policy issue, but a profound question concerning the kinds of research practices and ways of knowing that digital tools enable and promote.

This strategy of digital research capacity building is being put into practice, finetuned, and additionally developed through the research project on distributed archiving at IFDT, which collects, digitizes, preserves, interprets, and distributes information related to cultural heritage that various political and cultural authorities in Southeast Europe (SEE) have treated as "undesirable" or "unworthy" of support. Taking inspiration from the National Library of Serbia at Kosančićev venac in Belgrade, bombed and destroyed at Adolph Hitler's personal order in 1941, the project builds both the technical infrastructure and a community promoting libraries and archives that cannot be burned or destroyed at the order of a government or a single individual. This work is done through three work packages: 1) The Archive of Engaged Thought develops a collection of work by scholars and politicians who made a lasting impact on the history of SEE; 2) the Holocaust WP develops a collection of survivors' statements and archival data related to the German concentration camps in Serbia; 3) The Kosančićev venac Decentralized Library of Suppressed Heritage develops a collection of archival materials referring to the origins, destruction, and post WW2 neglect of the library at Kosančićev venac, as well as an XR mobile application that enables users to re-experience the destroyed library, access materials digitized by IFDT, and contribute their own materials.

The project builds **knowledge capacity** by supporting HSS students and researchers to develop digital expertise and skills by acquiring and implementing them during the project work, which is also academically recognized doctoral or post-doctoral work for a number of them. They gradually build knowledge capacity by completing a series of tasks essential for developing a digital collection as data (develop custom metadata, map out data organization, plan for cold vs hot data storage, etc.). In that way, we work on promoting and facilitating HSS scholars' sensibility to new ways of working and approaching objects of inquiry, while providing a safe space where they can freely ask, experiment, make mistakes, get frustrated, get excited.

The project builds **value capacity** by explicitly engaging first with thinking through how the

various sources to be digitized would be used in research, which kinds of questions would be generated. We want to do more than simply digitize, so we identify methodological and epistemological points of convergence and divergence among HSS scholars and disciplines, discussing epistemological, methodological, ethical and other challenges encountered in practice.

Finally, the project builds **technical capacity** with a user-centered approach implemented throughout the development cycle, ensuring that user input systematically informs the resource design process. We assist HSS scholars to articulate their discipline-specific methods and research objects in a way that best informs the design process through regular team meetings and discussions where both disciplinary and technological aspects of the project are discussed. We also work on documenting the design phases, methods, challenges, and decisions, with the goal to apply those insights to the development lifecycle and share with other institutions and teams engaged in related work.

Works Cited:

[1] Thomas Padilla, Hannah Scates Kettler, Stewart Varner, and Yasmeen Shorish. *Collections as Data: Part to Whole grant narrative*; accessed April 15, 2023 at https://collectionsasdata.github.io/part2whole/

[2] Smiljana Antonijevic (2015). *Amongst Digital Humanists: An Ethnographic Study of Digital Knowledge Production*. New York, NY: Palgrave Macmillan

[3] *Distributed Archiving at the Institute for Philosophy and Social Theory (IFDT)*; accessed April 15, 2023 at https://ifdt.bg.ac.rs/index.php/2023/03/03/distributed-archiving-at-the-institute-for-philosophy-and-social-theory-ifdt/?lang=en

Building the collections of tomorrow[1]

Beth Knazook, Digital Repository of Ireland

Mikala Narlock, Data Curation Network, University of Minnesota Libraries

Cultural heritage curation and data curation are information specializations that are, and should be, increasingly intersecting, especially with the rapid growth of digital cultural heritage resources and tools created by ambitious digitization programs and the rise of complex, computation-driven research in the digital humanities and adjacent fields. As the cultural heritage sector continues to build collections of invaluable and unique resources leveraged for a multitude of purposes – including academic research, citizen science, policy, industry, arts and education – it is becoming critical to consider how the curators (both of collections and data) act as "equal partners and important stakeholders in research activities."[2] As individuals with experience both in managing and preserving cultural heritage and advancing research data curation practice,[3] we are keenly aware of the similarities between the two specializations, and argue that efforts need to be made towards building a collaborative digital culture to build robust collections for the researchers of tomorrow.

Curators have immense power to shape collections – libraries, archives, and museums acquire, describe, interpret, digitize, preserve, and facilitate access to key government and business records, cultural heritage materials, and innumerable unique resources. In recent years our profession has been reckoning with the systems and biases that privilege certain types and sources of records, objects, and data, calling attention to the loud silences in the archival record that amplify the effects of historical inequalities.[4] The pressure to more equitably curate the digital record is amplified by the volume of born digital materials now pouring into libraries, archives and museums with a demand for collections processing that allows for machine computation with little human intervention. In her critical study of the American public library movement in the nineteenth century contrasted with the global digital library movement, Elisabeth Jones observed that "having access to the bits and bytes of a digital library does no good unless those data can be translated into comprehensible and usable knowledge. For digital libraries, then, an elision of technical access with access in general is untenable – and indeed, this perspective has deservedly fallen out of favor as digital library development has progressed."[5] In the same way, data curators and information professionals are all too aware of the complexity of providing machine-actionable but human-centered access. Curation practice is moving beyond the FAIR[6] principles into the CARE[7] principles, which recognizes and empowers the humans and communities often at the center of data collection. This shared area of investment by curators, both of cultural heritage and research data, is a space in which we can support and learn from one another.

Data curation, like museum curation, is a care practice. It requires empathy and understanding, compassion and an eye towards long-term access. Curation adds value to research outputs through description, structure, documentation, metadata linking and the enabling of automation, while the curation of traditional collections represents a selection, analysis, and interpretation of information sources for discovery and access – the importance of these activities as contributions to the academic

record is not well acknowledged and the ways in which their purposes overlap suggest that the work should be better aligned. Recognition of curation (in all its forms, including of physical cultural heritage materials) as a first class research effort is essential for achieving the vision of collections as data.

We see collaborations between collecting institutions and research data curators as critical to mediating and enabling computational access to cultural heritage as data. It is our job as information professionals and data stewards to encourage responsible engagement with our collections, regardless of their discipline or format, and to ensure we are providing access to these resources in a way that meets users where they are. Collections cannot become data without the end-to-end workflows that empower researchers and curators to acquire, interpret,share, and preserve the research data derived from collections alongside the collections themselves.

Across the datasets we've collected, curated, and created, what infrastructure, human or technical, is necessary to envision a more interconnected landscape? Can we envision a world in which our repositories and institutions, caring for both analogue and digital collections, can easily push both collections information and research data to high-performance computing centers to enable researchers to run scripts and analyses without needing a host of tools on their machines? Could the researchers, with the push of a button, submit an augmented dataset and accompanying publication back to the collecting institution/repository for review by an expert curator, bringing with it the essential contextual information about computing environment and provenance? To preserve the complexity of these resources, can we utilize established appraisal practices, which include reappraising collections, to consider how we might best curate and manage the growing volume of data as we continue to evolve? If our institutions are going to sustain ongoing expansion responsibly, we need to make it a priority to keep improving the overall quality of the collections data we preserve.

As a profession, we need to clearly define and recognize roles for curators involved in the shaping of the scholarly records. The items we accession into our data repositories, the collections we digitize and make computationally accessible, and the items we deaccession or leave in a physical format: all of these decisions impact future scholars and communities in ways that we need to honor. We shape research products. Scholars are making critical decisions about the data they share in repositories; curators need to be empowered to make the same decisions,and we need to communicate these decisions, these human interventions, in ways that acknowledge our expertise and biases.[8]

As a final reflection, we see the innovative spirit of our colleagues, our faculty and students, who are leveraging new models, techniques, and technologies in new manners at breakneck speed.[9] As our colleagues use data to create more data, accelerating the pace of change, we wonder how to balance our own desire for innovation and creating//curating new data. What are the resources that this work would consume?[10] How do we embrace collaboration and a slow mentality,[11] not to impede progress, but to ensure that we are promoting responsible, equitable, and sustainable engagement with the scholarly record?

[1] Thanks to Lisa Johnston and Matt Chandler for their comments on previous versions of this statement. [2] Georgia Angelaki et al., *How to Facilitate the Cooperation between Humanities Researchers and Cultural Heritage Institutions*. (Warsaw: Digital Humanities Centre at the Institute of Literary Research of the Polish Academy of Sciences, 2019). 10.5281/zenodo.2587481.
[3] We represent two institutions. The Digital Repository of Ireland preserves digital objects from collections stewarded by cultural heritage institutions (representing the GLAM sector - Galleries, Libraries, Archives, Museums) as well as from academic disciplines that typically interact with this data, including art history, literary history and archaeology. The Data Curation Network (DCN), based at the University of Minnesota, leverages a shared data curation workflow to support data stewards at our institutions in curating academic research data originating from a wide variety of different disciplines and formats. In addition to this shared curation workflow, the DCN prioritizes community engagement to solve shared challenges and develop practical resources.
[4] Dorothy Berry, in Berry, Dorothy, Kirt von Daacke, and Shaundra Walker, 2022. “Examining the Roots of Universities in Slavery and Anti-Black Racism.” Webinar from SPARC’s Knowledge Equity Discussion Series, July 27. https://www.youtube.com/watch?v=9SilcFvxrMU.

[5] Jones, Elisabeth. "The public library movement, the digital library movement, and the large-scale digitization initiative: assumptions, intentions, and the role of the public." *Information & Culture*, vol. 52, no. 2, 2017, p. 229+. *Gale Academic Onefile*, https://link-gale-com.proxy1.lib.uwo.ca/apps/doc/A494741834/AONE?u=lond95336&sid=AONE&xid=105 301bb. Accessed 30 Aug. 2019.
[6] Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg et al. "The FAIR Guiding Principles for scientific data management and stewardship." Scientific data 3, no. 1 (2016): 1-9.
[7] Carroll, Stephanie Russo, Ibrahim Garba, Oscar L. Figueroa-Rodríguez, Jarita Holbrook, Raymond Lovett, Simeon Materechera, Mark Parsons et al. "The CARE principles for indigenous data governance." (2020).
[8] Caswell, Michelle. "Dusting for fingerprints: Introducing feminist standpoint appraisal." *Journal of Critical Library and Information Studies* 3, no. 2 (2021).
[9] Thanks to Matt Chandler for his thoughtful conversation around this topic.
[10] e.g.,Kinnaman, A., & Munshower, A. (2022). Green Goes with Anything: Decreasing Environmental Impact of Digital Libraries at Virginia Tech. http://hdl.handle.net/10919/112392.
[11] Petrini, Carlo. Slow food: The case for taste. Columbia University Press, 2003.

Collections as Data as Destabilization

Jefferson Bailey, Internet Archive

It would be hard to address the topic of Collections as Data without acknowledging how the landscape of computational use of collections has changed since the prior forum in 2017. We refer not to the ample, admirable progress made in libraries, archives, and museums (LAMs) as far as advancing projects reconfiguring their collections as data. Instead we refer to the emergence of generative artificial intelligence (AI) and its immediate seizure of the news cycle, public imagination, and professional discourse. It is impossible to talk about computational use without also addressing the many ramifications of generative AI. Public and pundit reactions to tools like ChatGPT, Stable Diffusion, and others have ranged from guarantees of the impending apocalypse to promises of the impending singularity, with assurances of the golden age or dark age for work, politics, cultural production, (mis)information, and, yes, even libraries. Cue the obligatory mention of Skynet. While more established text and data mining operations for scholarly knowledge production will continue – and continue to rely on stewards making collections available as data – any future conceptual discussion of collection as data will also need to consider the role of artificial intelligence.

Untrained Transformer

Behind recent popular AI tools are massive datasets culled from a variety of sources. Generative text tools incorporate large language models (LLMs) built on huge troves of documents; generative image tools are built on massive datasets comprising billions of images. In some cases these tools are forthcoming about the provenance of their datasets, for instance the Pile or LAION,[1] while others sidestep the question or only obliquely reference some of their sources. It is widely understood that both LLMs and image datasets include a huge portion of material scraped from the web, from both reputable and questionable sources of various open access and rights statuses, as well as from institutional repositories, image collections, code repositories, and document archives. There is no doubt that these training sets and resulting models include collections from cultural heritage organizations. LAM collections on the web, either in native or derived form, as data or metadata or code – all are part of the data powering AI tools. How to reckon with this type of unexpected use of collection as data for unknown purposes in service to the creation of tools whose benefits and harms are as of yet unclear? For some non-heritage content stewards or creators, the response was legal action.[2]In other cases, AI organizations have worked directly with heritage organizations to find transparent and beneficial ways to prepare their collections for use in AI models, such as the BigLAM project.[3] But AI has also provoked a fair amount of anxiety around how collections are being used, what types of tools are being built on top of them, and what products, interests, or agendas these tools are serving.

Generative Collection Intelligence

The LAM community is working to share institutional experience in deploying AI tools and collaborating with the broader AI community via efforts like AI4LAM. The professional literature is exploring the policy, procedural, and ethical implications of AI in LAMs. At Internet Archive, like at

peer digital libraries, we have used ML/AI tools to enhance appraisal, description, discovery, and accessibility, from building models to identify scholarly literature, to using tools for voice to text transcription, image tagging, and automated metadata creation.[4] Some of these projects are already part of production pipelines and some are more exploratory experiments. At the same time, as a large digital library, we have worked with AI organizations, both academic and industry, to better understand how AI models themselves can be archived for historical preservation, how collection-specific "small models" can enable new forms of discovery, and the ethical implications of providing access to large original or derivative datasets. What is clear is that in many cases, AI will destabilize how we think of collections as data, both internally and in support of external use. Providing data for a model to learn against is different from providing data in support of scholarship. A generative AI tool is a different outcome than an academic article. And a multimodal training dataset – especially one in which your institution's data may be a miniscule portion – is different from a subject or topical research corpus. Much as digitization or the web destabilized LAM practices in ways both positive and fraught, AI, even in its infancy, has provoked a similar complex set of uncertainties.

Always Already Destabilized

Prior intellectual and operational inventions have had destabilizing effects on LAM practices. Postcustodialism challenged existing archival theories on the centralization of the management of records collections. Poststructuralist philosophies situated the archive as a locus for the negotiation of political power. Archival concepts such as the *fonds* have been twisted, disputed, and reconfigured over time. Recently, many projects have sought to address the exclusivity and homogeneity that characterized historical library collections by diversifying the voices and stories represented and preserved in contemporary collections. The Collections as Data movement has itself worked to challenge antiquated methods of access and related institutional workflows. All these actions represent historical waves of destabilization of prior frameworks, theories, and practices. AI has introduced another wave that will call into question everything from open access to staffing to infrastructure to non-consumptive use. Large AI datasets also prompt serious questions about the use of personal data culled from the open web. Access to, and use of, collection of data, is suddenly a critical part of the overall social discourse, and the boundaries between legitimate and illegitimate uses will often be opaque and disputed, with little legal clarity or ethical simplicity. Our provocation is that LAMs not react with the alarm and hand-wringing that has characterized so much of the recent public conversation around AI. We encourage the community not to let the yet-to-be-known positive or negative outcomes of AI impact our commitment to making collections accessible in any and every way possible. AI has evoked ample anxiety around use and access and we as a community must be vigilant and thoughtful that we remain committed to open access to information and to collections, even in times when access to, and use of, these collections can seem problematic. Collections always have and always will be used for public good and, at times, for public harm. The act of making collections publicly accessible, in their original form, and as data, has always been an act of faith – faith that most users will be well intentioned and that most uses will lead to positive outcomes, to knowledge production, and to the public good. Even if the pathway to that public good is, at times, messy, it remains critical to our collective mission that we maintain our commitment to

access.

[1] https://pile.eleuther.ai/; https://laion.ai/blog/laion-5b/
[2]https://techcrunch.com/2023/01/27/the-current-legal-cases-against-generative-ai-are-just-the-beginning/ [3] https://huggingface.co/biglam
[4] https://github.com/internetarchive/pdf_trio; https://openai.com/research/whisper;

Collections as data for machine learning: if we build it, who will come

Clemens Neudecker, Berlin State Library

In recent years, Staatsbibliothek zu Berlin - Preußischer Kulturbesitz (Berlin State Library, SBB) has experimented with multiple new ways to offer its digitized collections as data suitable for computational use, for a wide variety of users from humanities and social sciences to computer science. A library lab was established which first offered public documentation on the available APIs together with examples for its use, but also bulk downloads of datasets curated for particular themes (e.g. "gender" or "precarity"), events (e.g. "World War I", "German Revolution") or by modality (e.g. only OCR texts, only images extracted from documents, only metadata) were created. It was determined that all data offered this way must include (a) an open license (CC0/Public Domain Mark or CC-BY), (b) supporting documentation, (c) a contact in the library able to answer questions about it. The lab is complemented by a Zenodo community for sharing of datasets from SBB with a DOI. Furthermore, special-purpose datasets consisting of ground truth, i.e. manually transcribed and/or labeled data accompanied by scientific publications are published on specific platforms that were identified as being established and widely used in particular research communities (e.g. IAPR Technical Committee 11 for Computer Vision, European Language Resources Association Catalog for Natural Language Processing or HuggingFace for Machine Learning). All of the above activities were conducted mainly as a grassroots effort by volunteers from various departments in the library in addition to their regular duties, at times joined by staff from funded and temporary projects. With no dedicated resources available in the library to maintain these activities, there is considerable fluctuation with regard to how often the resources are updated, or new materials are being added.

Alas, only very little use of the above resources has been recorded over the last three years, especially in comparison to the time and effort devoted to it by the volunteers in the library. This has also made it more difficult to justify the allocation of permanent resources for such activities. While the number of enquiries for access to data has continued to grow over the same time period, the requests have since become much more specific and diverse at the same time. For example, where in the past users had enquiries about the availability of full text, now one user may be interested only in the stock market tables from a particular historical newspaper title while another needs only the printer marks from all German prints or only the pages from all digitized documents that contain depictions of castles with their geolocations. What satisfies one use case is insufficient for another. With new possibilities also comes more demand, and it becomes more challenging to prioritize available time and resources. The current way to deal with this is to collect use cases and evaluate them for common needs, which will then eventually be rolled out across all collections. But this removes a lot of the agility from the collections as data approaches.

At the same time, another major driver in the library for collections as data have been the advances in machine learning and artificial intelligence (AI). While AI has enormous potential to improve digitization (e.g. text recognition and information extraction from full texts) or traditional library tasks (e.g. subject

indexing), it also comes with some considerable risks attached. Most current AI tools and models are based on neural networks that are trained on large amounts of data to derive stochastic models. Therefore, the capabilities of these AI models are highly dependent on the type and quality of the data that they were trained on. But too often it is not clear what is the source of the data that AI models were trained with, and what may be undesirable implications and biases that the model draws from the data. Many questions come up when investigating how data for training AI models is assembled: What are the sources from where the data have been obtained? Have there been selection criteria, and if so, what were they? Are questions of data quality addressed? If the data contains problems, are they known and documented? Is there someone to reach out to in case of issues? A comprehensive and sobering survey of the practices of dataset development and use in machine learning gives many examples [1]. When libraries adopt and apply industry AI tools and models, particular care must be given to not introduce or amplify biases and issues in the collections in this way. Similarly, when collections as data increasingly become available also to train and fine-tune AI models, due to the predominantly historical nature of collections being digitized, it would be very desirable to establish a curatorial process that aims to identify, document and contextualize any problematic content (e.g. due to colonialism, racism, or underrepresentation of marginalized groups).

On the other hand, with their long tradition and great expertise in curation following established quality standards, libraries would be well equipped to adopt such data stewardship [2]. Curators are subject experts who are trained in various domains to provide contextualization, and have attention to detail. Also, libraries as public and non-profit organizations provide a higher level of transparency, trust, neutrality and reliability. Furthermore, libraries tend to show greater sensitivity and are more rigorous with regard to data sovereignty and the adherence to laws regulating privacy and personal data protection. Last but not least, research data repositories in libraries typically are established based on long-term preservation policies that also provide ways to track dataset changes and updates or even the depublication of datasets transparently and more reliably. But also the curatorial practices in libraries need to be continuously questioned and updated to build up awareness and reflect and resolve biases and issues.

This leaves us with the question whether such investment into good curation and quality data by public institutions should be free for private enterprises to exploit commercially via new AI systems? [3]

Works cited

[1] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, Alex Hanna (2021). Data and its (dis)contents: A survey of dataset development and use in machine learning research. In: Patterns 2 (11). https://doi.org/10.1016/j.patter.2021.100336.
[2] Clemens Neudecker (2022). Cultural Heritage as Data: Digital Curation and Artificial Intelligence in Libraries. In: Proceedings of the 3rd Conference on Digital Curation Technologies (Qurator2022). https://ceur-ws.org/Vol-3234/paper2.pdf.
[3] Jörg Lehmann (2023). On the Use of Licences in Times of Large Language Models. https://mmk.sbb.berlin/2023/03/24/on-the-use-of-licences-in-times-of-large-language-models/?la

ng=en.

Collection Datafication for Greater Access, Better Understanding, and More Inclusivity

J. Stephen Downie and Glen Layne-Worthey

HathiTrust Research Center, iSchool, University of Illinois Urbana-Champaign

Positive convergences of seemingly negative forces are a rare and beautiful thing: when apparently insurmountable restrictions of intellectual property lead directly to otherwise unimaginable advancements in data openness; when the brash "moon-shot" project of an uncontrollable industrial juggernaut inadvertently sows the seeds for a major grass-roots academic effort; when the overwhelming tidal wave of artificial intelligence unexpectedly reveals new purposes for human intelligence – a celebration is in order. More than clouds with silver linings, these are occasions that reveal the sort of hopeful connectedness that underlies our work in libraries, and our collective efforts to curate, preserve, share, and understand our cultural heritage.

The very existence of HathiTrust is owed to just such a set of convergences. This is true for all of its many guises – as a digital library consortium; as a massive collective digital collection; as a provider of emergency library access during the pandemic; and finally, in the guise we in the HathiTrust Research Center (HTRC) know best, as the tangible embodiment of the "Collections as Data" principles alongside which our Center came into being.

This brief statement will focus on three aspects of library collection datafication as we know it in HTRC. First, the origins, affordances, challenges, and opportunities of derived data. Second, the essential and growing importance of human intelligence, intervention, and care under circumstances of automation, mass digitization, and the dizzying pace of progress in artificial intelligence. And third, an unexpected return to page images as our next collections-as-data challenge.

1. Data into metadata: Spinning the straw of copyright-restricted collections into the gold of massively open data

One of the HathiTrust Research Center's primary activities is the creation and provisioning of derived data sets, principal among them the Extracted Features (EF) dataset.[1] We've also created more specialized sets of derived data that include geospatial, sociolinguistic, narrative, and other derived-data enhancements; and we're working on future forms of derived data such as word embeddings and language models. But EF is the parent of them all.

The HathiTrust collection currently includes more than 17.5M digitized volumes, about 60% of which are under copyright protection (or of uncertain copyright status). A book once digitize is data, of course – but even as such, this collection was largely unusable for data-intensive research because of its restrictions. In order to facilitate text-mining access to it, we chose to turn the data into metadata: word-level metadata of a linguistic, statistical, and page-positional nature which, because it consists exclusively of facts *about* the text rather than the text itself, is not subject to copyright. Relying on the legal understanding of *non consumptive* use (which was a byproduct of the decade-long Authors Guild v. Google (and briefer Authors Guild v. HathiTrust) lawsuit, we release this derivative data openly, for the

entire collection. In addition to being open access, it is also by its very nature, “always already computational.”[1]

But that does not mean it is easy to use: its bulk is cumbersome, and its JSON format is intended for algorithms more than humans. So to mediate among the rich data potential, its many potential uses, and its many hopeful users, we’ve undertaken a project called TORCHLITE (short for “Tools for Open Research and Computation with HathiTrust: Leveraging Intelligent Text Extraction"), which will provide open API access to the massive EF dataset at an arbitrarily granular level, allowing both ourselves and our user community to build lightweight tools to exploit that data.[2]

2. Keeping the humans in the digital humanities (and across the disciplines)

Although our datafied collections have machine algorithms as their focus, and although human readers – with the important exception of those taking advantage of the HathiTrust’s pandemic-era Emergency Text Access Service – we have learned (or rather, reconfirmed), that humans are absolutely essential as we increasingly employ both automation and increasingly powerful artificial intelligence tools. These human interventions are absolutely required for expert curation, domain knowledge, and social justice.

Although they are based on the solid foundation of decades and centuries of human-curated library collections, the mass digitization efforts that have created the HathiTrust corpus have been quite haphazard. In order even to understand these collections, to know their margins, to discover their lacunae – and all the more to improve them, make them more diverse, equitable, and inclusive – we have come to rely more than ever on human expertise. Our “Scholar-Curated Worksets for Analysis, Reuse & Dissemination” (SCWAReD) project, generously funded by the Mellon Foundation,[3] centers humans and social justice in two crucial ways: by relying scholar-curators who help us to identify and remediate gaps in the collection, and by focusing on historically marginalized and underserved textual communities.

3. Fast-forward to the past: The page as data

One of our next collections-as-data challenges is the algorithmic use of page images. Although this may be counterintuitive in the context of the sort of machine-actionable, text-centric, non-human-consumptive reading technologies that HTRC has pioneered, we are learning that page images include informational riches such as tables, charts, illustrations, captions, formulas, advertisements, typesetting, and many other features that are lost in the very first steps of datafication. We are beginning to focus more effort on recapturing those important information carriers, submitting them to the rigors and affordances of statistical and machine-learning methods, and making them more computable.

We are embarking on a new research agenda in exploiting page-image data in a number of different projects: one designed to algorithmically distinguish front matter pages (which, like Extracted Features, are largely not subject to copyright restrictions and can safely be opened for public viewing); another, to identify photographs from historical astronomy texts, and extract them together with their appropriate

captions and legends; and a third, to enable cultural historians better to identify and interpret trends in advertising and illustration in the popular press of the early post-colonial Middle East. The HathiTrust Research Center, together with its generous funders and its lively research community, is thrilled to be a partner in the Collections as Data effort.

[2] "HathiTrust Research Center receives NEH support for open research tools." https://ischool.illinois.edu/news-events/news/2022/01/hathitrust-research-center-receives-neh-support open-research-tools

[3] "HTRC receives $500,000 from Andrew W. Mellon Foundation," https://ischool.illinois.edu/news events/news/2020/10/htrc-receives-500000-andrew-w-mellon-foundation

COLLECTIONS AS TRAINING DATA?

Daniel van Strien, Hugging Face

daniel@huggingface.co

1 Introduction

A spectre is haunting GLAM collections — the spectre of Large Language Models. This spectre calls into question the role of GLAM collections and the values, processes and technologies underpinning these collections. The ever increasing demand for more data to train these models, alongside the desire to enable new forms of interaction with collections via machine learning, means that increasingly collections will be used as training data. There are implications for collections as training data. It changes; who, how and why collections are used. There are also unique challenges and opportunities in using collections as training data, meaning specific attention must be paid to this topic. Many institutions are already responding to collections as training data in various ways, but many opportunities for deeper collaboration and collective responses remain.

2 Collections as training data?

Below are some questions I believe are essential to tackle as collections are increasingly used as training data.

2.1 Collections as training data: how?

Broadly, we can characterise the process of training machine learning models into two broad regimes: supervised and unsupervised. Data for both of these training regimes exist in GLAM collections. Potentially different approaches are needed to facilitate (or discourage) the use of collections as training data for these two training approaches.

2.2 Collections as training data: who?

Mapping out the potential users of collections as training data will help institutions understand who might be a potential user of collections as training data and, importantly, which of those groups collections feel warrants the most investment in supporting.

2.3 Collections as training data: why?

As well as considering who might want to use collections as training data, it is also important to consider why collections might be of interest as training data.

2.3.1 The Collections themselves?

The appeal of a collection as training data may result from the collection itself. Collections may be interesting because they:

- contain material not readily found on the traditional datasets used for training machine

learning models

- collections may contain 'low resource' languages difficult to access from other sources.
- collections may also have already been transformed or presented in ways that make them readily useable for training various machine learning models, making them appealing because they are readily useable for training machine learning systems.

2.3.2 Scaling 'laws'

Without delving too deeply into technical details, 'scaling laws' refers to the observed phenomena of 'a optimal model size and number of tokens for training a transformer language model under a given compute budget.' [1] What this can mean in practice is that more data can be valuable. Some GLAM institutions hold significant volumes of data, making them an interesting 'resource' for accessing training data.

2.3.3 The Institutions holding the collection?

For some users of collections as training data, the institution holding the data might be incidental. However, for other potential users of GLAM collections as training data, the institution responsible for collections are a direct part of the appeal because, amongst other reasons, the work done by the institution to describe collections, consider legal and ethical questions, and the expertise institutions have around the limitations of the collections they hold.

3 Potential modes of engagement collections as training data face

There are a variety of 'modes of engagement' with collections as training data. Understanding these modes – and which modes we might want to encourage – will help shape how (or if) to make collections available as training data.

3.1 Do ask; don't tell

The users of the collections as training don't rely on legal access (requiring only some technical means of accessing collections) and because they don't report what is included in their training data is not easy to know if a collection was included in the training data.

3.2 Fair use

Another approach to gathering training data for machine learning models is to rely on fair use exemptions. Whether this approach will withstand legal scrutiny across different jurisdictions and settings remains an open question.

3.3 Data governance

Another mode of engagement is driven much more strongly by consent and collaboration. The

BigScience project[2] is an example of such an approach to gathering a dataset for training an open large multilingual language model. The BigScience project was an open science project involving more than 1000 researchers from 60 countries and over 250 institutions. The project developed approaches to data governance[3] for training data for large language models and created and documented the ROOTS corpus, a 1.6TB Composite Multilingual Dataset used to train the BLOOM language model.[4]

4 Practicalities

There are practical responses to collections as training data that are already being actively explored. These practical steps include:

- Documentation of collections that may be used as training data. This includes the use datasheets as a tool for documenting datasets,[5] and model cards for documenting machine learning models [6].
- Licensing: an open-eyed investigation if existing approaches to licensing materials remain appropriate as collections become training data.
- Access: innovating new approaches to accessing collections which better enable (or prevent) the use of collections as training data. For example, providing access to collections well suited for machine learning applications via the Hugging Face Hub.

How best to pursue these must be driven by what institutions wish to prioritise. There is also a lot of value in institutions continuing to openly share lessons learned in making collections available as training data.

References

[1] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models, 2022.

[2] Christopher Akiki, Giada Pistilli, Margot Mieskes, Matthias Gallé, Thomas Wolf, Suzana Ilic, and Yacine Jernite. ´ BigScience: A Case Study in the Social Construction of a Multilingual Large Language Model, December 2022.

[3] Yacine Jernite, Huu Nguyen, Stella Biderman, Anna Rogers, Maraim Masoud, Valentin Danchev, Samson Tan, Alexandra Sasha Luccioni, Nishant Subramani, Gérard Dupont, Jesse Dodge, Kyle Lo, Zeerak Talat, Isaac Johnson, Dragomir Radev, Somaieh Nikpoor, Jörg Frohberg, Aaron Gokaslan, Peter Henderson, Rishi Bommasani, and Margaret Mitchell. Data Governance in the Age of Large-Scale Data-Driven Language Technology. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2206–2222, June 2022.

[4] Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, Jörg Frohberg, Mario Šaško, Quentin Lhoest, Angelina McMillan-Major, Gerard Dupont, Stella Biderman, Anna Rogers, Loubna Ben allal, Francesco De Toni, Giada Pistilli, Olivier Nguyen, Somaieh Nikpoor, Maraim Masoud, Pierre Colombo, Javier de la Rosa, Paulo Villegas, Tristan Thrush, Shayne Longpre, Sebastian Nagel, Leon Weber, Manuel Muñoz, Jian Zhu, Daniel Van Strien, Zaid Alyafeai, Khalid Almubarak, Minh Chien Vu, Itziar Gonzalez-Dios, Aitor Soroa, Kyle Lo, Manan Dey, Pedro Ortiz Suarez, Aaron Gokaslan, Shamik Bose, David Adelani, Long Phan, Hieu Tran, Ian Yu, Suhas Pai, Jenny Chim, Violette Lepercq, Suzana Ilic, Margaret Mitchell, Sasha Alexandra Luccioni, and Yacine Jernite. The bigscience roots corpus: A 1.6tb composite multilingual dataset, 2023.

[5] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

[6] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229, 2019.

A community of practice – Looking for proactive users of cultural heritage data.

Alba Irollo - Europeana Foundation

Version 2.0

"Community of practice" is the definition that, nowadays, European policymakers tend to adopt when referring to users within large-scale initiatives based on data-sharing, in which data mainly comes from content providers (meaning institutions/ organisations). While it is evident that this definition reflects the need to prove the usefulness of such initiatives, their impact, and the purpose of allocating relevant amounts of public money to them, it is not always clear what a "community of practice" adds to the concept of users itself and what it should look like in reality.

The word "community" may allude to users that recognise an aggregation point in a specific initiative and find room to share the outcomes of data reuse. It may translate into particular features on the platforms designed for data-sharing. Nevertheless, a community will likely remain an elusive idea, and what we get to know is only a tiny percentage of users that may not even be significant in comparison to the needs and behaviours of those who reuse data on a regular basis or want to access and reuse it but find some sort of barrier, unless of finding rewarding ways of engagement.

Since tracking users has many questionable aspects (especially within the European Union), it seems to be more and more necessary to rely on users' willingness to make themselves visible if we want to understand who is within our "community of practice". One way could be to offer them the possibility to enrich metadata, adding the outcomes of their activities with the data. However, this example does not escape the trend when the catchment of users taken into consideration is mainly in academia and research. Users' engagement happens when they perceive their needs can be met or their problems can be resolved. In the case of enrichments, users may be interested in increasing the dissemination of their outcomes, especially when the platforms are designed to facilitate reuse beyond academia.

In what is called the "digital decade", the European Union is supporting the setup and development of fourteen data spaces in areas of public interest, one of which is the common European data space for cultural heritage. Unlike other data spaces, this one doesn't start from scratch.[9] With the Europeana Initiative acting as its steward, it benefits from the standards and frameworks for modelling, licensing and publishing data that have been developed and refined over twelve years within this initiative. By definition, Europeana is an initiative of the European Union and revolves around the platform of the same name. The platform currently provides access to 50 million digital items from approximately 4000 cultural institutions - mainly museums, libraries and archives - thanks to the intermediation of forty aggregators.[10] A network of 4000 professionals – the Europeana Network Association – complements Europeana making peer-to-peer knowledge-sharing concrete.[11] The word community also recurs in this context, as the network was restructured into seven communities in the past few years, and the one with the highest number of members is the Europeana Research Community.

[9] https://digital-strategy.ec.europa.eu/en/news/deployment-common-european-data-space-cultural-heritage
[10] https://www.europeana.eu/en
[11] https://pro.europeana.eu/about-us/mission

The common European data space for cultural heritage kicked off in September 2022. There is a strong focus on data reuse in the data space, and one of the largest catchment areas to consider is that composed of academic and research communities, which are, nevertheless, the most demanding. Bringing on board DARIAH (Digital Research Infrastructure for the Arts and Humanities) as an official partner, the Europeana Initiative has decided to strengthen its collaboration with the European infrastructure that represents the users potentially most interested in digital cultural heritage and to explore together the challenges and opportunities coming from the "Collections as Data" approach. In this transition phase, the Europeana Initiative also benefits from the experiences shared within the GLAM Labs Community, a group of cultural heritage professionals who animate labs at their institutions and recently tested a handy tool in the form of a "A checklist to publish Collections as Data in GLAM institutions".[12]

Beyond the publication of a "Collections as Data" workflow, for the benefit of cultural heritage institutions willing to become part of the data space, the Europeana Initiative and DARIAH will explore how to use this approach to inform product development within the Europeana environment. And here is where the idea of a "community of practice" is already evolving and translating into new features for the platform. Since the aim is to facilitate computational reuse of the data, we started reasoning around the concept of datasets, aware that it has a different meaning from content providers' and users' perspectives and that it may vary based on the academic disciplines the latter practise. Therefore, we launched a consultation leveraging the Europeana Research Community: What is a dataset to you?[13]

Think about a recent scenario where you needed to use cultural heritage data. Ask yourself:

- *What sort of data structure did you prefer?*
- *What sort of information did you need or want about the dataset?*
- *What kinds of digital cultural heritage objects were part of your dataset?*
- *Who created the dataset (you or someone else, e.g. a cultural heritage institution)?*
- *How large was the dataset? How typical is this for datasets that you have used in the past? How large do you imagine you could ever need them to be?*
- *When you were using that dataset, what made your research or teaching activities easier?*
- *What, if any, obstacle did you encounter in trying to reuse that dataset?*

The consultation is still ongoing. Perhaps not surprisingly, the first reactions didn't come from academics and researchers but from cultural heritage professionals solicited by the topic and the approach. Again, in their case, user engagement happens when they perceive their needs can be met or their problems can be resolved.

[12] https://glamlabs.io/checklist/
[13] https://pro.europeana.eu/post/what-is-a-dataset-share-your-thoughts

Community-owned infrastructure and collections as data praxis

Amanda Whitmire, Stanford University

In our marine science branch library, much of the collections as data work we do involves surfacing historical (mostly 20th century) oceanic and biodiversity observations. For example, we have extracted decades of oceanographic data from annual reports [1], published plankton population data from handwritten log sheets [2], [3], and transcribed biodiversity observations from a 1934 dissertation [4]. Most recently, we've been collaborating with colleagues across Stanford Libraries on Taxa, a project to extract species occurrences data (an observation of a given species at a particular place on a specific date) from the text of student research reports using a range of methods from named entity recognition to large-language models [5]. When we first started engaging in these collections as data projects, we took an admittedly narrow view of our process and infrastructure. We transcribed and transformed physical collections into datasets to make them computationally accessible and put them into the institutional repository, a straightforward workflow that didn't seem to warrant much reflection. More recently however, we've finally caught up to the idea that we need to "think expansively about the development of community-owned infrastructure that enables computational research" [6]. We have adopted a philosophy with a minor variation on this suggestion: where possible we should *contribute to existing community-owned infrastructure* as part of our praxis.

A crucial part of our workflow involves leveraging linked open data resources like authoritative taxonomic databases, georeferencing tools, and Wikidata to quality control and augment curated datasets. We rely heavily on these resources. The collections we curate are inextricably linked to the social data infrastructure of the web, and while our digital collections are designed to meet the needs of the research community we support [8], they can also contribute directly toward a global effort to share biodiversity data on community-owned platforms. When working with these materials, the data collection development we do should be an intentional contribution to the broader biodiversity knowledge graph, which is described as, "a network of connected entities, such as taxa, taxonomic names, publications, people, species, sequences, images, and collections" [7]. It turns out that where we share collections can be just as important as how, and the implications across our praxis are significant.

In pulling data from these open resources, we've noticed occasional content gaps. For example, in comparing a checklist of 459 scientific names of the plants and animals that are found in the nearshore environment around Hopkins Marine Station [9], we found that Wikidata is missing 7% of the species, some species have duplicate entries with unaccepted scientific names, and a third of species that were in Wikidata from our list are missing photographs. The work we do in support of marine science research is interconnected with resources like Wikidata, so it is well within the scope of our collection development work to contribute to this community-owned resource and help ameliorate these gaps. Contributing to a global resource like Wikidata also helps our local research community because they also use it as a resource. Our collection development and collections as data work are no longer siloed or limited to what is physically held in our library or our repository.

Looking ahead, I'm interested in discussing the concept of post-custodial approaches to collections as data praxis that is mentioned in the Always Already Computational: Collections as Data Final Report [10]. We deposit preservation-oriented datasets into Stanford Digital Repository with a light wrapper of documentation and provenance to contextualize it, but in the case of biodiversity datasets, they must reside in the Global Biodiversity Information Facility (GBIF) to gain context and utility by the research community. The data we contribute helps to build out the largest biodiversity catalog of species occurrences in the world, truly the most important bit of community-owned infrastructure in this research space. In contributing to collections that live outside of our organization, what we lose in control and local visibility is dwarfed by what is gained in context, discoverability, and overall usefulness. If we contribute data into community-owned infrastructure (both as datasets and as entities in Wikidata), does this adoption of a collaborative approach to open data constitute a form of post-custodial work?

I will close by adding that this critical realization of our status as a tiny node of the interconnected biodiversity knowledge graph has had a significant impact on how we approach our collections as data workflow development. Continuing with the example of the Taxa Project, we are working intentionally to develop a workflow that is as transferable to other institutions and settings as possible. For example, in the Stanford Digital Repository accessioning pipeline for scans of text-based documents, images are set to automatically go through OCR to produce an Alto XML file of each page of text. We could use this XML file as a direct input into our LLM pipeline, but since most of our colleagues with similar collections at marine science libraries would NOT have access to these kinds of XML files, we are instead incorporating an open software image OCR step into our workflow. This workflow will then be more directly portable to our community of practice.

Our collections as data praxis considers the needs of our local researchers, the context of our interconnected global research community, and our community of practice in cultural heritage institutions. Especially in the case of biodiversity data, we must be actively engaged in the open data community and contribute to community-owned infrastructure. Sometimes that takes the form of sharing data via community platforms, and sometimes it's contributing to shared infrastructure like Wikidata. It is also paramount that we grow our communities of practice by being thoughtful about how we build and share collections as data tools and workflows.

References

[1] R. Rodgers *et al.*, "Collections as Data Facets --- Always Already Computational: Collections as Data," May 2019, doi: 10.5281/zenodo.3066240.

[2] A. Whitmire, "Monterey Bay phytoplankton data from the Hopkins Marine Station CalCOFI program (1954-1968)," Aug. 2022. doi: https://doi.org/10.25740/pz376rp0413.

[3] A. Whitmire, "Monterey Bay zooplankton data from the Hopkins Marine Station CalCOFI program (1955-1963)," Aug. 2022. doi: https://doi.org/10.25740/nt620vn7810.

[4] W. Hewatt, M. Tabbarah, and A. Whitmire, "Data from: Ecological studies on selected marine

intertidal communities of Monterey Bay," Dec. 2021. doi: https://doi.org/10.25740/zd748qn2612.

[5] C. N. Coleman, J. Nelson, and A. L. Whitmire, "Taxa Project," *TAXA - a species verification tool*, 2023. https://taxa.stanford.edu/ (accessed Apr. 10, 2023).

[6] T. G. Padilla, "Collections as data: Implications for enclosure," *Coll. Res. Libr. News*, vol. 79, no. 6, p. 296, Jun. 2018, doi: 10.5860/crln.79.6.296.

[7] R. Page, "Towards a biodiversity knowledge graph," *Res. Ideas Outcomes*, vol. 2, p. e8767, Jul. 2016, doi: 10.3897/rio.2.e8767.

[8] T. Padilla, L. Allen, H. Frost, S. Potvin, E. Russey Roke, and S. Varner, "Santa Barbara Statement on Collections as Data," May 2019, doi: 10.5281/zenodo.3066209.

[9] L. Reyserhove *et al.*, "A checklist recipe: making species data open and FAIR," *Database J. Biol. Databases Curation*, vol. 2020, p. baaa084, Nov. 2020, doi: 10.1093/database/baaa084.

[10] T. Padilla, L. Allen, H. Frost, S. Potvin, E. Russey Roke, and S. Varner, "Final Report --- Always Already Computational: Collections as Data," May 2019, doi: 10.5281/zenodo.3152935.

Data Models and Library Collections: Observations from a Digital Scholarship Center

James Lee, University of Cincinnati

I am basing my position response to the "Collections as Data: Part to Whole" document from my vantage point as the director of a library-based Digital Scholarship Center (DSC), as well as my position as a university-level administrator in the Provost's office of my institution.

Context: The Andrew W. Mellon Foundation has supported the creation and expansion of the University of Cincinnati's DSC in order to address two problems that complement the Collections as Data mission: how to apply machine learning and data visualization techniques to digital collections to enable the investigation and analysis of scholarly research questions, and how computational analysis of large unstructured datasets using machine learning can enable a transdisciplinary culture of library-centered team scholarship involving mixed teams of faculty, librarians, IT professionals, and students. The DSC uses technical innovations in machine learning and data visualization on large, unstructured humanistic datasets in text, image, audio, and video formats to investigate complex, crosscutting research questions that call for collaboration among humanists, social and natural scientists, librarians/information professionals, and others. Our technical methodology in practical terms has been built upon our machine learning and data visualization platform, "Model of Models" (MoM), which can perform multiple machine learning modeling techniques in parallel on large text and image datasets, and presents results as a set of interactive visualizations that permit non-technical researchers to explore, annotate, and cross-validate findings across the parallel modeling methods.

In partnership with faculty and motivated by their research questions, our teams serve as a catalyst, to use the chemical metaphor, by synthesizing a reaction of different components into a cohesive product and by reducing the barrier to entry for such a reaction to commence. Our transdisciplinary research groups genuinely span multiple disciplines, drawing from the "team science" model used in biomedical research, with people trained to think very differently about every step in the research process: formulating research questions, gathering relevant data, analyzing information, and presenting conclusions. The DSC works with the team to acquire the appropriate datasets, conduct data analysis, perform hypothesis testing, develop plausible arguments, and disseminate our research findings in analog and digital forms. In contrast to previous efforts, our digital scholarship efforts directly spark digital modes of inquiry by connecting faculty with digital collections and analytical tools that enable them to approach research questions in ways that may not be possible with the standard methods of their field.

Data Models and Digital Collections:

A key distinction that has animated our thinking about the potential of defining library collections as data has been a recurring mismatch between library-based collections and research-oriented models. Our research partners in the research and teaching colleges have not demonstrated much interest in

collections as such, but rather, engage with our DSC by creating, analyzing, and validating the computational models we generate based on the data contained within the collections. To begin to bridge this divide, we are exploring ways to intentionally integrate the many possible machine learning models derived from a dataset as individual tests of a research question, into our definition of a digital collection. Whereas the collection is a stable base of information suitable for long term reproducibility and replicability, the models built from the data underlying the collection are formatted and deployed for ease of interpretation, testing, and rapid iteration, rather than access and preservation. We have been studying some of the model access and storage practices of Hugging Face and genomics repositories, and have incorporated the use of model cards as a way to categorize and link models with their underlying data collections.

Based upon my experiences in the DSC, I hope to contribute several opportunities and challenges to the Collections as Data initiative.

Opportunities:

- Reimagining collections as data for computational analysis has opened up a significant level of demand among faculty and graduate students interested in collaborating with the DSC to apply machine learning methods to their research questions. In this manner, the DSC's computational approaches have served as a gateway to increased researcher collaboration with the libraries overall.

- The DSC's "Catalyst" model has made the library's mission and value clearer to departments and individual researchers in STEM and other already data-driven fields, who do not engage with traditional collections to the extent that we can observe from faculty in, for example, Colleges of Arts and Science. In this sense, the DSC's scope has expanded the library's reach into new disciplinary domains.

- As the university's existing steward of information, the DSC specifically and the libraries more generally have emerged as an intellectual hub in the data science and artificial intelligence initiatives on our campus.

- Our development of the Model of Models machine learning and data visualization platform, and the projects that have been accomplished using this technology, demonstrate that the technology exists to enable a library-centric Collections as Data approach specific to research, including software packages and open-source code libraries, high performance computing, and cloud resources.

- A team science approach to computationally-driven digital scholarship has integrated librarians tightly into research teams through the entire project lifecycle. This direct engagement with the research process has led to librarians' receiving credit as co-authors in peer-reviewed publications, conference presentations, and grant funding. The recognition of librarians in the research teams has contributed to promotion and tenure cases, and greater awareness among research faculty of how librarians can contribute to the research process.

Challenges:

- The integrated Collections as Data paradigm articulated in the proposal may encounter cultural, rather than technical, difficulties. At our institution, the conventional organization of library departments that work in the digital space include reference, metadata, collections and acquisitions, digital repository, and IT. We have experienced challenges in attempting to reframe these organizational structures in more elastic terms aligned with user needs. The main cultural pressure points include how to redefine job descriptions beyond conventional library roles, how to attribute headcount to departments, and how supervisory / reporting lines can be distributed across units. As solutions, we have experimented with the formation of a cross-cutting research and data services unit, the implementation of a matrixed organizational structure, and allocating a percentage of librarian weekly time to grant-supported research activities to support their professional development.

- Position papers from the previous cohort of respondents have noted that the data formats underlying a Collections as Data practice can be difficult to reconcile. Data prepared for preservation in the digital repository is not necessarily amenable to computational analysis, thus introducing barriers to reproducibility and replicability of research results.

- Our collections budget has been a protected budget line, and is allocated by the Provost for the primary purpose of supporting the development of the physical collection, with digital collections purchased within the conventional acquisition workflow. We have debated the consequences of opening up our collections budget to a future model where digital, rather than physical, collections compose a significant portion of our acquisitions, with the biomedical journal collections strategy used by our health sciences library as one digitally-oriented model under consideration.

Extracting Black Humanity, Blurring Collections into Data

Dorothy Berry, Digital Curator, National Museum of African American History and Culture SI

GLAM institutions remain in thrall to technological possibilities that offer hints of innovation or that serve as shibboleths, letting everyone know who is the digitally-aware person in the room. For better or worse, with logic or none, *collections as data* has become one of those phrases that gets thrown into the mix more to represent being up on the times than to signify anything particularly concrete. This is not, necessarily, a flaw of collections as data but can signal a concerning direction in a number of related fields that have long histories of valuing the object over the person, the material representation over the less tangible, research value over anything.

From my particular vantage, this becomes an issue when dealing with large scale collections of archival material related to the lives of Black people and excitement expressed by the predominantly White GLAM tech leaders around scaling research in ways that further serve the individual blurring of thousands of "unknown Black men/women/children" into data, designed to serve whose research, if anyone's?

My questioning position on *CaD* as I hope I've articulated thus far is not related to the concept at its core, but rather to the ways we continually obfuscate complex humanity in archival collections for the sakes of neutrality, the pursuit of innovation-related funding, or simply to avoid the fact that our institutions may be more highly staffed with technological expertise than any knowledge focused on the histories of marginalized peoples. *Collections as Data* seems most successful in contexts where there is a clear connection between the desire for data and the collections content- a clear narrative of what can be gained in aggregation but also a clear view of the importance of the individual stories within that aggregation. The macro view only serves the individual who is already valued as an individual, and that feels like an unproven presumption across many GLAM institutions.

My position is a narrow one, driven narrower still by an awareness of my own professional limitations and focus. With various successes in the "how" of *Collections as Data*, I still see a need for a clearer understanding of what justifies the "why," especially when the "how" can play into some of our more concerningly distancing practices.

Future Directions

Gimena del Rio Riande (CONICET)

As February 2023 is passing by and I am writing these lines, the news regarding the conquest of our lives through artificial intelligence are overwhelming.[14]There is no website, newspaper, or social network where this is not mentioned either in a positive or negative way. There are almost no middle terms to describe the future that is already here. So, if I am to think about the *future directions* of the *Collections as Data* project - or directions that, in addition to being future, are both global, equitable, ethical and useful for those who work with humanities data, I must start with this new state of the art technology.

Last October I was honored with an invitation to participate in “A Roundtable on Machine Predictions and Synthetich Text”[15], which gave me the opportunity to work with some of the smartest voices in the data humanities field. It resulted in a series of responses[16] to the enlightening paper "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" by Emily Bender, Timnit Gebru, Angelina McMillan-Major and Margaret Mitchell.[17] Both Bender et al. work and this online roundatable reflected on the dangers of both artificial intelligence and large language models for humans and the planet. By that time I had already started investigating about the large language model developed by the National Library of Spain, which, although its model based mainly on the scraping websites from that country, it paradoxically proposed a model on a global scale. My findings pointed to the fact that, despite the fact that it is true that any Spanish speaker would understand a language model based on a variety from Spain, both identity and linguistic representation/non-representation would still be a problem when using this utterly biased model.

Today danger seems to lurk from another flank: Companies dedicated to the development of artificial intelligence have been working really quietly on everything, absolutely everything that is in the open on the web. None of these companies, which have invested billions of dollars to develop artificial intelligence by hiring some of the most talented programmers and developers in the world, are disclosing 100% of their goals, methods and objects of study. Hence, as an academic committed to the open access and open knowledge initiatives, I am experiencing a bitter feeling: we have been feeding the monster. Every article, every image, every datum we have struggled to make accessible and open, and every entry we have written to democratize knowledge on Wikipedia is now part of this invisible (semi)deity that seems to be everywhere and know everything.

In addition, all the systems underlying the construction of digital infrastructure and devices are tainted both by inequality and bias. Artificial intelligence sustains its automation and speed largely on the work

[14] The Real Academia de la Lengua Española (Royal Academy of the Spanish Language) chose *artificial intelligence* (*inteligencia artificial*) as the most important term from 2022.

[15] https://cdh.princeton.edu/updates/2021/10/03/join-us-for-a-roundtable-on-machine-predictions-and-synthetic-text/

[16] https://startwords.cdh.princeton.edu/issues/3/

[17] https://dl.acm.org/doi/10.1145/3442188.3445922

of annotators who tag millions of texts and are paid extremely low salaries. And, ironically, do you know where these human annotators are based? Well, mostly always in the Global South. This year in late-January, an article published in *Time* magazine showed how underpaid and exploited content moderators in Kenya were used to label harmful texts for the ChatGPT software.[18] and how outsourced initiatives related to OpenAI, etc. were growing in Latin America.[19]

In addition to millionaire investments, huge datasets and large troops of workers, artificial intelligence technology requires an enormous amount of computing power and cloud computing to train and operate these systems that, for the moment, only the giants of the Global North seem to be developing, as can be read in the policy of the National Research Cloud (NRC):

> Artificial intelligence requires vast amounts of computing power, data, and expertise to train and deploy the massive machine learning models behind the most advanced research. But access is increasingly out of reach for most colleges and universities. A National Research Cloud (NRC) would provide *academic and non-profit researchers with the compute power and government datasets needed for education and research*. By democratizing access and equity for all colleges and universities, an NRC has the potential not only to unleash a string of advancements in AI, but to help ensure the U.S. maintains its leadership and competitiveness on the global stage.[20]

Personally, I consider this scenario too big and too ethically, environmentally, and epistemologically compromised for me to include artificial intelligence tools or processes as part of the technologies involved in my research activities. However, I am against banning its use but absolutely for asking ourselves what those interested in the humanities and cultural heritage, or in humanities data, could do to improve this unequal context. How can *Collections as data* continue to work on globally valid models of responsible adoption and implementation of its resources?[21]

In these last stages of development of *Collections as data*, the project has gone from "part to whole", putting its proposal into practice in different cohort projects that have operated as implementing models and have made their collections as data. From what I could read in the final reports of each cohort project, I believe that the goals sought for this stage –variation in institutional infrastructure and staffing to support collection implementation; experimentation with repository centric and/or non-repository centric solutions; and commitments to utilization of open source technologies that interoperate with a broader open scholarly communication infrastructure– have been more than attained. Furthermore, I notice both the thematic and technological diversity of each of them, and the different scale at which they have worked. However, all of these projects were developed in institutions from the United States of America. Even taking into account the smaller or larger scale or team structure of the projects, these

[18] https://time.com/6247678/openai-chatgpt-kenya-workers/

[19] https://www.bbc.com/mundo/noticias-64827257?at_link_origin=BBC_News_Mundo&at_link_id=938D2328-B909-11ED-A6FE-77DCFF7C7F44&at_medium=social&at_link_type=web_link&at_ptr_name=facebook_page&at_format=link&at_campaign_type=owned&at_bbc_team=editorial&at_campaign=Social_Flow&fbclid=IwAR25TC34EWoPKlzdJfL9eUEnLNA-IvBxmNPv_ZxpB-EKTLmNKe_CDqB0tXM

[20] https://hai.stanford.edu/policy/national-ai-research-resource

[21] https://collectionsasdata.github.io/resources/

are initiatives that base their work on robust infrastructures and on experienced teams. But can these models be adopted and adapted to other regions with less (and utterly different) technological development? Can collections work on a small scale? How can this contribute to narrowing the technological gap that artificial intelligence threatens to widen? Do we need a/many (re)adaptation(s) and transculturation(s) of what has been done so far in the Collections as data project?

I personally believe that the ideal starting point is to go South. Investigating whether and how similar initiatives have been carried out in other latitudes can be surprising. Even if they are smaller scale initiatives or outdated projects, a reverse engineer plan should be put into practice. Moreover, both collaboration and partnering in projects is essential As the saying goes, actions speak louder than words. We are all know aware of the fact that, in most of the Global South countries, funding such as the one provided by the Andrew W. Mellon Foundation is non-existent. Ultimately, there is a big need to rethink standards, technologies and disciplinary or academic cultures. Most of the metadata models in current use in libraries around the world were either developed by Dublin Core, the Networked Digital Library of Theses and Dissertations, or the Library of Congress. These are imposed standards that are used every day and are case-adapted to a cultural, linguistic, economic and social context that is completely different from that where this technology originated. Besides, open source and access are two terms to have in mind. Many communities can't engage in projects in which work is done using proprietary software, or in which work is mainly done via cloud computing services. Lastly, it is always core to remember that disciplinary cultures vary around the world, therefore, many assumptions, values or even tasks that were part of these many cohort projects might differ in other countries.[22]

I strongly believe that Collections as data could be defined today, at least in the United States, not only as a project but also as a community of practitioners. While it is expected that the initiative can grow organically in North American institutions, paying attention to big –and very probably AI infused projects—the state of the field needs to be honestly assessed on an international level in which these future directions must be included, as I have already mentioned, together with other very different contexts of use for these project implementations while working on a set of possible shared commons. A case that could serve as example of what I propose is the use of Minimal Computing technologies in DH. Minimal computing can be a solution for the development of projects in the Global South, where access to infrastructure such as web hosting or even reliable and affordable Internet access is limited for humanities students and faculty since it can also represent a set of shared principles and open technologies to empower students and scholars to work autonomously on their own projects and solve sustainability issues.[23]

From my perspective, the *Collections as data* should now move in these future directions from whole to part, where this *wholeness* is a micro example different enough, open enough, technologically neutral enough, and participatory enough to be reproduced, but also human enough to critically understand collections as data.

[22] Asking for feedback on the Santa Barbara Statement of the 50 things to have in mind might be a good idea. Translations and living documents are always welcome.
[23] http://www.digitalhumanities.org/dhq/vol/16/2/index.html

Humanities Data: Lexical, Bibliographical, Archival

Barbara Bordalejo, Humanities Innovation Lab, Humanities Data Inquiry

Humanities Data Inquiry (HDI) has been researching humanities data as part of a SSHRC grant that aims to build a map of data use in the Humanities and to analyze current practices. In my background as a textual scholar and a practitioner of Humanities Computing, data were a fundamental and unquestioned part of my research. From the most basic units of morphological lexical data to the construction of textual apparatuses as analogue databases; and from virtual library collections to the metadata of the same, we are faced with subjects that can be classified, visualized, and understood. In other words, Digital Humanists recognize digital objects as data (Padilla 2017).

For other humanists, mainly analogue ones, the idea of "data" is more closely related to the sciences and social sciences than to their day-to-day work. Their understanding is disengaged from the idea of humanities data even though historically, lexical and bibliographic data have been the constituent basis of humanities research. The most obvious examples are dictionaries and encyclopedias as collections of lexical elements, but also library collections, whether general or particular. In literary and historical studies, as well as for textual critics interested in making editions, the use of representations of cultural objects (surrogates) for research exemplifies the use of what we have come to call representational data. Traditionally, these analogue representations existed as facsimiles, copies, editions, or comparative tables, but contemporary applications almost always result in digital representational data, which need to be classified, exchanged, and managed.

Whether those collections of representational data exist in a particular institution or whether they form part of a curated hub creating a virtual collection. For the purposes of HDI, we started from a hypothesis in which humanities data (Borgman 2015) might have been conceivable as different from, for example, data in the sciences. However, a year and a half into the project, it has become clearer that humanities data have a history (think of the textual annotation system first devised in Alexandria for Homeric texts) and that research data in the humanities and sciences are not intrinsically different. If this is correct, the problem is a problem of recognition rather than essence. The computational component of DH forces researchers to recognize they work with data, but for many reasons, this is not true for analogue humanists. Influential positional perspectives made public in and around 2012 (Fish 2012b; Liberman 2012; Marche 2012; Fish 2012a) created a chasm between digital and analogue humanists that persists to this day and has separated two factions that hardly engage with each other.

The process of understanding analogue humanists' data and information practices, including their resistance to data, exacerbated by the perception that the concept of data is foreign to the humanities, currently drives a large part of our research. Technophobias might be playing a large role in this, perhaps due to the idea, now reinforced by pervasive AI, that computation attempts to replace human thought and careful analysis.

Borgman, Christine L. 2015. *Big Data, Little Data, No Data: Scholarship in the Networked World*. Cambridge, Massachusetts: The MIT Press. ieeexplore.ieee.org/xpl/bkabstractplus.jsp?bkn=7040497.

Fish, Stanley. 2012a. "The Digital Humanities and the Transcending of Mortality - The New York Times." *The New York Times*, January 9, 2012. https://web.archive.org/web/20200715124709/https://opinionator.blogs.nytimes.com/2012/01/09/the-digital-humanities-and-the-transcending-of-mortality/.

———. 2012b. "Mind Your P's and B's: The Digital Humanities and Interpretation." *The New York Times*, January 23, 2012. https://bit.ly/3gKM3mC.

Liberman, Mark. 2012. "The 'Dance of the p's and b's': Truth or Noise?" *Language Log* (blog). January 26, 2012. http://languagelog.ldc.upenn.edu/nll/?p=3730.

Marche, Stephen. 2012. "Literature Is Not Data: Against Digital Humanities." *Los Angeles Review of Books*, 2012. http://www.lareviewofbooks.org/article.php?id=1040&fulltext=1.

Padilla, Thomas. "On a Collections as Data Imperative." (2017). digitalpreservation.gov/meetings/dcs16/tpadilla_OnaCollectionsasDataImperative_final.pdf

The neverending story: law and ethics challenges to data-driven scholarship

Timothy Vollmer, UC Berkeley Library

The Santa Barbara Statement on Collections as Data appropriately centers ethical considerations in building and using collections, with a particular focus on "respect[ing] the rights and needs of the communities who create content that constitute collections, those who are represented in collections, as well as the communities that use them" (Principle 2). The University of California Berkeley Library, in collaboration with librarians, legal experts, and information professionals in the U.S. and abroad have been grappling in parallel with understanding ethics questions as well as the broader law and policy issues in supporting both text data mining (TDM) and digitization of and access to special collections.

Supporting researchers to navigate law and ethics in text data mining

The UC Berkeley Library has led a few different projects to help researchers understand the law and ethics of using online or licensed collections as data. For instance, Samberg and Hennesy explored issues such as copyright, contracts, and ethics in text data mining, and discussed guidance for researchers in how to traverse the landscape of computational text analysis.[1]In 2020, the UC Berkeley Library hosted the *Building Legal Literacies for Text Data Mining Institute*.[2] The project brought together expert faculty from across the country to train 32 digital humanities researchers and library professionals on how to navigate law, policy, ethics, and risk within text data mining projects. The institute was a success in that it improved researcher confidence in navigating the substantive law, policy and ethical issues; encouraged library staff to improve institutional content licenses to specifically enable text data mining; and gave participants the tools to educate others about the TDM law and policy issues on their campuses.[3] We also published an openly-licensed eBook to guide others in understanding the TDM literacies.[4]

The institute uncovered additional challenges, including institutional risk aversion to supporting progressive text data mining practices, and confusion amongst researchers in engaging in TDM with content published or licensed abroad, or collaborating with researchers in foreign countries. We created a follow-on project, *Legal Literacies for Text Data Mining–Cross Border*, to explore the issues in greater depth through roundtable discussion between cross-border TDM practitioners and legal experts. We'll eventually produce an outline of key educational materials that could be created to guide international TDM research responsive to the legal and ethical challenges therein.[5]

One question posed by all of the instructional work we've done so far is: How can we understand and react to a dynamic legal system responding to the changing technological landscape for computational analysis? While crucial court decisions such as *Authors Guild v. Google* and *Authors Guild v. HathiTrust* have enabled TDM practice as a transformative fair use under U.S. copyright law, the explosion of new AI and machine learning tools have given rise to additional legal challenges, including several copyright infringement lawsuits relating to AI image generation.[6] Another complication is the intersection of copyright and contract law, in particular the *ML Genius v. Google* case.[7] The issue is whether a website

can prohibit copying via terms of use, thus overriding uses that might be permissible under limitations and exceptions to copyright law. The Association of Research Libraries has identified some possible pathways for navigating copyright and contracts issues.[8]

Regarding education of TDM law and policy issues, how can we adequately scale training for these legal literacies (copyright, contracts, privacy, ethics) for text data mining researchers and support professionals?

Collection digitization and use: balancing public access with ethical concerns

The UC Berkeley Library developed "responsible access workflows" to guide the library in evaluating copyright, contracts and donor agreements, privacy concerns, and ethical considerations in digitization of and access to special collections materials.[9] We gave specific attention to the ethics questions, particularly since these are incredibly context-specific and there is no general consensus about how to approach ethical dilemmas when working on mass digitization projects. A library working group produced a draft local policy for evaluating access when ethical questions arise.[10] The recommendation from the group focused on balancing expected value with potential harm: we ask whether the value to cultural communities, researchers, or the public outweighs the potential for harm or exploitation of people, resources, or knowledge.

The Library is actively working on projects that foster the ability for researchers to explore collection content as data. For example, UC Berkeley Library hosts materials related to the incarceration of Japanese Americans in camps across the western U.S. during World War II.[11] These types of materials include prisoner intake forms, photographs, diaries, and maps. While there may be few copyright, contractual, or privacy restrictions in digitizing and making the content available, there could be ethical questions that prompt the Library to consider whether it should be shared publicly, particularly content that is seen as sensitive or problematic by community stakeholders. The Library is engaging with these stakeholders to make informed decisions about ethical access and reuse of collections as data.

[1] Samberg, R. G., & Hennesy, C. (2019). *Law and Literacy in Non-Consumptive Text Mining: Guiding Researchers Through the Landscape of Computational Text Analysis*. https://escholarship.org/uc/item/55j0h74g [2] *Building Legal Literacies for Text Data Mining Institute*. (2019). https://buildinglltdm.org/institute/ [3] *Public Version—Building Legal Literacies for Text Data Mining: Institute White Paper*. (n.d.). Google Docs. Retrieved April 4, 2023, from https://docs.google.com/document/d/107Qmu595-7aOc2DPWc4vVDbz7PTULzn_20NOeNfBbWM/edit?usp=drive_we b&ouid=105265392674002944423&usp=embed_facebook
[4] Althaus, S., et al. (2021). *Building Legal Literacies for Text Data Mining*. University of California, Berkeley. https://doi.org/10.48451/S1159P
[5] *Legal Literacies for Text Data Mining – Cross Border ("LLTDM-X")*. (2022).

[6] Lee, T. B. (2023, April 3). *Stable Diffusion copyright lawsuits could be a legal earthquake for AI*. Ars Technica. https://arstechnica.com/tech-policy/2023/04/stable-diffusion-copyright-lawsuits-could-be-a-legal-earthquake-for-ai/ [7] *ML Genius Holdings LLC v. Google LLC, No. 20-3113 | Casetext Search + Citator*. (n.d.). Retrieved April 4, 2023, from https://casetext.com/case/ml-genius-holdings-llc-v-google-llc

[8] Klosek, K. (2022). *Copyright and Contracts: Issues & Strategies*. Association of Research Libraries. https://www.arl.org/wp-content/uploads/2022/07/Copyright-and-Contracts-Paper.pdf

[9] UC Berkeley Library. (2020, May 14). *UC Berkeley Library makes it easier to digitize collections responsibly with novel workflows and bold policy*. https://lib.berkeley.edu/about/news/responsible-access

[10] *UC Berkeley Library_Ethics Local Practices_DRAFT*. (n.d.). Google Docs. Retrieved April 4, 2023, from https://docs.google.com/document/d/10Ux--7GgrOvoYzTAlbTRtdVAbENYRsCr9HmUNGw9LC4/edit?usp=embed_fac ebook

[11] Friedman, M. (2021). *Using AI/Machine Learning to Extract Data from Japanese American Confinement Records*. https://doi.org/10.48448/c1rq-qf28

Overview of the Language Data Commons of Australia (LDaCA)

Robert McLellan, LDaCA

Australia is a massively multilingual country in one of the world's most linguistically diverse regions. Significant collections of this intangible cultural heritage have been amassed. However, these materials are held in different geographic locations by a variety of institutions as well as in digital repositories, meaning that discovery and access procedures are inconsistent and, at least in some cases, difficult. Thus, while Australia's rich linguistic heritage is well documented, much of this data remains inaccessible or under-utilised due to barriers to accessing collections, as well as a lack of access to tools and skills for analysing that data at scale. The LDaCA program is designed, as part of a broader Humanities, Arts and Social Sciences and Indigenous Research Data Commons (HASS & IRDC) funded through Australian Research Data Commons (ARDC) and the National Collaborative Research Infrastructure Strategy (NCRIS), to work across researchers, communities and institutions to secure and unlock this rich linguistic heritage, and to provide the tools and skills necessary for researchers to analyse language data at scale and exploit the full research potential of those collections for local and national good, and also providing secure access to fragile language records of the Pacific region.

The LDaCA program aims to further the following objectives:

1. Develop the social and technical foundations for a national, distributed archival repository, including:

    a. shared data governance and standards framework;

    b. shared data access, authentication and authorisation policies, procedures and processes;

    c. shared technical infrastructure for curation and storage of language data;

    d. shared technical infrastructure for collection and annotation of language data.

2. Secure vulnerable and nationally significant collections of Aboriginal and Torres Strait Islander languages, Indigenous languages in Australia's Pacific region, (varieties of) Australian English and migrant languages, and sign languages of Australia and its region.

3. Develop a national data portal for accessing and repurposing language data of significance to researchers and communities that is held in GLAM institutions, including libraries, archives and museums, alongside language collections held in the distributed archival repository.

4. Establish an integrated analytics environment (ATAP: Australian Text Analytics Platform) for researchers to create fully described, reproducible research on written, spoken, multimodal and signed text in accordance with Open Science principles, and aligned with community expectations for research of practical benefit.

5. Provide training and develop resources for researchers and communities to support best practice in accessing, analysing and archiving language data in line with FAIR and CARE principles.

The different components of the LDaCA infrastructure are summarised in Figure 1 below.

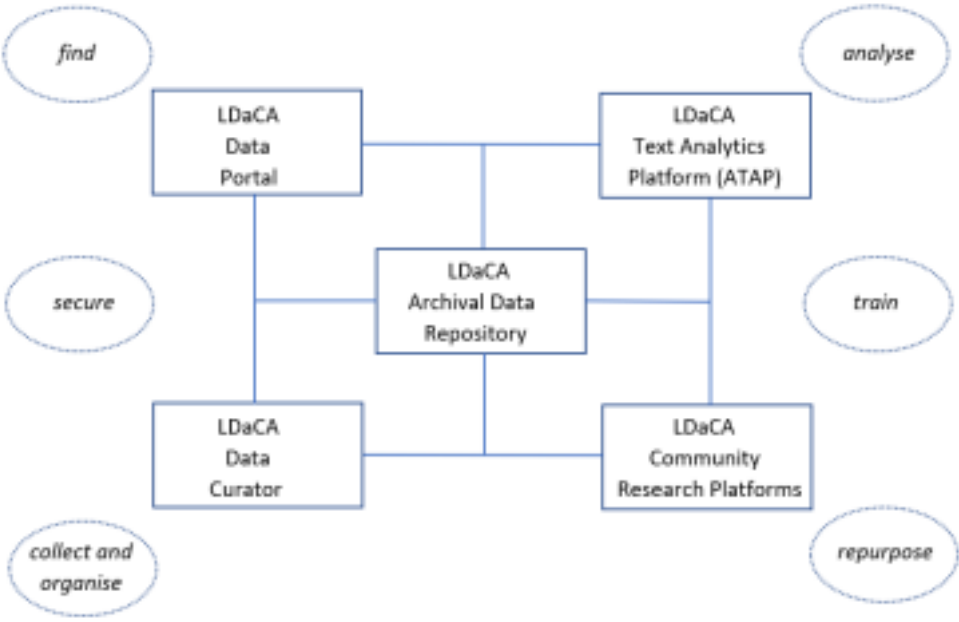

Figure 1:
Overview of LDaCA Infrastructure (2023-2028)

Benefits of LDaCA for researchers and communities

The resulting infrastructure, while aimed at language researchers, will also be relevant to cognate humanities disciplines, such as communication, history and Indigenous studies, as well as researchers who use text analytics in other fields including the social sciences, environmental sciences and health sciences, as its tools will be readily available to researchers across a range of disciplines interested in employing text analytics methods to analysed unstructured datasets. It will also benefit researchers in computing and information sciences wanting to access high quality text collections for the development of next-generation tools in NLP and text-based AI and machine learning.

LDaCA will also support a culture of greater community and industry participation in research through the development of a social architecture that enables co-design and co-development by researchers, communities, GLAM institutions and industry. In particular, the project will serve as a point of reference on how to implement multi-tiered access control (using the CILogon access control tool) and navigating multiple data usage licences in the same environment. Additional support will be provided to communities who already control and manage their own data to evaluate their existing data management practices and how to improve their organisational capability and self-sufficiency. In the longer term, this will lay the foundation of a national language data commons that supports the revitalisation of Indigenous languages across Australia and its region, as well as maintaining other community languages of national significance, thereby fostering greater community and cultural well being.

Commitments of LDaCA

When working with language material, LDaCA is guided by FAIR and CARE principles (Russo Carroll et al., 2018. Scientific Data, 2022). The CARE principles extend these data management principles, ensuring that Indigenous communities benefit from the data, that authority to control Indigenous data is held by Indigenous Peoples, we have a responsibility to share how data is used to collectively benefit Indigenous Peoples, and that Indigenous Peoples' rights and wellbeing are the primary concern at all stages of the data life cycle.

The LDaCA project will make nationally significant language data available for both academic and non academic use and provide a model for ensuring continued access with appropriate community control. It is primarily intended to meet the needs of researchers, but it will provide access to language resources for other groups. For example, people with a non-academic interest in languages will be able to use it to explore materials, and teachers and students at different educational levels will be able to find data relevant to their interests and needs.

Further information

For further information about LDaCA see: https://www.ldaca.edu.au/

For further information about the Australian Text Analytics Platform (ATAP) see: https://www.atap.edu.au/

Lessons Learned from Building, Documenting, Using, Supporting, and Expanding Collections as Data

Amanda Henley, University of North Carolina at Chapel Hill

My position on collections as data has been shaped by work over the past four years on a project called On the Books: Jim Crow and Algorithms of Resistance (OTB).[1] This project was funded through Collections as Data: Part to Whole (cohort 1).[2] We created a plain text corpus of North Carolina session laws between 1865-1967 and used machine learning to identify laws likely to be Jim Crow. We designed our project based on principles outlined in the Santa Barbara Statement on Collections as Data.[3] The work has been guided by ongoing ethical commitments and our workflows and product design focus on reusability. The project has used open-source solutions, created heavily documented code and tutorials to explain workflows, and made end products accessible. We use an iterative process for creating and improving products and methods.

Our latest iteration of this work is a new project funded by the Mellon Foundation. The project is expanding the work of OTB to two additional states through a partnership with teams at university libraries at the University of South Carolina and the University of Virginia. The project also assesses the needs of collections as data users by supporting fellows as they use OTB products in research and teaching.

The project has had many collaborators and this experience has been a growth opportunity for staff members and graduate student workers. The team has learned about designing reproducible or adaptable workflows, creating textual datasets (including parsing text and creating metadata), creating ancillary resources (essays, a website, and a data visualization), and assisting researchers and instructors in using them. Most importantly, it has taught us the importance of a team approach for this work, as the process has required diverse skills. Library staff have partnered with disciplinary scholars and worked across departments to do this work. Some of the lessons we have learned are summarized below.

Build a large, diverse team. The team composition specified in the Part to Whole CFP required the inclusion of a disciplinary scholar; a librarian, archivist, or museum professional with senior administrative responsibilities; and a project lead.[4] This guidance was integral to our success. Based on our experience, we would also recommend including a strong technical lead and a collections expert to ensure a comprehensive collection with potential for wide use.

There is a need for improved text parsing tools. The legal volumes we used for the collection are organized by law type (local, public, private), chapter, and section (sections are individual laws). To analyze the corpus at the law level, we needed to split the volumes. Even using automated approaches, this work took over a year to complete and was one of the most challenging aspects of the project. Improved text parsing methods will benefit collections as data by reducing the time needed to separate texts into meaningful sections and by simplifying creation of detailed metadata. This will ultimately allow users more flexibility in defining their unit of analysis and provide them with the option of using more detailed metadata for modelling purposes.

Focus on workflow adaptability, not reproducibility. State partners at USC and UVA have worked with the OTB team to adapt our workflows, methods, and scripts to create legal corpora for their own states. To support their work, the UNC team replaced hard-coded values in scripts to adjustable parameters. A focus on adaptable code rather than reproducibility allowed partners to use and innovate upon our workflow, even though the legal volumes are different for each state.

Instructors appreciate a user interface. We supported a cohort of three teaching fellows who created and integrated modules featuring OTB materials into their courses. Instructional use of OTB provided valuable feedback for product improvements. We found it noteworthy that none of the teaching fellows were interested in using data sets in their modules, preferring instead to have students interact with the project website and data visualization. Creating these products required significant effort beyond creating the corpora and were only possible because of our extensive text parsing and inclusion of metadata. In our case, some of these products were created during the second phase of the project.

Collections about sensitive topics may benefit from supplemental materials. Teaching fellows agreed that the sensitive nature of OTB added to the difficulty of teaching with the material, but the importance of teaching the topic ultimately outweighed their apprehension. Through interviews and focus groups, we learned that instructors valued supplemental materials such as essays and curated links to additional resources that contextualize the collection. This has implications for collections as data, as projects led by ongoing ethical commitments may best support instructors by providing ancillary materials for use in scaffolding assignments. All of the teaching fellows required their students to read at least one of the essays provided on the OTB website as background reading.

Choose your collections carefully. Teaching fellows talked about how the local focus of OTB made the collection more meaningful for students, as many used OTB products to identify Jim Crow laws in their own hometowns. Research fellows' inquiries are focused on North Carolina as well. It is noteworthy that, despite extensive outreach efforts, we received fewer proposals than expected for teaching and research fellows; conversely, we received seven outstanding proposals from academic libraries across the country interested in being state partners. It's possible that the local nature of the collection made it more meaningful to some than to others. One potential explanation is that data sets are still used less than traditional resources for many scholars likely to be interested in North Carolina Jim Crow legislation. Identifying which collections to use as data remains one of the most critical decisions to the collections as data enterprise; input from collections experts and public services librarians are critical for identifying collections that are both comprehensive and most likely to be used.

Be prepared to support users. Three research fellow teams are currently working on projects that use OTB products. Two are using the OTB corpora, and both have required considerable support from our technical team. At UNC, we support faculty interested in humanities data and we teach digital scholarship methods to promote innovative research. Based on our experience, outreach and assistance are both needed for successful scholar engagement.

Conclusion

Creating and supporting the use of collections as data is both labor intensive and rewarding. In doing this work, library staff have advanced skills, trained students, improved interdepartmental cooperation, and

strengthened relationships with faculty. A successful project starts with a carefully selected collection and a large, diverse team that includes collections experts, disciplinary scholars, a project manager, and a technical team with a strong lead. Improved text parsing tools will facilitate the creation of textual collections as data and improve the utility of corpora. Heavily documented code with micro-case studies and instructional exercises allow others to understand, adapt, and innovate upon workflows. A user interface and ancillary materials allow for collections to be more easily integrated into instruction. By using the datasets to create secondary products, library workers become more familiar with the final products and are better prepared to assist researchers as they use them. Projects guided by ongoing ethical commitments are meaningful to those creating and using the products, and can benefit from the addition of essays and other materials that contextualize the collections.

Works Cited

[1] *On the Books: Jim Crow and Algorithms of Resistance*. University Libraries at the University of North Carolina at Chapel Hill. Retrieved 4/1/23, from https://onthebooks.lib.unc.edu/

[2] *Collections as Data: Part to Whole*. Part to Whole. Retrieved 4/1/23, from https://collectionsasdata.github.io/part2whole/

[3] *Always Already Computational - Collections as Data*. The Santa Barbara Statement on Collections as Data V.2. Retrieved 4/1/23, from https://collectionsasdata.github.io/statement/

[4] *Collections as Data - Part to Whole*. Call for Proposals. Retrieved 4/1/23, from https://collectionsasdata.github.io/part2whole/cfp/

A different level of abstraction: Contextualising computational access

Dr. Leontien Talboom, Cambridge University Libraries

Access to digital material has been one of my main focuses in the last couple of years. This was the subject of my PhD thesis and during my studies I became aware of the Collections as Data project. Providing access to material as data, or computational access, quickly became the focal point of not only my thesis but also the talks and papers published alongside this work. Several of these projects were in collaboration with The National Archives UK and gave me an idea of what computational access may look like in practice (Beavan et al. 2021). My engagement with sector peers working to provide computational access in practice and carrying out my own practical work on this topic led me to writing a guide about getting started with computational access in collaboration the Digital Preservation Coalition (DPC) (Digital Preservation Coalition 2022).

This guide outlines four approaches to computational access. The first encourages digital preservation practitioners to create a 'Terms of Use' because an institution may provide some type of access (e.g., metadata in a catalogue) even if it can't provide access through computational methods. Providing any type of access opens up the possibility of web scraping or other computational methods being used on collections, and it is good practice to define and publish guidelines on how collection materials are allowed to be used, as I found out with a coworker when working with UK Government Web Archive material (Talboom and Bell 2022).

The other approaches are datasets, APIs and platforms. Instead of with the 'Term of Use' these three approaches provide access to the metadata or data of an institution. We found it important in this work not to think of these approaches in an evolutionary manner, meaning that one was better than the other. With help from individuals within the digital preservation field we were able to create a list of case studies and examples showcasing that different institutions need different solutions and that even trying to provide computational access is a great way forward. The focus also was on the practitioners providing access, not the researchers using the material. With this guide we showcased that there are many possibilities to providing computational access, but that the most important thing was for institutions to start thinking about their collections in this manner.

It is so fascinating to see how much has changed even in the last couple of years of working in this area. Information specialists are starting to realise that there are new ways of opening up collections that go beyond the more traditional approaches such as keyword searches using online catalogues. These have their places within our field, but are not necessarily the most optimised way to provide access to digital material. A lot of work has been done around this topic in the last few years (Gilliland 2016; Winters and Prescott 2019; Milligan 2019), but even with computational access becoming more popular, there are still a number of topics around this that need to be explored. This was also pointed out when writing the DPC guide, examples of these topics are ethics and security. One topic that really piques my interest is the documentation needed to provide computational access in a meaningful way.

It seems that a different layer of abstraction is needed to make computational access fully accessible and useful. When material is made available at scale, the questions that researchers need answers to should also be answered at scale. Examples of these questions are why only a part of a collection was digitised or what techniques were used during the digitisation process. Within the computer science field, documentation at this scale has been discussed by Eun Seo Jo and Timnit Gebru, who call their own field the 'wild west' in regards to documentation and look to archivists and other information specialists for answers (Jo and Gebru 2020). Emily Maeruma has taken the work that they have done around datasheets into our field to discuss a way of documenting datasets (Gebru et al. 2018; Maemura et al. 2018).

During the Archive of Tomorrow project (Archive of Tomorrow Project 2022) a datasheet was created to provide extra context to researchers using the JSON metadata files from the targets (archived websites and pages) collected. This was a very interesting exercise for members of the technical team on the project and made us think of the material in a more abstract way: instead of focusing on metadata that goes with specific targets within the collection, the collections as a whole was considered. There were also questions within this dataset that raised concerns around recording certain processes and choices earlier on. The datasheet was created at the end of the project, when most targets had been collected and documented, prompting questions around what processes and methods should be documented throughout the creation of records to make them more usable for computational access.

The digital preservation field is moving in the right direction, making more data available at scale, and a large number of institutions are sharing their approach to computational access. However, thinking about documentation still seems to be at an early stage. The introduction of the datasheet has been great, but does open up further questions around when processes should be documented and at what scale.

Work cited:

Archive of Tomorrow Project. 2022. 'Archive of Tomorrow'. National Library of Scotland. 2022. https://www.nls.uk/about-us/working-with-others/archive-of-tomorrow/.

Beavan, David, Fazl Barez, Mark Bell, John Fitzgerald, Eirini Goudarouli, Konrad Kollnig, Barbara McGillivray, et al. 2021. 'Discovering Topics and Trends in the UK Government Web Archive'. Data Study Group Final Report. London: Alan Turing Institute.

Digital Preservation Coalition. 2022. 'Computational Access: A Beginner's Guide for Digital Preservation Practitioners'. Digital Preservation Coalition. https://www.dpconline.org/digipres/implement-digipres/computational-access-guide.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2018. 'Datasheets for Datasets', March. https://www.microsoft.com/en-us/research/publication/datasheets-for-datasets/.

Gilliland, Anne J. 2016. 'Designing Expert Systems for Archival Evaluation and Processing of Computer-Mediated Communications: Frameworks and Methods'. In *Research in the Archival Multiverse*, edited by Anne J. Gilliland, Sue McKemmish, and Andrew J. Lau. Social Informatics.

Clayton, Victori, Australia: Monash University Publishing.

Jo, Eun Seo, and Timnit Gebru. 2020. ‘Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning’. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, January, 306–16. https://doi.org/10.1145/3351095.3372829.

Maemura, Emily, Nicholas Worby, Ian Milligan, and Christoph Becker. 2018. ‘If These Crawls Could Talk: Studying and Documenting Web Archives Provenance’. *Journal of the Association for Information Science and Technology* 69 (10): 1223–33. https://doi.org/10.1002/asi.24048.

Milligan, Ian. 2019. ‘Historians’ Archival Research Looks Quite Different in the Digital Age’. *The Conversation* (blog). 19 August 2019. https://theconversation.com/historians-archival-research-looks-quite-different-in-the-digital-age-121096.

Talboom, Leontien, and Mark Bell. 2022. ‘Keeping It under Lock and Keywords: Exploring New Ways to Open up the Web Archives with Notebooks’. *Archival Science* 22 (3): 393–415. https://doi.org/10.1007/s10502-022-09391-6.

Winters, Jane, and Andrew Prescott. 2019. ‘Negotiating the Born-Digital: A Problem of Search’. *Archives and Manuscripts* 47 (3): 391–403.

Library digital collections as data

Wyne Nekesa, Public Procurement and Disposal of Public Assets Authority/IASSIST

According to Smith-Yoshimura (2006) digital collections is one of the types of data published as Linked Data. Libraries publish Linked Data in order to expose their data to larger audiences on the Web and also demonstrate how the users can use library digital collections as Linked Data.Linked Data technologies have supported the availability of information resources on the Web without any retrieval restrictions. However, most of the Linked Data technologies are being used in developed countries, and not in developing countries because of lack of awareness (Bryne, 2015).

Although libraries have been digitizing their materials for years to provide enhanced access of the library collections and meet the needs of the multiple users, Some of these information resources are seldomly used (Warraich and Rorissa, 2018:2). Uganda has over 53 university libraries, however, only 14 have adopted library integrated systems such as DSpace, Koha. Other universities have developed in-house systems which seem to limit access to library collections among the university libraries (Mwesigwa, 2016; Mulumba, 2020). The open software technologies they are using are customized to meet their needs. They prefer open sources software technologies because of they do not require a big budget to implement. Users can freely access library resources without limitation. These academic libraries have sections which oversee the digitization of these library collections, on its done, then they are uploaded onto these systems for wider access.

I conducted a research recently on the same, and the findings showed that indeed the academic libraries in universities store large volumes of library collections that can be shared and reused to support research and innovation. However, these libraries need to create visibility and accessibility of their collections by transforming them into data. Linked Data plays a such a significant role and will help to create web visibility and accessibility of this data. Linked Data technologies in libraries are significant to making their rich data available on the Web (Warraich and Rorissa, 2018)) The librarians also expressed the need for collaboration with other institutions that have implemented such technologies in order to improve digital scholarship. Currently, each university libraries in Uganda, has its own institutional repository that can only be accessed by its users/faculty, yet if there was a central data repository where these university libraries can deposit their collection, this will not only improve access, but it enable sharing and re-use of data among these universities and their users/faculty instead of working in silos. This will improve discoverability of library collections/data. Currently, there is no inter-university library exchange of Linked Data in Uganda, yet it seems this exchange has benefitted other universities in the developed countries. Adoption of Linked Data among the university libraries will enable university libraries to share their library collections/data for further research. so that it can be re-used by other university. low levels of awareness hinder the adoption and implementation of Linked Data in these libraries (Warraich and Rorissa, 2018). There is need strengthen research and innovation through sharing information resources nationally and internationally.

Linked Data also allows for the integration of library data and data from other resources like scientific research, government data, commercial information, or even data that has been crowdsourced. Linked Data increases the sharing of library metadata and other information resources. Therefore, making library collections/ data available for reuse eliminating unnecessary duplication of data that is already available elsewhere, through reliable sources (Borst et al., 2010; Alemu et al., 2012; Gonzale, 2014).Through the Consortium of Uganda university libraries (CUUL) partnership, the libraries in Uganda provide access to a wide range of scholarly e-resources (**Electronic Information for Libraries (EIFL)**, n.d). EIFL has supported the launch of open access repositories in the country (OpenDOAR, n.d). The collaborative efforts of the universities and national policymakers are underway to make all peer-reviewed research output openly available (EIFL, n.d).

References:

Alemu, G., Stevens, B., Ross, P., & Chandler, J. (2012b). Linked Data for libraries. *New Library World*, *113*(11/12), 549–570. https://doi.org/10.1108/03074801211282920

Byrne, G., & Goddard, L. (2010). The strongest link: Libraries and linked data. *D-Lib Magazine*, *16*(11/12), 2.

Mwesigwa, A. (2016). *Exploring online repository software suitability for resource sharing at Makerere University*.

Mulumba, O. (2022). *Library Services and Linked Data at Makerere University : Prospects of a DigitalCommons @ University of Nebraska - Lincoln Library Services and Linked Data at Makerere University : Prospects of a Research-Led University*.

Smith-Yoshimura, K. (2006). *Linked Data Implementations: Who, What and Why. What and Why.*

Warraich, N. F. (2016b). Linked Data Technologies in Libraries: An Appraisal. *Journal of Political Studies*, *23*(697), 697–707. http://pu.edu.pk/images/journal/pols/pdf-files/Nosheen - 22_v23_2_16.pdf

Lowering Barriers to Working with Big Web Archived Data: The Archives Unleashed Project

Ian Milligan, University of Waterloo

"We want researchers to be able to use the archived web in their research."

This has been the central mantra of the "Archives Unleashed" project, dating back to our first datathon in 2015/16, and then to the two Andrew W. Mellon Foundation-funded phases of the project between 2017 and 2023. This goal stems from my work as a contemporary historian, where I have argued that it is hard to imagine a historian being able to do justice to so many research projects from the mid-to-late 1990s onwards *without* drawing on archived websites. Consider: a cultural historian will be able to draw on user generated webpages from sites as varied as GeoCities.com or later social media; a business historian the pages of the dot-com boom and other corporate sites; a political historian, everyday political activity.

The Problem of Working with Web Archives at Scale

It is currently too hard to draw on web archives for research at any level beyond individual replay. If one knows what they want to find, a Wayback Machine is more than enough: a keyword search query (on home pages) or URL entered, coupled with contextual information, and increasingly sophisticated abilities to compare pages. For larger-scale research: textual analysis at scale, complicated keyword searches across all of the content of a web archive, network analysis, or any combinations of the above, the Wayback Machine is not enough. For example, the Wayback Machine works if you want to say look at the evolution of the World Health Organization's homepage or a particular program over a decade or two. But if you wanted to do exploration such as "find all pages on the World Health Organization mentioning 'Canada,' and all of the links that they link out to," you are operating at a level of complexity which requires sophisticated data exploration.

Until relatively recently, the only way a scholar could answer those questions would be to work with the "raw material" of a web archive itself: the WebARChive files, or WARC files. WARC files are an international standard, specified in an ISO, that serve as a container file format for all of the rich material that goes into a web archive. The difficulty is that if a scholar wants to work with WARC files directly, they run into difficulties: (a) they are large, terabyte-scale files; and (b) there's no large user community.

Enter the Archives Unleashed Project

In 2017, the Archives Unleashed Project started with funding from the Andrew W. Mellon Foundation. Led by an interdisciplinary team — a historian (myself), a librarian (Nick Ruest from York University), and a computer scientist (Jimmy Lin from University of Waterloo) – our mission was to "make petabytes of historical internet content accessible to scholars and others interested in researching the recent past." We had three inter-related projects:

- **Archives Unleashed Toolkit**: An open-source platform for analyzing web archives built on Apache Spark, which provides tools for analytics and data processing. It is a command-line based tool.

- **Archives Unleashed Cloud**: We recognized that the Toolkit was not user friendly. Accordingly, the main method of interacting with the Toolkit would be through the "Cloud." This would be a centrally-hosted version of the Toolkit with a GUI on the front end. Users would "sync" their Archive-It collections with the Cloud to analyze them. Instead WARCs, they could generate:

    - **Full-text files**: A CSV of the text that would be displayed on HTML pages.
    - **Hyperlink networks**: An extract of all the domain-to-domain web links, arranged by date.
    - **Domain information**: An overview of what domains were collected.
    - **Archives Unleashed Datathons**: Finally, we wanted to ensure researchers knew how to use our tools! Between 2017 and 2020, we ran four datathons. These datathons provided an opportunity to build a sustainable community around Archives Unleashed tools, scholarly discussion, and training for scholars with limited technical expertise to explore archived web content.

Overall, these three projects would together help lower the barrier to web archival research. We also, of course, were cognizant that our initial grant from the Andrew W. Mellon Foundation was to get this started, but not to last forever. How could we keep it going?

Partnering for Sustainability: The Development of "ARCH"

Throughout our first 2017 - 2020 grant, sustainability was top of mind for the Archives Unleashed team. Our computing infrastructure was supported by Compute Canada, which is accessible only by writing and successfully winning a grant from that organization. We explored setting up a separate non-profit corporation that could accept small payments from organizations to process their web archives, but realized the back-end overhead would not scale. Also, none of us had become researchers to run a small web archiving analysis company (it also turns out that running a business is hard!).

Fortunately, we knew there was a natural partner: the Internet Archive. As the Cloud already worked with Archive-It, we saw an opportunity to partner with the Internet Archive to bring a collection-to analysis ecosystem completely together. Users could create web archives with Archive-It, and then analyze them on the same system. This would also have benefits to reduce data transfer times and having to duplicate the data in two different places. It was the epitome of a natural relationship. Crucially, the Internet Archive and Archive-It brought a community ethos to the relationship, from their Community Webs program to considerable outreach activities.

The Andrew W. Mellon Foundation agreed, and generously funded our project between 2020 and 2023. Currently, as of 2023, we are preparing to publicly launch the successor to the Archives Unleashed Cloud: the Archives Research Compute Hub, or ARCH. ARCH unlocks the research potential of web archival collections with its ability to generate over a dozen datasets (including domain

frequency statistics, hyperlink network graphs, and extracted full-text). Users can select their Archive-It collections for exploration and prompt dataset creation with the push of a button.

From Datathons to Cohort

The final problem, of course, has been user adoption. In general, web archiving suffers from a lack of tools… which is often because there are not enough users… which, we believe, is in part because there are a lack of tools. Our original vision for the Archives Unleashed datathons was to break this cycle by helping to bootstrap researchers over an intensive two-day event. However, we realized that many of our researchers had a rewarding datathon, and then went back home and could not continue their projects.

For the second phase of our project, we launched the **Archives Unleashed Cohorts** program. In each of the 2021-22 and 2022-23 years, five research teams would be sponsored by our group. Each would receive a small monetary award, as well as bi-weekly meetings with our team and access to Internet Archive support. The projects are going exceptionally well, and a full listing of projects can be found at https://archivesunleashed.org/cohorts2021-2022/ and https://archivesunleashed.org/cohorts2022-2023/.

Conclusions

Our project has been working to "make petabytes of historical internet content accessible to scholars and others interested in researching the recent past." Through our collaboration and programs, we think we are getting there. Barriers are lowered through technical development – providing derivative datasets – but also thinking seriously about sustainability and how we can best reach our community.

Ian Milligan is professor of history at the University of Waterloo, where he is also Associate Vice President (Research Oversight & Analysis). He would like to thank his colleagues on the Archives Unleashed team, notably his co-principal investigators Nick Ruest, Jefferson Bailey, and Jimmy Lin, as well as our project manager extraordinaire Samantha Fritz. Thanks also to fantastic Internet Archive colleagues who we have been lucky to have on our team, notably including Thomas Padilla, Helge Holzmann, Kody Willis, Alex Dempsey, and Karl Blumenthal.

Making Collections as Data Tractable: Examples of Iterating on Small Web Archives Datasets

Trevor Owens, Library of Congress

Over the last decade a lot of progress has been made to support a wide range of computational use of cultural heritage collections. The work supported by *Collections as Data: Part to Whole* has played a big part in making progress on this front.

With that noted, it feels like we are still in an early state of clarifying and establishing the operational patterns that will broadly support use of cultural heritage collections as computational data for scholarship. What kinds of datasets do we need to have so that researchers can dip their toes in and try out new research? Does everyone need to be establishing big collaborations to do this kind of work? What kind of support is reasonable to imagine comes from librarians or other experts to enable computational research from collections? We have a lot of good work suggesting potential answers to these questions, but it feels like all of these are still, to a large extent, open. It's also my sense that these kinds of questions are some of the biggest ones facing us to make collections as data kind of work more mainstream.

In my position statement, I run through some of my own work with some small datasets extracted from much larger sets of web archive data to illustrate how smaller datasets, roles for collections as data librarians, and support for new kinds of collaborations can all work together to help get us closer to a world where computational research with collections becomes more common and routine.

Roles for smaller/easier to use test sets of data

One area that I think can be of general use as a model is creating smaller test sets of larger corpora that can be used to test and try out various research ideas. As an example, the Library of Congress Web Archiving Team and LC Labs teams have made available a number of smaller sample sets of data from the petabyte scale data of the LC Web Archives.

Most of those datasets are small enough to be able to do meaningful work with them on a commercial laptop. With 1,000 rows, it's possible to try out ideas with manual coding of data. Along with tabular data about the files, the sets of the files themselves which were extracted from government websites are that small enough that it's possible to work with them at a manual scale. Because they are random samples, they also come with the value of being representative of the broader set of material they come from.

As examples of the kind of research that can be accomplished with these kinds of resources, Slide decks as government publications: exploring two decades of PowerPoint files archived from US government websites which was recently published in the journal *Archival Science*, uses one of those datasets as a way to conduct novel research without really using any particularly sophisticated computational

methods.[24] Close reading of individual files alongside basic analysis on derived metadata was sufficient to support new basic research into studying changing publication practices of the U.S. government. This kind of work has the potential to scale up further with bigger corpora of data, but it's still viable and useful at that small scale.

Roles for Collections as Data Librarians/Digital Scholarship Librarians

From discussions with a wide range of researchers interested in computational use of collections, the skills necessary to process data into a format that is usable and then actually do computational analysis is broadly lacking. There are a small number of researchers who have those kinds of skills, and while that skill base is growing for some of the next generation of researchers it is still going to be a relatively small portion of humanities and social science scholars that have the computational research chops to be able to work with command line tools and run the kinds of processes and jobs necessary to do computational collections as data research. Having smaller datasets can help lower the bar, but it's worth noting that it takes a lot of work to winnow down some huge set of data into something tractable and potentially useful for that kind of work. In the case of those smaller web archives datasets, that was all only possible because of the work that expert staff on the Library of Congress Web Archiving Team did to produce each of those smaller datasets.

I imagine we are likely to see more and more tools develop that can lower those barriers to entry, but it also seems increasingly likely that for this kind of work to really take off we also likely need to support the creation of more roles, like "Collections as Data Librarians," "Digital Scholarship Librarians" or "Data Services Librarians" who as part of their work engage in processing collections into usable data in support of specific research requests and in general as resources that could invite different kinds of use for scholars. The derivative web archive datasets from the LC web archives are a mixture of these kinds of resources. Some developed directly in response to individual researcher requests but then shared more broadly and some developed intentionally to support a range of potential use cases.

Roles for larger teams and collaborations

Along with thinking about how to support work of individual researchers and scholars, it seems increasingly clear that a lot of the future of collections as data research and scholarship is likely to become something that requires larger teams and collaborations.

Significantly, I think the previous two points in this piece also fit with this third one. That is, if we make more and more smaller sets of data available for testing out ideas and doing initial exploratory work, and if we create more roles for digital scholarship librarians and or collections and data librarians who can help to support some requests from researchers and also produce more of the datasets that are user

[24] Owens, Trevor, and Jonah Estess. "Slide Decks as Government Publications: Exploring Two Decades of PowerPoint Files Archived from US Government Websites." *Archival Science*, November 21, 2022. https://doi.org/10.1007/s10502-022-09406-2.

friendly, then we will have a growing body of work that can be considered and reviewed for potential scaling up.

As an example of how some of that can work, again drawing from the web archive datasets, without any specific dedicated funding or resourcing, Ben Lee and I engaged in some research using a test set of 1,000 PDFs from government websites to produce Grappling with the scale of born-digital government publications: Toward pipelines for processing and searching millions of PDFs, which was published in the International Journal of the Digital Humanities.[25]

By focusing our project on a representative but small data set like this, we were able to produce useful and publishable results, but we also were able to demonstrate a proof of concept that with further funding and resourcing could likely scale from 1,000 files up to corpuses of hundreds of millions of files. Importantly, by starting with the 1,000-file set, it's possible to do the work that really demonstrates that there is reason to scale this all up further.

[25] Lee, Benjamin Charles Germain, and Trevor Owens. "Grappling with the Scale of Born-Digital Government Publications: Toward Pipelines for Processing and Searching Millions of PDFs." *International Journal of Digital Humanities* 3, no. 1 (April 1, 2022): 91–114. https://doi.org/10.1007/s42803-022-00042-x.

Masters of the universe

Duncan Loxton, University of Technology Sydney

When I first encountered the phrase *always, already computational* I was immediately seduced by its call to possibility. Not only could we develop our materials and interfaces to do cool things with computers, it was already something we could act upon. Best of all it didn't necessarily require recoding. We could better ourselves and our services by reorienting our thinking: our collections had always been in a state of becoming computational. I felt an intoxicating sense of optimism and abundance. It felt empowering. I was reminded of Chris Chesher's declaration that "Computers should not be called computers, but invocators". We could bridge the technical and the magical.

I couldn't help but start imagining the affordances of the computational in the collections I worked with. *Always, already computational* offered itself as both a premise and a conclusion to my work as a data curator. I shared Tim Sherratt's desire for more data, whatever its quality or scale. I felt voracious. When Laila Shereen Sakr ventured that our speculation about materials and interface design is analogous to worldbuilding, my power to shape collections as data felt decisively and creatively potent. My thinking had changed now that I'd started thinking with a different type of media.

I currently work with a small number of qualitative research datasets at the Aboriginal and Torres Strait Islander Data Archive at the University of Technology Sydney. Access to most of these datasets is carefully mediated so that we are able to support the rights of Indigenous peoples to maintain, control, protect and develop their knowledge in the data. Work with this data nearly always involves the community of interest. I wondered what worldbuilding I could offer if it was requested.

I read with interest Hart Cohen's affirming reflections about how digital media and digital technologies were helping First Nations peoples resume relationships with records in the Strehlow Archive in Australia. Digital media, he argues, "appears to afford the promise of furthering the goal of Aboriginal knowing while at the same time challenging the orthodoxies of research practices". He describes databases being used to support and mobilise narratives and surface a multiplicity of voices and representations of archival records. The database can affect, he writes, and "will inevitably alter the experiential qualities of how the reader enters into the spaces of the story and, in this case, conceives of and sings up the scenes of landscape and country". He acknowledges that computer-mediated technologies such as databases are limited in what they can represent and the invocations they afford us. It's an important reminder that our worldbuilding powers are constrained and loaded with different cultural contexts and paradigms. There are only so many narratives and descriptive traditions that the computational can support; we cannot encode the totality of a world.

I occasionally encounter the idea that records can speak for themselves, that there's 'no wrong door' of encounter wherever you delve. This idea assumes that whatever the record represents or references materially should be immediately apparent. That we need not consider, inform or mediate the context in which something is read. That the mechanism of encounter and interpretation isn't circumscribed or political or hegemonic or subversive or deceitful. As a counter to this idea, I take Grant Bollmer and Katherine Guinness' view that data can be theorised as a trope or figure of speech, namely metonymy

and synecdoche. Their construction reinforces the idea that data is rhetorical in nature and that it can be incomplete or referent or stand in for somethings else. Conceiving of data as a literary device also sustains the acute feeling of empowerment I felt when first learning of *Always Already Computational: Collections as Data*. We can actively shape and participate in these worlds and should invite our communities to do the same.

Works Cited

Bailey, Jefferson, Chassanoff, Alexandra, Clement, Tanya, Foreman, Gabrielle P., Mookerjee, Labanya, Fowler, Dan, Green, Harriett, Guiliano, Jennifer, Hardesty, Juliet L., Harlow, Christina, Jansen, Greg, Marciano, Richard, Johnston, Lisa, Lincoln, Matthew, Liu, Alan, Miller, Matthew, Neatrour, Anna, Posner, Miriam, Rabun, Sheila, … Zwaard, Kate. (2019). Position Statements --- Always Already Computational: Collections as Data. Zenodo. https://doi.org/10.5281/zenodo.3066161

Bollmer, G. (2019). *Materialist Media Theory: An Introduction*. London: Bloomsbury Academic. https://doi.org/10.5040/9781501337086

Bollmer, G. and Guinness, K. (2018). 'Do You Really Want to Live Forever?': Animism, Death, and the Trouble of Digital Images. *Cultural Studies Review*, 24:2, 79-96. https://doi.org/10.5130/csr.v24i2.5995

Chesher, C. (2003). Layers of Code, Layers of Subjectivity. *Culture Machine, vol. 5*, http://culturemachine.net/index.php/cm/article/view/255/238.

Cohen, H. (2017). *The Strehlow Archive: Explorations in Old and New Media* (1st ed.). Routledge. https://doi.org/10.4324/9781315145464

Is Minimal Computing the answer? Or how can developing countries catch up with Collections as Data?

Paul Jason Perez, University of the Philippines Diliman

In small Zoom breakout rooms, I asked my students to examine the websites of different libraries, archives, and museums. It is early in 2023, and we are still conducting some of our classes via Zoom. We discussed the features of a digital library as defined by Calhoun (2014, p. 18). The task was for them to look at the digital objects, the metadata, and the overall platforms. One by one, we examined the DPLA, Europeana, Trove, and the Internet Archives as examples of excellent digital cultural institutions. Afterward, we looked at the digital collections[26] of the National Library of the Philippines, unfortunately, as a counter-example, and our discussions moved to what was lacking or what could be improved. This is a familiar scenario I have whenever I teach digital libraries in my class in the Philippines.

We have a long history of cultural institutions in my country. Our national library, museum, and archives trace back its roots in the late 1800s during the Spanish colonization. One of the earliest recorded formal training in library science in the Southeast Asian region happened here in 1917, conducted by Mary Polk and other American librarians. The Philippine Librarians Association, Inc., just celebrated its 100th anniversary, a professional organization older than IFLA. With such long history and tradition, I often wonder why we, as a country, never had "*decades of digital collection practice*" from which we can "*build upon"* (Padilla et al., 2019, p. 7).

This is a question that I, along with two of my undergraduate students, tried to answer in 2021. In particular, we looked at the state (or lack thereof) of open data in Philippine Museums (Perez et al., 2022). Given that many museums closed their doors during the pandemic and transitioned online, we wanted to know to what extent the museum collections are accessible using Roued-Cunliffe's (2020) Open Heritage Data framework. Out of the 85 museum websites we examined, only one, an elite, private university, exhibited the feature of having its collection searchable through its metadata—a feature that only satisfies one of the five qualities of open heritage data.

We interviewed eleven private, public, and volunteer-led museums to ask about their general perceptions of open data. While we, as the research team, talk about catalogs, most museum staff refer to their inventory—which they only use internally and not for the public. There are also questions on copyright and how people will use their digitized collections should they make them available publicly. Overall, the shortage of funding and personnel remains the top reason why Philippine museums could not make their collections as data available, especially during the pandemic, when physical access is limited or non-existent.

As GLAM institutions in developed countries continue to enhance and make available their digital collections, we need to ask how developing countries, having issues with their digital infrastructures, can do the same.

[26] TechnoAklatan (https://nlpdl.nlp.gov.ph/TechnoAklatan.htm) is the digital collections of the National Library of the Philippines, which appears to be just digitized pages without browsable metadata.

As COVID-19 pushed many conferences and workshops online and mostly free, I chanced upon CollectionBuilder (Wikle et al., 2021) at the Digital Humanities Summer Institute in 2021. It was also the first time I heard about the minimal computing paradigm stemming from the digital humanities. CollectionBuilder uses open-source web development languages. It can be deployed on the web for free through GitHub, albeit limited. Furthermore, perhaps more importantly, it can get organizations off the ground even if all they have is a spreadsheet of their collections, making their metadata available in open formats.

I used this tool in the projects I was involved with directly or as a consultant. We were able to use the minimal technology approach for developing the Philippine Commission of Human Rights Digital Archive. In one of my classes, a group of students transformed a Google Drive collection of banned or threatened books in the Philippines into a digital library with the help of CollectionBuilder. Lastly, in a workshop I conducted for a volunteer group of artists fighting against the War on Drugs in the Philippines, participants were able to think more about the metadata of their collections. So far, CollectionBuilder and the minimal computing approach seem to be serving as my hammer.

However, this minimal computing approach is not a silver bullet. "*Collections as data is an ongoing process* (Padilla et al., 2019, p. 4)," and given the labor conditions of cultural heritage workers in the Philippines, serious questions on staffing and the sustainability of the technology infrastructure still need to be addressed. For example, many of the museum workers we interviewed are non-permanent staff. Some volunteer-led organizations do not even have a website or server and use social media to "store" their digital collections.

There are many arguments for why cultural heritage data should be available and ready for computation. However, as institutions from developing countries grapple with their material conditions, we must develop more ways to get organizations lacking resources on board and make the available data on cultural heritage more diverse.

Works Cited:

Calhoun, K. (2014). Exploring digital libraries: foundations, practice, prospects. Facet Publishing.

Padilla, T., Allen, L., Frost, H., Potvin, S., Russey Roke, E., & Varner, S. (2019). Always already computational: collections as data. https://doi.org/10.5281/zenodo.3152935

Padilla, T., Allen, L., Varner, S., Potvin, S., Roke, E. R., & Frost, H. (2018). Santa Barbara Statement on Collections as Data. Always Already Computational Collections as Data. https://doi.org/10.5281/zenodo.3066209

Perez, P. J., Escueta, J. K., & Jequinto, P. M. P. (2022). Digital Cultural Heritage Inequality: Philippine Museums During the COVID-19 Pandemic. Collections, 18(4), 568–585. https://doi.org/10.1177/15501906221121624

Roued-Cunliffe, H. (2020). Open Heritage Data: An introduction to research, publishing and programming with open data in the heritage sector. Facet Publishing.

Wikle, O., Williamson, E., & Becker, D. (2021). “Creating Digital Collections with Minimal Infrastructure: Hands On With CollectionBuilder for Teaching and Exhibits”, DHSI 2021, https://osf.io/5q976/.

FAIR GLAM: Information and data in museums

Saskia Scheltjens, Rijksmuseum

Information and data are crucial to a museum's mission, connecting people with art and history or science. Managing and organizing the different types of information and data enables the institution to preserve its collections, engage with the public, fulfill its educational mission and conduct or enable research.

The responsibility for information and data management in museums is sometimes centralised [1], but usually it is spread across multiple departments and teams, depending on the size and structure of a museum. It is essential that there is a coordinated and strategic approach to managing this important museum asset.

However, in much of the recent literature and research projects, the breath and specifics of museum information and data is not taken into account as a whole. This hinders future development and engagement with the broader public, makers and researchers.

Back to basics

It looks like stating the obvious, but it is important to recognize the difference between data and information and their interrelation within the DIKW[2] model, also in museums. Data refers to raw facts or figures that are collected and processed, while information is the result of analyzing and contextualizing data. In a networked world, data without information is meaningless, and information without data is useless. It is crucial that they have to be treated in relation to each other. Even more so since recent literature and research have highlighted the ethical challenges of data and information bias in museums as well.

Moreover, information and data in museum does also not only involve the collection. The wide range of information and data in museums includes object-related, interpretive, conservation-related, bibliographical, archival, research-related, marketing and communication-related, exhibition-related, visitor-related, administrative, operational, and financial information and data. Managing all this in relation to each other will enrich how a museum can engage with the world. And the world with a museum.

This requires a mature overall information and data strategy, a pragmatic information architecture and a robust digital infrastructure, preferably using open standards and - systems by a dedicated team of information & IT specialists. All too often, that is not yet the case.

OPEN GLAM

The Cooper-Hewitt Design Museum in New York was the first museum to issue an open data policy in 2010, followed shortly after by the Brooklyn Museum in 2011. They were the earliest examples of museums making their collection data openly available for public use and reuse. The first museums to issue an open data policy in Europe were the British Museum (CC-BY - 2011) and the Rijksmuseum in Amsterdam (CC0 - 2012). There was also the rise of (public) digital platforms like Wikimedia and Flickr (early 2000s), Europeana (2008) and the Digital Public Library of America (2013) to provide online access to millions of items from libraries, archives and museums with the aim to promote wider access to and use of freely available digital resources for education, research, and public engagement. The unsurpassed Open GLAM survey [3] by Douglas McCarthy and Andrea Wallace lists as of now 1.608 institutions who offer (some of their) collections with an 'open access scope'.

The adoption of open data policies marked a significant shift towards greater openness, transparency, and collaboration in the cultural heritage sector. But while all this supported the rise of new disciplines and fields of expertise (digital humanities, cultural network analysis, cultural AI, UX design, ..), it also gave rise to a conflation of concepts related to all those efforts that by now threatens to hinder further innovation and collaboration.

The discourse about OPEN GLAM en 'Open Access'[4] in cultural heritage institutions, especially museums, in the literature of the last decade, almost always referred to the practice of making digital images (more specifically 'digital surrogates of public domain objects' [5]) and metadata about those images and objects from an institution's collections freely and openly available to the public. This might have been what was available a decade ago, but by now the field has evolved and requires more nuance [6].

In museums, next to object collection data, there is a wealth of other kinds of information and data available which up till now remains mostly invisible. The degree of openness of research data, conservation data, marketing and communication data, the scholarly output of museums curators and conservation scientists, etc… etc.. are not (yet) listed in surveys, nor are they very often used in research. And they don't do well with a binary approach of open or closed.

My point is that sharing this kind of information and data requires more nuanced and culturally and technologically advanced licences, policies and infrastructures than is now being used in the (art & history) museum world.

Towards FAIR GLAM

The FAIR data principles were developed in 2015-2016 [7] under the eagis of the European Commission to make research data more findable, accessible, interoperable and reusable. They have become an important part of the Open Science movement, which aims to make research more transparent, collaborative, and accessible. Applying the FAIR data principles to museums can help enhance the impact and value of museum collections and research, ultimately benefiting both the museum community and society as a whole.

I propose to move from an Open GLAM discourse to a FAIR GLAM discourse. It allows museums and other cultural heritage institutions where this is not yet a discussion, to build on what has been developed in the past decade, all the while align more with research and the current technological, political and ethical challenges we are faced with now.

And last but not least, to allow for conversations in museums about the need for organisational changes regarding the responsability for integrated information and data management and unglamorous, more robust digital infrastructures and open digital standards to make FAIR GLAM possible.

Works cited

[1] As is the case in the Rijksmuseum's Research Services Department (2016 -)
[2] https://en.wikipedia.org/wiki/DIKW_pyramid
[3] https://douglasmccarthy.com/projects/open-glam-survey/
[4] Used as an overarching term of openness, for example "Implementing Open Access in GLAMs - Open GLAM Medium" by Anne Young. February 17, 2020
https://medium.com/open-glam/implementing-open-access-in-glams-3bda60941188
[5] idem
[6] For example, the Rijksmuseum currently has an integrated Digital Strategy (not yet public), an overall Information Plan, a renewed Collection Information and Data Architecture, an Open Access policy for the scholarly output of its staff and own Rijksmuseum Bulletin, and an Open Data policy for the collection data it openly shares via a CC0 licence. A Research Data Management policy and a FAIR Data policy, the latter of which will replace the current Open Data policy, will be published by the end of 2023.
[7] Wilkinson, M. D., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3 https://doi.org/10.1038/sdata.2016.18

No Size Fits all: Publishing Collections as Data

Steven Claeyssens, KB, National Library of the Netherlands

In considering how to develop digital cultural heritage collections to support computationally-driven research and teaching, we must acknowledge that cultural heritage institutions (CHIs) are increasingly acting as *publishers* of their digital collections, both born digital and digitised. Just like conventional publishers, CHIs are not the creators of a work, but they do however make the decision to make a digital work *public*. They take responsibility for much more than simply making a scan and putting it online. They oversee the entire process of selection, production and dissemination, just as any other publisher does. Recognizing this is important. It makes it clear that, as publishers, CHIs make choices about which objects they want to make available, to which audiences, in which ways, and, not unimportantly, to which end. If we make these decisions more thoughtfully and less implicit, it will gain us a better understanding of the complexity of the Collections as Data mission, and will, in time, result in better services to facilitate computational use of cultural heritage.

Publishing as an activity has significant scholarly attention. Possibly the most versatile anchoring theory of publishing has been formulated by Michael Bhaskar in *The Content Machine*.[27] He proposes a metaphoric vocabulary to understand publishing as a content machine that fills specific *frames* according to certain *models*. Simplified, the *frame* determines the shape and size of a publication and through which channels it is distributed, the *model* provides the '*raison d'être*', the 'goals, motivations and ideological underpinnings.' For example: an *Introduction to AI* is produced as a paperback and sold through mainstream bookstores (*framing*) with the intention of disseminating knowledge to a wider audience and making money (*modelling*). The concepts explain how (*framing*) and why (*modelling*) something is published in a certain way.

In line with this way of thinking, the classic online graphical user interface searching digital collections of cultural heritage is a *frame*, a specific way of making digital objects discoverable by people. Underlying the frame is a *model*. For Delpher, the Dutch national platform publishing digitised books, newspapers and periodicals free of charge on the web, the *model* is related to Open Access principles: making printed material available for the widest possible human audience as open as possible. Facilitating data-level access to entire collections of digital material is something else entirely. The *model* might be pretty straightforward – the intention to enable and foster computationally driven use of the collections – but the *framing* turns out to have many faces. Data Services at the KB, the national library of the Netherlands, for instance, consists of documented API's, ZIP files, instructional notebooks, and ad hoc assistance by a librarian.

Over the years we accumulated quite some expertise in publishing digitised printed material. We have to draw the conclusion: there is no one-size-fits-all *frame* for publishing collections as data. What other

[27] Bhaskar, Michael. *The Content Machine: Towards a Theory of Publishing from the Printing Press to the Digital Network.* London/New York/Dehli: Anthem Press, 2013.

insights have we gained? And especially, are there insights that might assist us in publishing collections as data?

The content

1. A digital cultural heritage object is never a surrogate. Once published, it's a publication as much as the so called 'original' or 'master'.
2. We publish *objects* (books, newspapers, etc.) as well as *collections* (of books, newspapers, etc.). Both are publication efforts in their own right. Both require associated metadata and/or documentation.

The model

3. We want to publish as open as possible and as closed as necessary. (Although there are compelling reasons to re-evaluate this and make an exception for Big Tech.)
4. We want to publish as FAIR as possible, i.e. Findable, Accessible, Interoperable and Reusable, for both humans and machines.
5. We want to publish for the widest possible audience. Everyone can be a researcher or a teacher, whether they received any formal training or not, whether they are able to code or not.

The Frame

6. An API is not enough. A human-friendly front end is also required.
7. The main focus should be publishing to facilitate research, not research itself.

With all of this in mind the KB is reforming its efforts as a publisher of collections as data, with a sharper emphasis on usability and accessibility. The remake puts a stronger emphasis on the FAIR principles and documenting data provenance. It includes the introduction of a data registry to make the collections more easily findable for both humans and machines, and a series of data sheets and/or data cards as standardised documentation, encouraging transparency and potential bias identification. The KB also aims to build a corpus selection interface, offering advanced functionalities of data discovery and selection to support the creation of research corpora, as such functioning as a more intuitive user interface to the existing API's.[28] Finally we are addressing the need for giving access to in-copyright and in commerce materials by creating an onsite mining facility and an online tools-to-data solution, providing ways to mine data without violating the rights of copyright owners.

[28] Kemman, Max, Nick Jelicic, Guido de Moor, Marenne Massop, and Tommy van der Vorst. *User Needs for a Text Suite for Advanced Digital Research*. Utrecht, 2022. https://doi.org/10.5281/zenodo.6591572.

On data interpretation

Kim Pham, Max Planck Institute for the History of Science

It is heartening to read the reports from Part to Whole as it demonstrates the diverse and evolving perspectives and practices needed to create Collections as Data. Some immediate future directions come to mind, such as what possibilities are there for cross organisational collaboration and bridging infrastructures? A common response from cultural heritage institutions trends towards imposing standards and structure onto data for the sake of interoperability and scalability, reducing a messy reality only to invite it to be blown up again for reinterpretation. Which makes me consider another future direction: how do we approach the interpretation of collections as data?

"*To collect photographs is to collect the world*," writes Susan Sontag in 1977, a time when amateur photography became popularised, and commercial photography was on the cusp of becoming digital. The proliferation of data seems to follow a similar industrial arc as the proliferation of the image. Its mass production and consumption passes for an approximation of the world.

Working in digital humanities, my experience is that the historian's interpretative framework desires a resistance to codification. Data is as much a historical object of study as it is evidence. Data, like the image, too can range from the "*objective*" to the "*psychological science fiction*". Historical meaning never wholly relies on data, because data taken at face value is never what it purports itself to be. Meaning is created through a synthesis of evidence and different orders of thought drawn from patterns in history and derived contexts - things that aren't so explicit. Theories of intention, motive, and constructed attention are considered - a solemn acknowledgement that knows not to pass data off as truth, ironically bringing in more trust.

Why not another passage from Sontag, where we again find a kindred spirit to the historian's perspective on data:

"Photographs, which cannot themselves explain anything, are inexhaustible invitations to deduction, speculation, and fantasy. Photography implies that we know about the world if we accept it as the camera records it. But this is the opposite of understanding, which starts from not accepting the world as it looks. All possibility of understanding is rooted in the ability to say no. Strictly speaking, one never understands anything from a photograph."

I hope these models and workflows from Collections as Data opens the discourse on hermeneutics. Can we anticipate how data will be used, accounting for differing approaches to interpretation? Do we also come up with models for use and interpretation? I invite you to poke fun at my off the cuff description of a data-wine interpretation model (dreamt on a Friday night ��). It asks the data producer to consider the following: Do we hold some responsibility for how the data we put out there gets used, or how it is interacted with, and does responsible use have a minimum training and experience? Are there complementary pairings of data and research and methods to recommend?

Can we make sure data matures well or do we know it to have an expiration date? Does it inherently communicate certain notes, or is there a way we can already speculate its interpretation?

Consulted works

- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?��. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610-623).
- El-Hajj, H., Zamani, M., Büttner, J., Martinetz, J., Eberle, O., Shlomi, N., ... & Valleriani, M. (2022). An ever-expanding humanities knowledge graph: The sphaera corpus at the intersection of humanities, data management, and machine learning. *Datenbank-Spektrum*, *22*(2), 153-162.
- El-Hajj et al. (in review) The Historian XAI: Explainability and Transparency in the Realm of Digital Humanities, *International Journal of Digital Humanities (IJDH).*
- Siebold, A., & Valleriani, M. (2022). Digital perspectives in history. *Histories*, *2*(2), 170-177. - Sontag, S. (1977). In Plato's cave. *On photography*, *3*.
- Sontag, S. (1966). *Against Interpretation and Other Essays.*

Moving Collections as Data from a Patchwork of Projects to a Sustainable Program

Elizabeth Russey Roke, Emory University

Over the past several years, the collections as data initiative has encouraged the GLAM (Galleries, Libraries, Archives, and Museums) community to reimagine how we collect, create, and manage digital collections. From the initial Institute for Museum and Library Services grant that defined what computational data looks like and outlined use cases, to the most recent Mellon grant that put these principles in action, the movement towards collections as data has been a gradual process of learning to integrate data science principles into librarianship. Formerly, access to a digital surrogate in most digital repositories was the norm (what I have been known to call 'expensive photocopying') without much consideration of what digital access might facilitate beyond remote access. Thanks to collections as data, the creation of machine-readable data in tandem with human-readable digital resources has become more common in GLAM-supported digital repositories, expanding the potential of digital resources to reveal patterns and trends that would not be evident using a single page-view focused approach.

The next challenge is to expand the tools, workflows, and vision of collections as data beyond the traditional research data use case. Particularly in this age of rapid development of Artificial Intelligence-based technology that depends upon large data sets, providing collections as data should be a core expectation for libraries. Library-based collections as data could potentially feed applications both within and beyond the library and research setting. For instance, collections as data make libraries more accessible by supplying the machine-readable data needed for assistive tooling. Collections as data can dramatically improve search algorithms and facilitate precision searching for better discovery. Collections as data can support the development of large language models that automate library tasks such as identification of photographs and automated metadata creation. Collections as data can facilitate and encourage data sharing and reuse between research projects.

From Project to Program

Operationalizing collections as data requires not just best practices and metadata schemas, but also a defined custodial model for this type of material. What is the role of libraries in creating and maintaining collections as data? In many ways, collections as data are inherently archival. They are unique, irreplaceable resources that require long-term preservation as well as documentation of the provenance and context around their original creation and use. They are meant to be consumed and understood in aggregate, where the sum of the whole is just as important as description of a single data point or document. Defining collections as data as archival enables libraries to extend well-established archival principles and practices, particularly recent efforts to center communities the records/data set documents. Born-digital archival communities, which struggle with these same questions around how to preserve and make digital data accessible, could be a generative partnership for the collections as data

community. Post-custodial archival frameworks might also offer a model for developing ethical donor partnerships around the creation and stewardship of collections as data.[29]

In addition to broadening our understanding of custodial models for collections as data, collective efforts to develop scalable technological solutions could make a significant impact towards universal adoption of collections as data principles. Currently, collections as data are published and made accessible in a variety of ways, including downloadable bulk extracts, individual file downloads, and API access. Most library systems, both commercial and open source, do not support collections as data capabilities, meaning that libraries choosing to enact collections as data approaches must do so on their own through boutique and homegrown projects requiring maintenance over time with local resources. Software products, both open source and vendor-driven, must supply more out-of-the-box solutions for collections as data that build upon existing systems and software. The community should also be thinking about impacts of the shift to cloud computing where storage is cheap, but regular computational access to that data can be costly. At my own institution, for instance, regular fixity checks on our AWS-based preservation repository require additional financial resources to support. Cost analysis and comparisons of the options for providing collections as data will result in more scalable solutions and result in a more even and predictable user experience for researchers. This also means developing defined, long-term commitments to maintain and update the resource over time, ensuring reliable long-term access to the original artifact.[30]

Finally, we must find ways for collections as data to be more discoverable alongside other library collections. Currently, researchers need to search specific discovery systems designed for data sets or visit local repository web portals to find and download collections as data. In some ways, collections as data are almost a "hidden" resource requiring expertise to find and use. This results in encapsulated collections that are hard for researchers to situate in cross-institutional contexts and privileges one single authoritative context and use case over others. One path forward might be exploration of a linked data metadata framework for collections as data, both to describe and to structure the data set itself. A primary advantage would be a natural alignment with Google's Dataset Search, which relies on Schema.org, a linked data-based metadata model. As the place where most researchers begin their search, finding ways to make all library-based collections as data more visible in Google is crucial. Linked data would also potentially put collections as data into conversation with other resources beyond data sets, as is being explored in systems such as Yale's Lux discovery platform.[31] These new linked data-based systems enable resources to be simultaneously discoverable within multiple contexts and communities, creating new paths of discovery and inquiry.[32]

---

[29] An example of post-custodial approaches in practice is at the University of Texas: Alpert-Adams, Hannah, David A. Bliss, and Itza Carbajal, "Post-Custodial Archiving for the Collective Good," *Journal of Critical Library and Information Studies,* vol. 2, no. 1 (2019). doi: https://doi.org/10.24242/jclis.v2i1.87

[30] The British Museum offers a good example of why service models around collections as data need to be carefully defined for long-term stability and availability. The SPARQL endpoint to their resources was suddenly removed in 2022, breaking applications built on the service. https://twitter.com/bm_lod_status

[31] See: https://lux.collections.yale.edu/

[32] See "The Power of Parallel Description: Wikidata and Archival Discovery," in Mark A. Matienzo and Dinah Handel, eds. 2021. *The Lighting the Way Handbook: Case Studies, Guidelines, and Emergent Futures for Archival Discovery and Delivery.* Stanford, CA: Stanford University Libraries. https://doi.org/10.25740/gg453cv6438.

In a 21st century library, the data-based demands of researchers and the tools they use to access our collections depend upon a programmatic approach to collections as data, with specified resourcing, processes, and staff, such as a program manager for data-based collections to coordinate cross-library activities. The next steps for collections as data must be to broaden the vision of collections as data as a core program of libraries rather than a special initiative to meet a specific research need. This will empower more libraries to enact collections as data as a strategic goal and put into place workflows and staffing to support the development and preservation of these collections.

With thanks to Emily Porter and Jennifer Doty for providing input into this statement.

Collections as Data Position Statement

Dr. Kari L. Jordan, The Carpentries

Many institutions offer data skills training through semester-long courses, short courses, and other formats. However, discoverability remains challenging because the curriculum spreads across universities and instructors. Additionally, some universities need more qualified instructors for these topics, so it is difficult to integrate them into the curriculum and even more challenging for instructors to develop their training materials. While semester-long courses integrate well into training for undergraduates and often graduate students, they are usually not the best option for researchers who need training that fits their limited time and is available more frequently. A key challenge in this space is finding high-quality, relevant training matching researchers' time and resource limitations. It may seem that some training is better than none; however, training delivered poorly with no cultural appreciation, at inappropriate levels, or not relevant can demotivate researchers. That lack of motivation limits their ability and interest in data-driven discovery rather than empowering them to use data effectively in their research.

Centering libraries as data skills training providers, not only for quantitative data but collections as data, is a strategy that should be the norm, not the exception. Library effort allocated to collections as data implementation has yielded interest in and discoverability of historical artifacts, stories, imagery, and cultural practices on the verge of extinction, as evidenced by Collections as Data: Part to Whole and its 12 institutional regrantees.

While reading the regrantee's final reports, two questions piqued my interest:

1. What aspects of implementation models are most sustainable to support broad adoption and participation in collections as data?
2. What impact does politics (e.g., the Stop Woke Act) have on maintaining historical collections, and what role can libraries play in that preservation? Students from specific cultural and socioeconomic backgrounds have a strong advantage in learning from collections as data. In underserved communities, inequities directly affect teaching. Research, textbooks, case studies, and data examples often exclude the interests of underserved groups and geographies.

By prioritizing collections as data, one could dismantle the many barriers to accessing the data that influence decision-making, particularly decisions that impact the everyday lives of people of color. By preserving collections as data and preparing learners to use open-source tools and datasets, we will empower a future generation of social scientists of color. Further, partnering with leading inclusive communities teaching data and coding skills brings extensive experience in inclusive education to the libraries and catalyzes the growth of regional communities of practice.

What perspectives are missing? How about a discussion on gatekeeping? Many artifacts constituting tangible or verbal collections are in closed communities. Are there inequities related to

incentives for making collections available to research institutions? How is trust built with cultural heritage organizations to ensure prioritization of the CARE Principles? And what about collections in languages other than English?

It takes a concerted, collaborative effort among diverse stakeholders to implement collections as data. If accomplished, there is no end to the scholarly possibilities.

Collections as Data Position Statement

Martin Tsang, Independent Scholar

With scant few noteworthy exceptions, data, digital objects, and relating to Cuba, or the wider Caribbean are sorely missing within the actionable framework of repositories and institutions in the US. To a sizable degree, the data available and becoming computational fits within narrowly defined topics and disciplines, such as dominant periodicals, historiographical collections of data, or political events such as the Cuban Revolution and Cuban-US political relations. The reasons for such a state are well-known and unsurprising: cross-border embargoes with Cuba, lack of investment in computational equipment and expertise in the Caribbean as a whole, and the pursuit of Euro-American forms of knowledge gathering and dissemination.

Moreover, in such lacking scenarios, the histories, experiences, and presence of diverse and marginalized peoples and communities are summarily further invisibilized by their absence in a two-fold problematic scenario of 1. lack of awareness and inadequate foci on collecting, and 2., the paucity of opportunities to present related material digitally by and for members of the community.

I have gathered from direct experiences of working in the Cuban Heritage Collection at the University of Miami, the largest publicly available repository of materials available on Cuba and Cuba in the world that data are primarily analog with a sizable lag in opportunities to make parts of it readily computational. A methodical approach to digitizing Cuban newspapers and Cuban books is ongoing at the University of Miami. In 2017, the University of Miami team which included Dr. Paige Morgan completed the excellent project to digitize almost 50 years of the nineteenth-century Cuban newspaper, *La Gaceta de La Habana*.[33] Through this initiative 50,000 pages of OCR'd text was made available via the university's server and GitHub for researchers who might be interested in using text mining to explore the collection. Both on the island and in the US, researchers who gather, consult, and work with data connected to marginalized communities are going about these ventures in micro and independent means. On the island where access to computers and related data technology as well as the internet are creating and sharing records through hybrid networks and relay chains of uses and creators that includes USB drives, smart phones, the borrowing of equipment intended for other projects. Included in these networks are scholars and community members who can physically transport memory storage devices and equipment between Cuba and the world. In this manner, smaller pockets or packets of information that touch on a variety of silenced and discarded histories become available digitally. The desire of Afro-Cuban and Chinese-Cuban cultural memory keepers to document and share their data is inspiring and their successes are laudable considering the hurdles that must be navigated to do so. From my vantage point in South Florida, I am keenly aware that these clandestine modes of creating a hybrid and rhizomatic data landscape involves the entirety of the Caribbean and that one of our challenges ahead of us is how to acknowledge and invest our time, knowledge, and technology in centering these stakeholders and communities into our current and future work.

[33] https://www.diglib.org/preparing-la-gaceta-for-text-mining/

Collections as Data Position Statement

Laurie Allen, Library of Congress

I'm writing this position paper from a couple overlapping, but distinct perspectives. One is as the new head of the Digital Innovation Division (LC Labs) at the Library of Congress, where I recently joined a team who have been experimenting with collections as data, and providing leadership in this area for many years. And the other is as a professional in the larger digital library landscape, working for the past 3 years in Digital Strategy at the Library of Congress, but fascinated by the possibilities and problems with understanding our collections as data for a lot longer than that.

With those two perspectives in mind, I'll describe a little bit about the aspects of collections as data work that are being explored within LC Labs, and then I'll reflect on the field with a few broader thoughts. After my thoughts, I've included an excerpt from a longer report that gives more info on a current LC initiative to explore methods and services for providing access to collections as data within cloud computing environments.

A full history of the Library of Congress' approach to the topics touched on in this prompt would fill many books, and would need to be written by someone other than me. But, suffice it to say that the Library of Congress has been engaging with the thorny topics that arise from approaching collections as data for a long time. Colleagues across LC are constantly acquiring, transforming, moving, creating, and sharing collections as data in their day-to-day work. In one corner of the Library, where I work, a small team has been experimenting with novel approaches to collections as data work over the last several years. The Digital Innovation Division (LC Labs), a division of the Library of Congress Digital Strategy Directorate in the Office of the Chief Information Officer (OCIO) has found that, no matter what kind of innovation, experimentation, re-mixing, or re-imagining we want to invite people to do with our collections, the data offered up tends to defy user expectations for data. The heterogenous nature of our materials plus the idiosyncrasies introduced by historic and contemporary library practices prove to be a huge barrier to creative use. Even when it's standardized metadata, our data is messy, surprising, and confusing for people who are attempting to use it as data. So, "collections as data" work, or efforts to make our data more readily usable with computers, has been at the heart of the work of LC Labs for years.

In 2019, the Library was awarded a $1 million Mellon Foundation grant to support the *Computing Cultural Heritage in the Cloud* (CCHC) initiative. CCHC is designed to explore service models for using cloud technologies to enable computational research with large-scale digital collections within the unique context of the Library of Congress. For most of its duration, the initiative was designed and led by Senior Innovation Specialist, Dr. Meghan Ferriter, who envisioned an expansive approach to the many dimensions of CCHC and integrated CCHC into a tapestry of overlapping initiatives, some of which pre-dated or extend beyond CCHC. CCHC built on the Library's previous and contemporaneous work making collections available as data, from the public release of the loc.gov API through the creation and distribution of experimental datasets on "LC for Robots," and the work of digital collections colleagues who are developing methods for acquiring, ingesting, storing, and sharing datasets through the "selected data sets" initiative. It also relied on a constantly evolving and expanding set of affordances offered by

cloud vendors. Meghan's leadership centered a collaborative and coalition building approach to the work, inclusive of the expertise of colleagues throughout the Library. Further description of some aspects of that work are described in fuller detail below.

Formal outcomes from the CCHC grant funded initiative are in development, and will be shared toward the end of 2023. So far, I will share some of my big takeaways from the work of CCHC. I should note that I've only recently fully re-engaged in the CCHC effort, and these reflections on my part represent years of work from the CCHC team and others, but also might mis-represent some of that work. I'm still learning here. My takeaways.

1. In LC Labs, we are designing experimental data packages to describe, as much as possible, the intended uses of data, its provenance both as a collection, and as data, and any transformation it has undergone. This means a considerable amount of writing goes into each data package, and designing templates to capture and share the expertise of curators, technologists, vendors, and others has been a necessary component of this work. Given our experience with the questions that people bring to these data, and the increased ability of computers to parse text, this investment in thorough, thoughtful, and context-specific documentation is something we'll need to keep balancing against the requirements of scale.
2. Augmenting data has a lot of powerful possibilities, and the pipeline back into our discovery systems should be part of the plan, in some cases. Cloud services offer opportunities to enhance data in various ways, and our work will continue in creating tools to help ourselves verify the credibility of augmented data, and to make and document decisions about that augmentation. Version control is going to be key. (others know more about this part than I do)
3. Building and supporting communities for data experimentation both inside and outside of organization is vital. The data package is not the end, nor is the augmented data in our systems – the community of re-mixers and re-imaginers is the goal, and the more we keep them at the center of our work, the stronger it will be. Currently, our materials are often so difficult for them to use, we are probably going to need to continue to pay people to make the effort to use our materials for a while more. And we need to respect and time for the expertise of the people who know the collections best from every angle (technical, historical, subject, etc), and the needs of the people who want to make the collections into something new.

I'll close my section with a flurry of thoughts, that are a little less connected specifically to CCHC, or to the work of LC Labs:

- Sometimes I think "collections as data" is so tricky because "data" is supposed to be generalizable and generalized, and yet the materials held in our collections are held there because of their specific value – because each item stands as a specific and valuable piece of an infinite number of puzzles, and staff at our organizations selected this item to gather, describe, and hold onto over generations because of its very specific value. Maybe treating collections as data shouldn't work?
- Our collections are organized in ways that are inextricably bound to the institutions they support, and so, by putting collection materials out into the world "as data," we decontextualize them and make them part of the world of computers and the expectations of contemporary

users, which might be a delightful and powerful way to enable their recontextualization in support of new meaning-making, to undo or re-imagine them.

- Relatedly, the contemporary digital-age creates expectations about how information is formed, stored, shared, and/or exchanged that our materials frequently just don't match. Which is kind of awesome! But it means that people who want to understand the stories and lessons that our materials might offer have to spend a lot of time learning not just the assumptions of the creators of those materials, in whatever place and time they were creating, they also have to learn the practices of intervening librarians and institutions in storing those things. This can feel like an insurmountable barrier. To the extent that we can use the "collections as data" umbrella to help bridge the divides between the expectations of regular people who live in a datafied, digital age, with the stories within our materials, it's worth pushing on.

The following description of CCHC activities and lessons is excerpted from a S**ummary of *Computing Cultural Heritage in the Cloud*-supported Research Activities written by Meghan Ferriter and Innovation Specialist Eileen Manchester,** who has been deeply involved in supporting CCHC activities.

Excerpt Begins:

To test technical pathways and design scaffolding to support users, the CCHC team orchestrated two authentic research experiences involving outside scholar-practitioners**.** First, from May 2021 to January 2022, the Library of Congress contracted with three teams (two individuals, one group of two) for a ninth-month period to carry out independent research projects involving Library of Congress digital collections and computational methods. Second, in October 2022, the team refined a second iteration of our "computational access model" through a data jam with expert users. What follows is a summary of these CCHC-supported research activities and the resulting findings.

Overview of Research Activities

CCHC Researcher Cohort: May 2021–January 2022

As laid out in the relevant Broad Agency Announcement, the Library of Congress sought to award contracts to four researchers in three areas: early-stage exploratory research, advanced topical research, and advanced technical research. Only three researchers were selected, and all three projects fell into the early, exploratory category. The CCHC team's intent was to select proposals that represent a diversity of formats, research methods, and support needs. Overarchingly, proposals were evaluated for their ability to address technical as well as content-specific questions, and work across multiple types of Library materials. Fundamentally, all proposals had to make use of "computational methods that require large-scale access to collections and computational power."

The contract recipients were Andromeda Yelton, Lincoln Mullen, Lauren Tilton and Taylor Arnold. All three groups (Tilton and Arnold worked as a team) hailed from university settings and had disciplinary

ties to the Digital Humanities. Yelton, software developer and librarian, built a corpus of roughly 590,000 full-text items from the Reconstruction period (1865-1877) and applied machine learning methods to visually cluster the textual corpus by similarity. Mullen, professor of digital history at George Mason University, expanded his existing work to identify Biblical quotations within the Chronicling America database; his CCHC project refined the software he built to continuously crawl loc.gov and added an additional language detection component. Tilton and Arnold were the only researchers to work with historic photographs instead of text. Their CCHC project prototyped applying machine learning algorithms to detect faces, objects, poses, and regions within five documentary photograph collections at the Library. The resulting proof-of-concept tool shows readers not only the conclusions drawn by the algorithm but also its "level of certainty." More information can be found in the table below.

Researcher	**Area(s) of exploration**	**More information**
Andromeda Yelton	Analysis of ~590k full-text documents from 1865-1877; technical project to cluster documents by similarity of form and content	https://github.com/thatandromeda/lc_etl
Lincoln Mullen	Developed software to build and continuously update a local copy of LC digital collections; created scripts for detection of quotations and languages	https://github.com/lmullen/cchc
Lauren Tilton and Taylor Arnold	Adapted and applied computer vision algorithms to analyze photographs in five LC digital collections via face detection, pose detection, object detection, and region segmentation; developed prototype interactive visualization showcasing results	https://github.com/distant-viewing/addi/ https://distant-viewing.github.io/addi/06_interactive_viz/build/?id=2017877547 https://distant-viewing.github.io/addi/06_interactive_viz/build/?id=2014712029

CCHC Data Jam: September–October 2022

In October 2022, the Computing Cultural Heritage in the Cloud (CCHC) held a virtual Data Jam featuring speakers from higher education, library, and museum organizations around the world who presented feedback on working with Library of Congress data in the cloud. To provide access to the new material, our team established the LC Labs Data Sandbox, comprised of both a cloud hosting space (s3://data.labs.loc.gov) and a user-friendly interface on the LC Labs website (//data.labs.loc.gov). Each data package consists of a bundle of digital files ("assets") as well as technical documentation about how the dataset was compiled and contextual documentation describing the digitization and collection history of the material at the Library.

After compiling the data packages, the CCHC team recruited seven experienced professionals with both data and research skills who could provide feedback on accessing Library data predominantly via computer programs by participating in the CCHC Data Jam. We moved the data packages to the experimental s3://data.labs.loc.gov sandbox space in the cloud and assigned each data wrangler a data package and computational access method (REST API, SDK, or AWS CLI). The goal was for each user to provide as much real-life, authentic feedback on the experience of accessing, analyzing, and representing Library of Congress data in this way. More information can be found in the table below.

Data Jam participant	**Access pathway**	**Data package**	**Presentation title & time stamp in Data Jam recording**
Quinn Dombrowski, Academic Technology Specialist in the Library, at Stanford University & Nichole Nomura, PhD candidate, Stanford University Department of English	AWS REST Application Programming Interface (API)	Selected Digitized Books–166,218 .txt and JSON files containing full text from 90,414 books in the Selected Digitized Books collection on loc.gov.	2:16:10 "Juvenile Fiction in the LoC Digitized Collection"
Zoe LeBlanc, Assistant Professor in the School of Information Sciences, University of Illinois Urbana-Champaign	AWS Command Line Interface (CLI)	Stereograph Cards–39,526 stereograph card images from the 1850s through 1924.	26:17 "Subject Unknown: Stereograph Cards, Historical Metadata, and Feedback for Data Jam 2022"
Vikram Mohanty, PhD candidate in the Department of	Software development kit (SDK)	Austro-Hungarian Maps–4,998 images in TIFF format	50:55 "Computing Cultural Heritage in the Cloud Data Jam:

Computer Science, Virginia Tech University		representing non-georeferenced map sheets and corresponding GeoTIFF formatted images that are georeferenced.	Austro-Hungary Map Dataset"
Tim Sherratt, Associate Professor of Digital Heritage in the Centre for Creative and Cultural Research, University of Canberra	AWS REST Application Programming Interface (API)	Selected Digitized Books–166,218 .txt and JSON files containing full text from 90,414 books in the Selected Digitized Books collection on loc.gov.	1:48:15 "The American World Gazetteer or too much text, too little time..."
Aaron Straup Cope	AWS Command Line Interface (CLI)	Stereograph Cards–39,526 stereograph card images from the 1850s through 1924	2:36:22 "Feedback on Stereograph Cards Dataset" by Aaron Straup Cope, Head of Internet Typing, San Francisco Aviation Museum and Library
Daniel van Strien, Digital Curator, British Library (at the time of data jam)	Software development kit (SDK)	Austro-Hungarian Maps–4,998 images in TIFF format representing non-georeferenced map sheets and corresponding GeoTIFF formatted images that are georeferenced.	1:07:47 "LOC Data Jam: Maps and Python SDK"

Current Findings to date

The CCHC research team conducted two rounds of analysis on users' feedback in these sets of research activities. The first phase of analysis took place immediately following the research engagement. These internal-facing findings benefitted from their proximity to the engagements, and allowed, through their specificity, for immediate action to be taken around feasible next steps. For example, the CCHC

researchers' feedback on the JSON/YAML API led the Deputy Director of the Library's IT Design and Development Directorate and the Chief of LC Labs to establish a shared internal initiative to improve the API documentation at https://loc.gov/apis in April 2022.

A second round of analysis was conducted in March 2023. Distance from the execution of the engagements and the addition of new staff perspectives allowed the team to synthesize and summarize higher-level findings and to situate these outcomes in relation to the team's forthcoming recommendations for the grant as a whole. The following themes emerged from careful rewatching, note-taking, memoing, and dialogic reflection on the two CCHC-supported research engagements:

Finding 1: Sociotechnical approaches require institutional knowledge

A sociotechnical approach, which meets the needs of researchers, requires gathering institutional knowledge about how collections came to be acquired, cataloged/described, digitized, and put online. Some institutional knowledge may be readily available online; some of it may not.

Finding 2: Computational research is cumulative

Many of the researchers associated with CCHC repurposed past methodological approaches and technical toolkits. The ways in which these computational users approached digital materials reflected their disciplinary and methodological training, as well as time spent on building technical research tools, such as software or computer code.

Finding 3: Computational research methodologies often include data enrichment

Computationally intensive researchers do not rely exclusively on existing metadata. Often, researchers engage in data enrichment, augmentation, and/or enhancement strategies to provide broader and deeper information about the materials under study.

Finding 4: Multiple approaches for enabling bulk access to digital materials are possible but many require human intervention/mediation

Multiple approaches for enabling bulk access to digital materials are possible. Some, like prepackaged datasets, are done on the front end of a workflow by librarians and technical experts. Others, like APIs, require selection and compilation of content by a user on the back end of the internal workflow. Regardless of approach, bulk data access consistently benefits from intervention, communication, and guidance by Library staff including but not limited to those with subject matter, technical, and problem-solving expertise.

Collections as Data Position Statement

Chela Scott Weber, OCLC

In my role as a senior program officer for the OCLC Research Library Partnership (RLP), and in preparation for participation in *Collections as Data: State of the Field and Future Directions,* my colleagues and I convened two discussions sessions on the topic of collections as data with members of the RLP. This statement summarizes key takeaways from those discussions.

An open invitation was extended to all employees at RLP partners; 70 individuals participated from 39 separate RLP institutions in 6 countries (Australia, Canada, Hong Kong, Netherlands, UK, & US), from academic, museum, federal, and independent research libraries and archives. Participants reflected the diversity of roles and skills necessary to facilitate collections as data work—the conversation included people with technical services, public services, research support, digital projects and scholarship, and special collections responsibilities, and ranged from administrative and managerial roles to individual contributors.

The discussions were structured around the following questions:

- How has your institution supported collections as data efforts?
- Where is your institution on the project to program spectrum with regard to collections as data? Is it a priority and what kind of resources do you now have for this effort? Do you expect that prioritization to change?
- What are you hearing from researchers about collections as data needs and approaches? Are their needs or expectations changing or evolving?

Experience with and support for collections as data varied across the group, but none of the participants described their institutions as having a full-fledged program. Most participants talked about being in an early stage with this work, had limited staffing resources to devote to it, and were dealing with requests for data sets on an ad hoc basis or as a part of discrete, one-off projects. While people described successful responses to requests, they also discussed the challenge of a reactive stance to serving this research need. One participant explained, “[We are] trying to make interventions at point of need when there's a researcher who needs access to a data set. And there's a long time-lag between that initial request and the ability of the library to obtain the data. Oftentimes it's too late, it just didn't meet the need.”

People also described working to advocate for collections as data, approaching it on a policy and planning level as well as an operational one. A key challenge here was making confident assessments of just where collections as data should be their priorities. Many participants did not see consistent or predictable demand for collections as data, and were unsure if this was because it isn’t a widespread need or because researchers aren’t aware the library offers services in support of computational access.

This uncertainty impacts advocacy, as it can make it difficult to know what resourcing is warranted. As one participant put it, “The real challenge is figuring out, okay, if we're going to roll out infrastructure for this, is it worth it?”

Working with researchers was central to concerns and challenges raised in the discussions. Many participants felt a disconnect from researchers in this space, and described a need for greater understanding of researcher needs. They are unsure how to provide data to researchers in ways that will be useful to them and what kind of tooling (if any) might be most beneficial to provide. They also described a disconnect between the skills researchers had and those they need to work with library data sets. In many cases, researchers are interested in large data sets, but do not have the subject knowledge to understand what questions to ask of the data, or conversely, might have the subject knowledge but lack the technical skills to work computationally. There was a sense that in addition to providing data sets and tooling, outreach to make datasets visible and understood was also needed.

Issues around navigating rights and restrictions also surfaced as a challenge, one that impacts access to collections and ability to scale up work around collections as data. With regard to vendor supplied collections, there was frustration with DRM or other mechanisms put into place that limit or slow programmatic access to materials, as well as a desire for access via APIs. Participants discussed a desire to make mechanisms for computational access a commonly understood expectation for vendor-supplied systems and collections. One participant reflected this desire by asking “How do we make this an expectation within the systems that we have? So that this isn't a unique ask that [an individual institution] needs, it's something that's just expected of what our systems do for us.” For collections that originate in archives and special collections, participants voiced concern about being able to provide computational access to digital collections in a way that minimizes copyright, privacy, and confidentiality concerns, and allows the archive to keep any contractual obligations to the collection donor or creator.

In general, scaling up work around collections as data was described as a challenge throughout the discussions. Some institutions are leveraging projects to build program, learning from experimentation, and folding it back into creating workflows and frameworks for the future. Pilot projects that involved both library staff and researchers were seen as particularly valuable. More mature programs are building tasks that will support collections as data into routine workflows in cataloging, digitization, and data management. One challenge to this is the many different people and departments that may need to be involved in collections as data work. In many cases, this work was like the parable of the elephant in the dark room, with each person only seeing and understanding the portion they are close to, making it challenging to collaborate, communicate, and understand holistic need.

Participants in our conversations clearly aspire to do more with collections as data, but still have much to figure out to move this work forward. They described a need for shared knowledge across the LAM sector that could help support collections as data work, to make the lift lighter at individual institutions. Examples included best practices for an MVP for data packages, pedagogical frameworks to help people understand data sets and what they can do with them, or functional requirement examples to help advocate for needs with systems and collection vendors.

Collections as Data Position Statement

Kevin Hawkins, Opioid Industry Documents Archive, Johns Hopkins University

The Opioid Industry Documents Archive (OIDA) is a collaborative undertaking between the University of California, San Francisco and Johns Hopkins University. The mission of OIDA is to collect, organize, preserve, and make freely accessible documents from the opioid industry to enable multiple audiences to explore and investigate information which shines a light on the opioid crisis. These documents have been made available through on a web portal called the Industry Documents Library (IDL) that also includes documents from tobacco and other industries that influence public health. These documents in OIDA were selected for disclosure by a court, a settlement, or other legal process, and care has been taken to redact personally identifiable information and personal health information that may appear in these documents.

The IDL offers a traditional digital library interface that allows users to search and browse items by metadata and full text. The IDL software developers (part of the OIDA team) have prioritized offering this conventional digital library functionality over adding support for data science, but they are eager to see other members of the OIDA team build out our support for researchers to use computational methods.

We envision offering a "toolbox" of various options for users to work with the OIDA dataset, such as:

- analyzing the data in the SciServer virtual environment or in AWS
- downloading the data to one's local grid computing environment or even personal computer
- querying and manipulating the data through an API and through aggregators and other tools that make use of the API

This menu of options is meant to accommodate users of varying technical skills, with or without their own access to computing infrastructure. For beginners in particular, we believe SciServer, which offers a browser-based environment for analyzing datasets using Jupyter notebooks, will provide a useful starting point.

The OIDA team is developing support for data science not just for users of the archive but also for our own team members in order to help us improve access to the collections. With a rapidly growing collection of millions of documents that cannot possibly be reviewed by hand, the OIDA team needs to use computational methods to create and enhance metadata, detect meaningful images within documents, and otherwise process these documents. The toolbox will be useful to members of our own team in providing new ways to query and explore the OIDA dataset, and we will need to balance priorities in developing the toolbox for both external and internal users.

Broadly speaking, all of these use cases for applying data science to OIDA involve a user processing documents, whether directly or through an API. However, the emerging challenge facing the OIDA team—and those stewarding other collections—is whether to attempt to promote or restrict use of these collections by large language models and other rapidly emerging AI technology built on web-scale training data. We are considering training an open source LLM on the OIDA corpus so that we can offer something to the public that is transparent in its operation (including its built-in biases) and that, by relying on this corpus in particular, will be more useful than general-purpose LLM-based tools like ChatGPT.

At the same time, OIDA team members have been considering our role as an independent provider of factual information. We are committed to disclosure of these documents to the public even though that means that some uses of the content will be unexpected and will contradict our aim to shine light on the opioid crisis. For example, we are aware that Juul used tobacco industry documents in the IDL to devise a marketing strategy targeting children. Just as we do not feel that we can discriminate among users and uses of our digital library interface, we are reluctant to prevent AI technology from using our content.

Collections as Data Position Statement

Andrew Cox, University of Sheffield

The momentum behind the collections as data movement reflects that many research libraries are attracted to the idea of deeper exploitation of the special collections they own through working in projects with digital humanists (and others) and through this developing better search services for all users of those collections.

But many libraries do not have strong special collections. They may be better operating as gateways to wider data resources for emerging cross disciplinary data scientist research communities within their institutions. So, in the longer term it would be better to define a much broader data role for libraries, where the insights from work around collections as data is widened to broader tasks:

1. Data discovery across a complex data search landscape beyond the institution
    - Prompting all researchers to do data searches as a routine part of their literature reviewing
    - Assisting data scientists to discover secondary data in a very fragmented data search landscape
2. Data use
    - Advising on data provenance and quality
    - Advising on uses of data within copyright and other legal (eg GDPR) restrictions.
3. Data sharing
    - Promoting data sharing as a routine research practice
    - Advising on where to share data, how to describe it and license it
4. Data preservation and destruction
    - Designing the policy/ organisational and technical infrastructure to preserve derived data
    - Reminding staff to destroy data according to ethical/ legal requirements

Major barriers are:

1. Limited technical development resources of many libraries
2. Weak collective action across the library community
3. The power of monopolistic publisher platforms
4. Rapidly evolving user expectations driven by Google, chat GPT etc
5. “Data” vocabulary and thinking is still not deeply embedded in library vocabulary and culture

Collections as Data Position Statement

Tamar Evangelestia-Dougherty, Director, Smithsonian Libraries and Archives

Background

The Smithsonian Libraries and Archives (SLA) is an international system of 21 library research centers and the Institutional Archives of the Smithsonian Institute. Previously two separate entities, in 2019 the *Smithsonian Libraries* (SIL) and the *Smithsonian Institution Archives*(SIA) joined forces and integrated to better serve researchers, curators, educators, and learners of all ages at the Smithsonian and around the world. Tamar Evangelestia-Dougherty became the inaugural director of the integrated SLA in December of 2022. Within the field of library, archival and information studies, SLA is an Association of Research Libraries (ARL) affiliated research library which is uniquely categorized as a museum library and archives system. We are also classified as a special library.

The SLA's collections include over two million library volumes in subjects ranging from art to zoology. Consisting of more than forty-four thousand cubic feet, our archival records chronicle the growth and development of the Smithsonian throughout its history.

The SLA enjoys a dynamic and multifaceted role within the Smithsonian Institution:

- As the largest museum library system in the world which consists of 21 branch libraries and the institutional archives of the Smithsonian Institution.

  As a service provider to the Smithsonian, SLA services the research needs of the Institution's scholars, curators and researchers around the world.

- The SLA is a public educator with an active exhibition program and progressive digital library initiative, bringing important, rare and valuable works from the collections to the widest audience possible online.

  The Libraries and Archives together serve as the chief steward of the Smithsonian's institutional memory, capturing, preserving, and making available its history.

SLA Projects Employing a Collections as Data Framework

The Smithsonian Institutional Archives' (SIA) is a regular contributor to SLA's Biodiversity Heritage Library (BHL) which is the world's largest open access digital library for biodiversity literature and archives. BHL is revolutionizing global research by providing free, worldwide access to knowledge about life on Earth and the environment. Over 3,000 archival field books authored by scientists and researchers of the Smithsonian, have been published out as part of the BHL dataset of Smithsonian Libraries and Archives.

It is available through the Smithsonian Figshare system for researchers to work with computationally. There are no restrictions on the uses of datasets. The dataset includes all the descriptive metadata plus full text content data for the published material. All content is OCR capable and the handwritten field books have accompanying transcripts. As more and more of the SIA contributions are transcribed, our datasets have improved.

The Smithsonian Office of Chief Information Officer (OCO) Research Science group has collaborated with SLA on enhancing discovery of our collection materials related to women. Many of these projects are funded under the umbrella of the Smithsonian's American Women's History Initiative (AWHI). Using Artificial Intelligence (AI) this initiative enables SLA with the resources and technology to identify women within our collections and link staff data to collections data, especially in the National Museum of Natural History where many female scientists were hidden within the documentation of male scientists. "Because of Her Story" the "Funk List" was also part of this project. Work is currently underway to prepare our Smithsonian Women's Council archival holdings for similar use and we believe that this makes a great illustration of how our collections at the Smithsonian work as data.

As part of a National History Day programming in 2015, the Smithsonian Libraries and Archives separately made collections metadata openly accessible to Hackathons. The Archives did so in 2014 for the National Day of Civic Hacking together with the Smithsonian Cooper-Hewitt Museum.

A current instance and opportunity to incorporate collections as data at the Smithsonian Libraries and Archives is with the Smithsonian's Under Secretary for Science and Research (OUSSR), DR. Ellen Stofan's "Big Data" initiative granting program. SLA resides within the resides within the portfolio of OUSSR. "Big Data" Projects were sought with the potential to unlock well-defined science objectives by enabling digital access to a currently unavailable data set or improving access to an existing digital data set though database organization, enhancing queries or meta data. Projects to enhance a specific data science technology that will enable science were also be considered. SLA's proposal was successful and was recently presented at the Biodiversity Heritage Library conference in Paris, France. A presentation entitled "Enhancing the Museum Data Ecosystem: Mining Research Publications for Museum Specimens" was delivered by Richard Naples, Data Manager, of Smithsonian Libraries and Archives at the BHL Fostering Data Driven Natural Science through Open Digital Libraries.

Within our Freer Sackler Library for the National Museum of Asian Art, the Smithsonian Libraries and Archives followed participated in data enhancing projects for National History Day as depicted in this journal posting:
https://blog.library.si.edu/blog/2015/11/13/national-history-day-resources-and-a-hackathon/#.ZELhGHbMKHs .

Reflections on Challenges to Collections as Data at the Smithsonian Libraries and Archives:

Programming, marketing, and educating staff on how to facilitate researchers' use of our collections this way are our three biggest needs. Our staff is challenged with role overload and burnout as well as not

being equipped with adequate assessment tool or staff that can enable understanding of how to strategize. What goes into producing collections as data sets, and who can help researchers approach our collections computationally is a paramount concern. Riccardo Ferrante, SLA Associate Director of Information Systems, Digital Lifecycle & User Experience has discussed building and making datasets from primary *and* secondary sources across SLA available to researchers to work with through tools like Jupyter Hub. What seems to be the roadblocks to this work are: 1) it's not seen as a priority until a researcher comes to ask for it, and 2) the feeling that staff are burned out with the mundane daily tasks of librarianship and archival processing that there is a feeling of "we already have enough to do…too much to do to be visionary about collections as data."

A Smithsonian-centric challenge is that while it is great to be the Smithsonian…We are the Smithsonian. There is a shared sense of pride which means that we feel we do everything well but this approach can be quite siloed and not open to collaboration. The archives of the museums are administered under the collections or curatorial staff of each museum, not the Smithsonian Libraries and Archives. This creates a 'wild, wild west' approach to collections discovery and digital transformation. Many museums catalog archival collections in TMS, while SLA follows traditional discovery channels. At SLA we feel that researchers will go elsewhere until we have openly demonstrated our collections can be used innovatively as data in the manner that this "Collections as Data" initiative suggests. We have the ability (capacity) to help the museums work with the data, but that help is not always welcome or utilized.

In terms of educating and up-skilling our staff, it would be great to bring in Thomas Padilla to talk with our staff about this, perhaps as a kickoff to a discussion about how we can do this fearlessly and with excellence.

List of Smithsonian Libraries and Archives Library Research Centers by location with subject domains:

Washington, DC

American Art Museum / National Portrait Gallery Library / Renwick Gallery Library
Subjects: American art, portraiture, American history and biography, and American crafts.

Anacostia Community Museum Library
supports work on history and culture of the African diaspora in the D.C. area and more broadly in the Western hemisphere.
Subjects: Upper South, African American women, slavery and abolitionism, and religion and the African American community.

Botany and Horticulture Library (National Museum of Natural History)
Subjects: plant systematics, botanical history, ethnobotany, botanical art/design/illustration, floriculture, arboriculture, integrated pest management, gardening, plantscaping, etc.

Freer Gallery of Art and Arthur M. Sackler Gallery Library

Subjects: Artistic traditions/cultures of the peoples of Asia. Chinese and Japanese art represent about half of the collection.

Hirshhorn Museum and Sculpture Garden Library
Subjects: Modern and contemporary art, including painting, sculpture, drawings, prints, photography, video, and emerging art forms.

John Wesley Powell Library of Anthropology, National Museum of Natural History
Subjects: Physical anthropology, archaeology, cultural anthropology, linguistics, forensic science, area studies.

National Museum of African American History and Culture Library
Subjects: All aspects of the African American experience.

National Museum of American History Library
Subjects: History of technology, all aspects of American history—social, cultural, political, and economic, history of everyday American life, etc.

National Museum of Natural History Library
Subjects: General science, biology, ecology, evolution, biodiversity, geology, paleontology, conservation, etc. Includes sub-branches/satellite libraries in Invertebrate and Vertebrate Zoology, Mineral Sciences, Paleobiology.

National Postal Museum Library
Subjects: postal and philatelic history.

National Zoological Park Library
Subjects: veterinary medicine, pathology, genetics, nutrition, behavior, husbandry, wildlife conservation, biodiversity, zoo and aquarium horticulture.

Smithsonian Environmental Research Center
Subjects: Global change, population and community ecology, coastal ecosystems. Emphasis on Chesapeake Bay area

Warren M. Robbins Library, National Museum of African Art
African visual arts, including architecture, painting, sculpture, prints, pottery, textiles, popular culture, photography, rock art

Maryland

Museum Support Center Library:
Subjects; collections storage, research, and conservation

Vine Deloria, Jr. Library, National Museum of the American Indian
Subjects: All aspects of American Indian history and cultures, including architecture, health, law, education, music, dancing, religion, languages and literatures, pow-wows, etc.

New York, New York

Cooper-Hewitt, National Design Museum Library
Subjects: Design and decorative art from the Renaissance to the present.

Virginia

National Air and Space Museum Library at Udvar Hazy
Subjects: Space and aviation history, air transport, astronomy/astrophysics, terrestrial and exogeology, remote sensing, spacecraft design and instrumentation, etc.

Republic of Panama

Smithsonian Tropical Research Institute Library (Republic of Panama)
Subjects: Tropical biology, ecology, conservation, pharmacognosy, ecotourism, etc. Main site is in Panama City; branches in research stations on Barro Colorado Island on Gatun Lake.

Collections as Data Position Statement

Patricia Peña, Communication Faculty, University of Chile

patipena@uchile.cl

The Collections as Data proposal is an absolutely innovative vision and approach in relation to what we have and what is currently being developed in Chile, in initiatives and proposals in areas such as access to knowledge and free culture, public libraries and their online digital version, heritage institutions, access to open educational resources, whether from public, academic or cultural institutions.

In my work related to open data from communication and journalism, called data journalism and information visualization, a complex problem and challenge is that it has not been established as a field and area of development, neither from research or academic training nor from professional practice. This, in particular, has several levels because it ranges from the lack of data literacy to the lack of vision and editorial support to create work teams in these areas. There is also no consolidated work in the area of digital humanities, practically nonexistent in universities, or in the dynamics related to open science or open knowledge projects. Currently we have an important challenge in Chile, to bring back to the center the discussion on free culture, open access or even the ethics of the "commons", and the experience was very complicated.

Through the cohort initiatives that have been promoted in this project, they are an example of what it implies to broaden and complexify but also to make the possibilities of what is considered data more diverse. This, of course, is associated with challenges in its implementation and scope. It seems to me a major contribution, the whole proposal of guiding principles for projects of this type for those who want to implement and manage them. I will mention two challenges:

One of the main challenges in projects that mix computational, the world of data and its use has to do with the engagement and participation of users and communities, especially in relation to why a project is important to develop or in relation to its impact. It is not the same for a project that is focused on the scientific or research community or for journalists, but neither is it the same when we talk about a "community", thinking of users for example, when we think of users of a project that is focused on the scientific or research community or for journalists. thinking of users, for example, of a library, a cultural center or an information access platform. This is very variable, and this also implies different work strategies and project design. In particular, this is related to the type of co-design or participatory design methodology that can be applied from the beginning of a project that seeks to involve the needs and expectations of a specific group of people or any community (understood as people who share a common interest, need, expectation). In addition, as I mentioned before, it will often be necessary to accompany people with more basic training and support processes, especially in more vulnerable communities, who do not have the understanding, skills, competencies or knowledge even in relation to the meaning of heritage, memory or cultural heritage.

A second challenge has to do with technological sustainability, because although there is a vision that

this type of development should be implemented under the logic of using open-source software, openess or self-managed platforms, there is currently a strong tendency even in academia to rely on platforms that are not necessarily open, but rather proprietary systems. It is also necessary to think at this level in the training of those who develop systems, that is to say, of those who produce the solutions and computational tools to have a training and vision on these issues. Currently, in countries like mine, this training is totally oriented to the insertion in the industry and not necessarily in social or cultural projects.

And a third challenge is associated with openess and the current debate on "data justice". This is related, especially when data involving individuals, groups of people or communities and how in addition to the legal considerations that exist on protection and privacy (eg: images, audiovisual material, personal files) is that it is a consented or consented process, but also that it is not a process in which whoever designs that research that will lead to develop a collection applies only an extractivist logic (to see that data obtained as a resource for the benefit of their idea or research), but that it implies an opportunity for those from whom those data were obtained, also to be considered as part of the management, evolution and sustainability of the projects. This point is related to the previous one.

Several people, colleagues and friends who have promoted and worked on the vision of promoting collaborative projects of the type proposed by Data as Collections, had the opportunity to open this discussion about openness and commons resources in the recent constitutional process we had in Chile, to bring forward a New Constitution, in order to think about a future that would allow to review legal frameworks and regulatory notions on copy right or cultural and educational frameworks about knowledge and creation, for example. The debate that we were able to open at the political and social level on these issues was minimal and, as you may know, very disappointing since the proposed constitutional proposal was rejected. The opportunity to participate in this workshop with you, is also an opportunity for us to collaborate in the future to enhance the debate and training in these areas.

Prompt: We strongly welcome bridging, divergence, and provocation in your position statements. Is there something concrete or conceptual we are missing? Are there questions and communities we aren't currently considering? This is an opportunity to highlight aspects of your experience that relate to collections as data and will help stage interaction at the face-to-face meeting as we collectively consider the state of the field and future directions together. As a whole, these position statements will form a collective resource to be shared publicly with any community interested in collections as data.

Collections as Data Position Statement

Yvonne Ng, Senior Program Manager, Archives, WITNESS

Coming from outside of an academic institution or cultural heritage sector, the questions and experiences that I bring to this meeting may differ from the other collections-as-data initiatives represented here. I work at a global nonprofit organization that supports people using video to protect and defend human rights. The collections we deal with are primarily contemporary audiovisual records that document human rights violations, often filmed by people experiencing or witnessing violations in their communities. Data, in our context, are deeply intertwined with living people and communities, and data practices can have a direct impact on their agency, rights, and safety.

Our work centers the people and communities impacted by violations, and supports local advocates and experts who are using video to document. As such, we are aligned more closely with "data creators" or "providers" than with "data consumers," although we do aim to support human rights defenders to use their collections most effectively to achieve their justice and accountability goals, and to help them create collections that can be used by other stakeholders in support of human rights. To that end, we are interested in ways to democratize access to computational tools and techniques for data analysis that can be employed sustainably by non-institutional actors with sensitive audiovisual collections in less-resourced contexts. We are also interested in exploring models for controlled (vs. default to open) access that benefits and protects data creators and subjects, and gives them agency over what happens with their collections.

Democratizing access to computational tools and techniques for data analysis would help groups we collaborate with like Rohingya Vision, a Rohingya media organization dedicated to amplifying the voices of the ethnic minority community who have been persecuted in Myanmar for decades. Starting in 2018, we partnered with Rohingya Vision to help them create the Rohingya Genocide Archive, a Rohingya-led project to collect and preserve video evidence of genocide and crimes against humanity. They built a collection of over 600 videos that were submitted to various international courts and accountability mechanisms. We are now collaborating with them on a next phase, which will include making information about the collection more public, publishing project workflows, and training new archivists. They are also interested in building their capacity to analyze and glean information from their collections, so that they can use the collection in new ways in their ongoing advocacy and support for their community.

Developing access models that benefit and protect data creators and subjects, and that give them agency over what happens with their collections is also a key need. For example, we have been part of ongoing efforts and engagement to explore viable large-scale approaches to preserving critical human rights content recorded by human rights defenders, civic journalists and the general public in situations of armed conflict or generalized violence, that have been uploaded to (and possibly subsequently removed from) social media platforms. While preservation is challenging, the question of access to this content by external parties other than domestic law enforcement is an even more complicated conundrum that needs to be hashed out. Access models are also needed even in smaller-scale human

rights archiving initiatives. In a (still ongoing) survey we have been conducting to understand current needs of grassroots and activist archives and archiving initiatives, developing access and security protocols have been repeatedly cited as the main challenges by respondents across multiple geographic regions.

The Collections as Data initiative's aim is to support the scholarly use of collections as data, and as such, the duty of care and responsibility as outlined in documents like the Santa Barbara Statement on Collections as Data, are primarily oriented towards scholarly users / data consumers. For example, the principles that "Collections as Data should be made openly accessible by default, except in cases where ethical or legal obligations preclude it," and "Collections as data stewards aim to lower barriers to use," clearly privilege the needs of consumers (although passing reference is made "to respect the rights and needs of the communities who create content that constitute collections, those who are represented in collections.") This disposition may be understandable, given that many collections involved in this initiative may have clear donor agreements, be historical (with long deceased data subjects), consist of non-human subject research data, be not particularly sensitive in nature, or be part of larger institutions that exist primarily to serve a scholarly community. However, the prioritization of data consumers should not simply be taken as a given, since other collections and other approaches do exist. For example, the data trust model described by the Open Data Institute "involve[s] one party authorizing another to make decisions about data on their behalf, for the benefit of a wider group of stakeholders." In contrast to the principle of open access by default, the steward in the data trust model has a fiduciary duty to "represent and prioritize the rights and benefits of the data providers when negotiating and contracting access to their data for use by data consumers."

If the Collections as Data initiative chooses to prioritize scholarly users and data consumers, it could be more explicit about it and provide more context to justify the choice. The initiative could also still undertake a more thorough analysis of potential threats and harms to data creators and subjects, especially to vulnerable and marginalized communities, and identify ways to reduce harmful impacts. An example of such a process is a harms modeling that WITNESS led as part of our participation in the Coalition for Content Provenance and Authentication (C2PA), and its development of a technical specification for tracing the provenance and authenticity of digital assets. My colleagues led a task force to map potential harms, misuses, and abuses, and mitigation strategies for multiple stakeholders.

The significance of the pre-process: Building Collections as Data

Linda Garcia Merchant, PhD, Public Humanities Data Librarian, University of Houston; Co-Founder, Chicana por mi Raza Digital Memory Collective

This position paper calls into question the following statement from "Collections as Data: Part to Whole"

"Data driven scholarship depends on data that are computer readable and machine actionable. Unfortunately, most cultural heritage institutions struggle to meet this need because they do not think of their digital collections as this kind of data. Rather, items in digital collections are generally treated as surrogates of physical objects and their organization and the expectations for their use are based on the metaphor of a physical bookshelf, gallery wall, or listening booth. A collections as data orientation suggests different approaches to collection production, documentation, and access."

This statement suggests many elements about collections produced by cultural heritage institutions that are assumed to be true, specifically the character of cultural heritage institutions and the nature of production and publishing forms of digital scholarship within them. As cultural heritage institutions, we do mediate, analyze and publish content however, our bigger issue remains rooted in what is required of these processes to support and maintain the goal of building a corpus. Historically, cultural heritage institutions have not been given large scale production grants to generate complete runs of publications or collections as more than digitized and identified. Historically the understanding of what defines an entity as a cultural heritage institution implies ongoing streams of funding, self sustainability, and designated structures both in brick and labor dedicated to the preservation of heritage. Who we are and how we exist, by institutional definition, fails to speak to the challenge both of consideration as legitimate entities that can support themselves and any practice of preservation.

I offer this observation as both practitioner of the collecting process and as founding member of a non-traditional (read not supported by the institution) collective that exists to preserve a cultural heritage of a historically marginalized community. For the first 11 years we had no institutional affiliation. Our collective operated using small internal faculty funding, student research opportunities, and the private donations of time and labor from myself and other practitioners. During those 11 years our project was consistently rejected by every grant applied for–the primary reason being our lack of support from institutional infrastructure. For 11 years funders did not consider us viable, sustainable or significant enough in our field to fund, primarily because we lacked institutional support. And yet, we are considered by the communities we represent to be a cultural heritage institution holding a critical and valuable resource, even if our collection is post custodial and housed virtually, outside of a traditional library or archive. We continue to work with marginalized communities, and now we work with institutions willing to partner with us to build a post custodial resource that remains accessible to the communities it represents.

While we have yet to receive foundational grant funding, we have along the way, learned other strategies to create a sustainable and viable end product, still available to the broader communities outside of the academy as a resource. At first, we would work from small grants or no grants to digitize and convert to text, monographs, short runs, small scale bodies of work. We would leverage the

classroom experience with the collected data to produce metadata and create discoverable documents and assets, that foreground language and culture in approach. This presents the other issue when accomplishing this kind of discoverability–the issue of language and how that exponentially adds time and labor to a digitization project. For example, as a postdoc, I worked on a project to digitize a manuscript written at the end of the 19th century. the project included scanning the original book to TIFF images, converting those images to TXT using OCR, cleaning up the text to change incorrect characters and add spanish and spanish equivalent characters where necessary. The project was completed during CoVid where everyone was working virtually and when we could go to the office to scan materials, had to keep the number of staff in the building to 25%. Even with students working 20 hours a week on this one project, it took four months to finish one book and even then, several scholars had to review the final work and still make corrections to create an accurate match to the original.

We are still addressing discoverability of subjects and topics in existing collections. We are still addressing the misconceptions around the accessibility of archives, special collections, and what we know as publicly available collections in city and non university libraries. We are still struggling with the necessary advocacy to consider and make room and hold space for these histories of marginalized and underrepresented communities that are seemingly hidden. These histories are not so much hidden in as much as they are a challenge to uncover because of community resistance. The challenge represents many communities and subjects experiences with collecting institutions that are extractive and pre-defined in project scope and prohibitive in representational agency. This relationship perpetuates the suspicion and caution communities have when they have no active role in the acquisition of their materials.

Our histories exist because we have found ways to make them available to our communities often because access to libraries and archives are conveniently not made available to us and when they are, occur on terms and verbiage that define our culture and language and histories as special and separate. The history of this country is not black, latino, south asian, asian, or native, it is american, so to suggest that our histories and any acknowledgement of them must occur in three or four months of the year reinforces the same issues of exclusion and separation that hyphenating our identities does. Our contributions to the narrative and nature of the United States are as tragic and rich as all the acknowledged contributions to this long and fraught story.

Lastly, because we are also not actively employed in those spaces as researchers, faculty, archivists, librarians, and technologists, gaining access to the active use of archives is still accompanied by the stigma of suspicion of us. We come to these archives often recognized as subjects studied but not in the study of subjects. When we come to the archive to seek out subjects hiding in plain sight–as ancillary, often in relation or connection to better known figures in history, archives have been closed to us, determined to be unavailable, or accessible with no assistance by the staff. Archives collected without an understanding of their importance are sometimes deaccessioned, or even lost. Those of us familiar with a hidden history through actual experience and exhaustive research of the networked member organizations and related subjects, will know how and where to look for us in the archives. Those kinds of finding aids

don’t yet exist and there is no financial support for this kind of transgenerational translation that needs to occur to create robust discoverability of materials and new scholarship. Even though there is still plenty to do and it all has to happen at the same time. I will say we are not deterred, even given our tenuous relationship to the academy, to the archive, to the field, we get what we can done, when we can and in ways that are beginning to revolutionize our fields.

Exploring the publication and reuse of Collections as data

Gustavo Candela, University of Alicante, Spain

During the last decade, GLAM (Galleries, Libraries, Archives and Museums) institutions have been exploring new ways to make their collections available for computational use using a wide diversity of approaches. Initiatives such as the International GLAM Labs Community [1] and Collections as data [2] have paved the way towards a new context in which GLAM organizations need to be prepared for the future concerning the use of advanced computational methods [3]. However, there are still "unsolved problems" and challenges regarding the publication and reuse of collections as data:

- How can we publish a collection as data? Is there a simple way to do this? Are best practices and guidelines available that can be easily used? How can a collection be structured to facilitate its reuse?
- What are the requirements needed in terms of data quality to make a digital collection available for computational use? How can data quality be described?
- How can we encourage researchers to approach our collections and to better understand the content provided?

Unfortunately, there is no one-fit-for-all solution for these issues. In this sense, my research is focused on several aspects concerning the publication, reuse and data quality assessment of collections as data from an institutional and collective approach.

Inspired by, and in collaboration with the International GLAM Labs Community, a webinar was presented in October 2022 to bring together theory and practice on how to make GLAM digital collections as data available for publication and re-use by researchers, GLAM professionals, creative workers and other users. The webinar was focused on a checklist to publish Collections as data including 10 items to consider when preparing a digital collection for computational use. In addition, real-life examples were introduced from the National Library of Scotland, the German National Library, the Library of Congress, Europeana, the Royal Danish Library, KBR - Royal Library of Belgium and meemoo [4]. This led to the publication of a research article (*preprint*) describing the checklist creation process, its refinement thanks to the community feedback and the inclusion of several examples of use based on GLAM institutions that can be useful for other institutions willing to adopt the Collections as data principle [5].

Fig 1. Webinar presented in collaboration with the International GLAM Labs Community.[1] [1]
https://glamlabs.io/events/collections-data/

GLAM institutions are exploring new ways to show researchers how to reuse the content of their digital collections. In addition to the dedicated websites, Labs, data repositories and documentation (e.g., README files), Jupyter Notebooks have emerged as a powerful tool to provide documentation and code that can be run in cloud environments without the need to install additional software.[2]In this sense, we have designed a framework to extract digital collections as Collections as data that include the use of Jupyter Notebooks and has been applied to a wide diversity of digital collections made available by GLAM institutions [6].

In addition, as part of my research, I recently collaborated with the National Library of Scotland in the framework of the National Librarian's Research Fellowship in Digital Scholarship 2022-23 to reuse the digital collections provided in the Data Foundry.[3]In this context, I explored the use of the Semantic Web as well as the enrichment of the collections using external repositories such as Wikidata. A detailed framework based on several steps was provided in order to automatically transform digital collections into Linked Open Data that was applied to several metadata collections made available for the public in the Data Foundry [7].

Advances in technology have provided a new context in which data quality has become crucial for the application of computational methods. Natural Language Processing and Machine Learning methods require high quality data for specific tasks such as the training of models. In this way, I have worked on several research works concerning the semi-automatic and automatic data quality assessment of the collections by using data quality criteria based on several categories such as accuracy, completeness or interoperability [8].

An additional aspect to consider is the decolonization of digital collections in which computer methods based on Artificial Intelligence and the Semantic Web can help make the indigenous content more visible and reusable for researchers. In this context, I am part of the team of the *Unlocking The Colonial Archive: Harnessing Artificial Intelligence for Indigenous and Spanish American Collections* project funded by the AHRC, UK Research and Innovation (UKRI) and the US National Endowment for the Humanities (NEH).[4]

All this leads us to the following set of questions: What is the current state of institutions in terms of the adoption of the Collections as data principle? Can we think collectively in order to help institutions to make available their collections as data? Do we need a data space to standardize the publication process? How can organizations integrate research outputs into their daily workflows?

I would like to thank Sarah Ames from the National Library of Scotland, the personnel of the Biblioteca Virtual Miguel de Cervantes, the International GLAM Labs Community and the Unlocking Colonial Archive team for their help and support in writing this position statement.

Works cited

[1] Mahey, M., Al-Abdulla, A., Ames, S., Bray, P., Candela, G., Chambers, S., Derven, C., Dobreva-McPherson, M., Gasser, K., Karner, S., Kokegei, K., Laursen, D., Potter, A., Straube, A., Wagner, S-C. and Wilms, L. with forewords by: Al-Emadi, T. A., Broady-Preston, J., Landry, P. and Papaioannou, G. (2019) *Open a GLAM Lab.* Digital Cultural Heritage Innovation Labs, Book Sprint, Doha, Qatar, 23-27 September 2019.
http://dx.doi.org/10.21428/16ac48ec.f54af6ae

[2] Padilla, Thomas, Allen, Laurie, Frost, Hannah, Potvin, Sarah, Russey Roke, Elizabeth, & Varner, Stewart. (2019). Final Report --- Always Already Computational: Collections as Data (Version 1). Zenodo.
https://doi.org/10.5281/zenodo.3152935

[3] Research Libraries UK. A manifesto for the digital shift in research libraries. 2020.
https://www.rluk.ac.uk/digital-shift-manifesto/

[4] Gustavo Candela, Nele Gabriëls, Sally Chambers, Sarah Ames, Abigail Potter, Ellen Van Keer, Alba Irollo, Peter Leinen, Victor Harbo, Katrine Hofmann, & Olga Holownia. (2022, October 25). Towards implementing Collections as Data in GLAM institutions. Zenodo.
https://doi.org/10.5281/zenodo.7789480

[5] Gustavo Candela, Nele Gabriëls, Sally Chambers, Thuy-An Pham, Sarah Ames, Neil Fitzgerald, Katrine Hofmann, Victor Harbo, Abigail Potter, Meghan Ferriter, Eileen Manchester, Alba Irollo, Ellen Van Keer, Mahendra Mahey, Olga Holownia, Milena Dobreva. A Checklist to Publish Collections as Data in GLAM Institutions.
https://doi.org/10.48550/arXiv.2304.02603

[6] Gustavo Candela, Dolores Sáez, Pilar Escobar, Manuel Marco-Such: Reusing digital collections from GLAM institutions. J. Inf. Sci. 48(2): 251-267 (2022).
https://doi.org/10.1177/0165551520950246

[7] Gustavo Candela: Towards a semantic approach in GLAM Labs: the case of the Data Foundry at

the National Library of Scotland. CoRR abs/2301.11182 (2023). https://doi.org/10.48550/arXiv.2301.11182

[8] Gustavo Candela, Pilar Escobar, Dolores Sáez, Manuel Marco-Such: A Shape Expression approach for assessing the quality of Linked Open Data in libraries. Semantic Web 14(2): 159-179 (2023). https://doi.org/10.3233/SW-210441

Publication risk and the intentional re-introduction of friction

Sonia Ranade, The National Archives (UK)

The National Archives holds a rapidly growing collection of contemporary digital records, whether transferred within days of creation or harvested through automated crawling of live online content. The availability of contemporary digital collections greatly increases the relevance and utility of the Archive. But it also increases the risk that sensitive content will become discoverable. Without a robust response to this threat, the shift to digital which could offer so much promise for research, may instead bring about a reduction in access.

Compared with physical records, digital collections are highly discoverable. They are available at scale with great immediacy and are intrinsically amenable to computational processing. This opens up exciting opportunities for research but the converse is also true: many of the same characteristics which make digital archives a rich resource for research also increase the potential for harm. This will inevitably prompt greater caution in approving records for release. Factors such as the relative ease of copying and distribution; near-impossibility of recall; scope to compute and combine to derive new insight; and the ability to process large volumes of data, all contribute to the risks of publishing contemporary digital records and making them available for computational re-use.

For many libraries and archives, access to digital collections is regarded as synonymous with publication on the Web. However, while it is technically feasible for a public archive to publish everything it holds, it is neither lawful nor responsible. Archives have always made decisions about *what* we make available and *when* but in our increasingly complex digital environment we must also think about *how* to release records and consider what form(s) this access should take. The measured reintroduction of some of the friction inherent in working with physical records could help us manage publication risk. In parallel with our work to improve computational access to collections, archives will need to introduce carefully designed constraints that allow us to strike an appropriate and managed balance between the extremes of irresponsible publication and precautionary closure.

Fortunately, digital formats also give us greater scope to engineer solutions beyond simply ‘withhold’ ‘or ‘publish’. We will require finer-grained controls. For example, we might permit reading but not computation; limit the proportion of a collection that can be downloaded or copied; restrict access to particular physical locations (such as archives’ service or library reading rooms); offer mediated computational access via managed platforms or safe environments; create redacted versions that can be more freely re-used; or grant licenses for different modes of use such as reading, indexing or computation. The aim would be to create a nuanced access regime that brings about a reduction in publication risk without a corresponding reduction in utility, enabling us to widen computational access to archival collections as data in a responsible way. We must also anticipate that, despite these controls, sensitive content will sometimes be published - whether due to error, imperfect information for

decision-making, or legislative and cultural change over time. Responsible access requires ameliorative mechanisms for 'take down' and recall of content previously deemed suitable for release.

Much of the risk arises from ongoing technical innovation including the ready availability of large-scale processing capacity and the rapid growth of artificial intelligence and machine learning systems. These enhance the potential to surface information that would not otherwise be evident. For archivists, it is increasingly difficult to understand and predict where sensitivities will emerge. Facilitating the use of more sophisticated computational techniques for researchers is already an important aim of the *Collections as Data* programme. It is no less important to broaden access to these tools for archivists, librarians and records creators. The same capabilities that can expose sensitive data in published records could help us improve the quality of our decision-making on *what* to release and *how* and allow us to do so confidently at scale.

Find Case Law

We have been considering how to put some of these ideas into practice. Our response to the challenge of providing responsible computational access to archival collections as data started in 2022 with our *Find Case Law* service for Court judgments.

This collection is well suited to computational access:

- It serves a highly engaged user community drawn from the judiciary, legal profession, press and publishers, as well as academic researchers and citizens.
- The records are contemporary, often transferred to the archive within days of a judgment being handed down.
- This is a discrete collection of semi-structured documents which are reasonably well represented by a single data model.
- The service started with lower-risk judgments from the higher courts and tribunals (i.e. with an emphasis on points of law). We will extend it to more courts over time.

Some of the ideas outlined above are applied to help manage publication risk:

- Publication risk is assessed as part of our archival workflow.
- We aim to learn from our decision-making and apply this to develop a statistical model of publication risk.
- Our licensing model seeks to maximise access while protecting the proper administration of justice. It recognises computational processing (Transactional Licence) as distinct from storage, sharing and use (Open Justice Licence).
- Actively computing over the records has improved our own understanding of the collection.
- Crawling and indexing is not permitted, access may be blocked for sites that do not respect *robots.txt* or *noindex* instructions.

The service offers additional features to support computational access:

- Records as data in formats that support different modes of use.

    - PDFs created from the original MS Word documents to deliver a consistent visual representation and for printing.
    - HTML for reading and browsing online.
    - XML in LegalDocumentML (Akoma Ntoso) format which adds structural and semantic markup and can encode additional enrichment.
- Records are computationally parsed to identify links to related material, we will continue to enrich the collection over time.
- A choice of API or browser access to data.
- Notifications for newly added material that can be consumed by people or by automated processes.

We hope to generalise our learning from creating this service to help us extend computational access to more of the digital archive. We believe that this challenge is best addressed through collaboration and we welcome the opportunity to actively contribute to initiatives, such as this one, which aim to widen computational access to heritage collections.

Links

Find Case Law caselaw.nationalarchives.gov.uk
UK Government Web Archive www.nationalarchives.gov.uk/webarchive
The Transactional Licence for computational re-use of court judgments data
caselaw.nationalarchives.gov.uk/transactional-licence-form

Scale, thinking in different ways, new users, and new collections.

Alexis Tindall, University of Adelaide Library.

Years ago I took new digitisation volunteers on induction tours through wet collections of a natural history museum, explaining to them that every animal, and each tiny, ethanol soaked piece of paper in that room was data, and our goal was to make that digitally useful. So recently introduced to the millions of diverse items that make up our natural history museums, minds boggled at what that meant. Since that time I've been lucky to experience the challenge from diverse angles, in collections roles and user-focussed roles, and I believe that, to a large degree, the challenge remains the same.

Collections as Data: Part to Whole summarises the contemporary challenges of making collections computationally useful. In that challenge there exists a surfeit of potential, which can be simultaneously inspiring and inhibiting to the persons and organisations involved. I'm interested in exploring issues of scale, expanded ways of thinking, new users and approaches to new collections.

I am very interested in the potential of collections as data, and the hurdles to achieving and exploiting that potential, for collections *and* users. I worked in the early 2010s on a museum-based digitisation project that contributed to the *Atlas of Living Australia*, key informational infrastructure for biodiversity research communities. More recently I have contributed to support for the Australian digital humanities research community[34], before delivering a wide-ranging analysis of the Australian resource landscape and user community that influenced government investment in humanities, social sciences, and Indigenous research infrastructure[35]. Presently at the University of Adelaide Library, I support research data management, as well as progressing digital preservation in an institution that includes special collections, archives, recordkeeping and other unique digital collections, alongside library services.

Since 2020 I have delivered relevant community initiatives including a series of Digital Humanities Lab webinars[36], and contributing to the Artificial Intelligence for Libraries Archives and Museums (AI4LAM) Australia and Aotearoa New Zealand Chapter[37].

Scale and new ways of thinking.

Approaches to digitisation and collections discoverability have matured over decades, but capacity of organisations to participate in that space is unevenly distributed. I'm interested in exploring the achievability gap, and how thinking differently can be enabling.

In the abstract, our professional community recognises that collections can be delivered as data beyond well-resourced mass digitisation of large collections; that smaller, curated, and targeted digital projects are useful. But the tools and methods that make it possible for smaller collections or collections with limited scope for digitisation to participate often depend upon technical skills or resourcing that are rarely found in smaller organisations. In suggesting a 'surfeit of potential', I draw on experiences of small

[34] Tinker Data-Enhanced Virtual Laboratory (2018, 2019), selected resources
[35] ARDC. (2020). Humanities, Arts and Social Sciences Research Data Commons - Final Report.
[36] University of Adelaide Library (2021, 2022) Digital Humanities Lab playlist
[37] AI4LAM Working Groups: AU and ANZ Chapter

organisations, and large organisations with limited capacity for digitisation, where the challenge of delivering interoperable, machine-readable data can be overwhelming.

Users can also be inhibited by the achievability gap. In my experience working with researchers that were new to the digital humanities there was not limit to their interest and enthusiasm. In many cases the hurdle came at the steep incline in the learning curve between interest and enthusiasm, and scoping a research project that makes effective and compelling use of data and computational approaches.

I wonder whether wider familiarity with computational thinking can aid both collections and users. Not to suggest that all GLAM professionals need to code, but that increased exposure to elements of computational thinking could help, such as breaking a problem into parts, issues of abstraction, considering patterns, and identifying problems that benefit from quantitative consideration.

These experiences can help collections professionals ensure the usefulness and relevance of their choices making collections available as data, within the means of their organisation. It can help avoid project pitfalls and inform a user-centred perspective. It can help them evaluate the effectiveness and functionality of their outputs, many of which depend on vendor-supplied or community-developed tools, rather than in-house development. Improved computational thinking from collections professionals can also help users make most effective use of our data.

New users and new collections.

Users of collections as data won't necessarily be digital versions of our traditional users. At the University of Adelaide Library, efforts to encourage text mining and computational analysis in humanities related disciplines led us to observe growing use of text and data mining in STEM disciplines. Computational analysis was more mainstream in these disciplines, who were becoming more engaged with text as a useful source, and our collections as a relevant resource.

Understanding these communities, and supporting and influencing their methods of accessing text and image collections may be a mutually beneficial experience. This is also evident in my experience with AI4LAM. LAM professionals identify emerging challenges, and areas where our information management expertise would be beneficial, but are in the early stages of finding ways to apply that expertise beyond critique.

Expanded ways of thinking are also necessary as we become responsible for more and more 'Data as Collections'. In our University context, many of our traditional collections (correspondence, field notebooks, research data, experimental records and the like) are now created digitally. This is before mentioning our responsibility for massive collections of born digital business records, email and web archives. While parts of our profession are well established in digital recordkeeping, there is more potential to be realised if we can apply traditional principles like arrangement and description in new ways, and think about programmatic access when we consider discoverability.

Finally, I am also interested in the matters that give us pause. Firstly, we know better than anyone the work involved in maintaining collections. Do we rush forward, creating ever larger digital collections, knowing the work and cost involved in their sustainability? Secondly, custodians know the rich cultural

information and deep personal stories embedded in our collections. How can we embed the protections and mediations we know are essential as we consider providing access to these collections as data?

I'm very excited to participate in this forum, and appreciate the opportunity to connect with like-minded colleagues from around the world.

Show, Don't Tell: Marketing Collections as Data in Academic Libraries

Jenn Riley, McGill University

Nobody knows what we do

What academic librarian of the last fifteen years has not had this experience—You are speaking with a campus colleague from outside the library about their research or a discovery in their field making the news. The conversation turns to an area in which the library provides a service beyond the basics of discovering a book or journal article and obtaining it. As you start to describe your service and how your colleague might make use of it, their eyes widen, and they ask you with surprise "The library ***does*** that?!?" It certainly happens to me with some regularity, and it seems to be pervasive throughout our field. It can feel as if no matter what we do to reach our campus and bring patrons in (physically or virtually) to our libraries, our campus is still uninformed as to what a modern research library provides.

Services surrounding Collections as Data are no exception and are perhaps even more difficult to match to users as the corpus available is relatively small. While the library will have traditional print or electronic collections relevant in some way to every researcher on campus, the same cannot be said for Collections as Data. It is not enough to highlight a publisher's text & data mining package on a library subject guide or provide a link in a catalogue record taking the user to a data set for the original content represented by that record. We simply must put developing services front and center in as many places as possible and in as many ways as possible; researchers will not stumble upon them on their own.

New spaces, new services, new skills

At McGill University, we are currently in the planning stages for a down-to-the-bones renovation of our 1960s brutalist cube and 1890s heritage building somewhat chaotically added to and linked together that serve as our main library. This project to fully reimagine our library is called "Fiat Lux" (Latin: "let there be light"), taken from an inscription on a stone on the façade of the heritage part of the library building. The inscription was a serendipitous find, setting a tone of bringing knowledge out of the darkness and into the light as inspiration for our work.

The design for the remodel features a full technology-heavy floor devoted to digital research and learning. It will have all the feature you see today in newly built and newly remodeled academic libraries: makerspace, technology sandbox, virtual reality, a/v recording and editing, advanced computing, workshops, moveable furniture, open spaces for group work, screens and power everywhere, a sense of play… it's our instantiation of the model we've seen emerge in the last decade.

New spaces mean new library services and therefore new skills are required to offer those services. Collections as Data will be a key part of our digital scholarship services. We will need staff who can discover, engage with, market, and teach tools and methods for using Collections as Data in research, learning, creative, and civic work.

Marketing of Collections as Data could be done via a variety of traditional and emerging methods, and likely many different approaches will be needed to be effective. Visiting departmental meetings or research events, offering workshops, visualizing data sets on public display screens, advertising in campus newsletters, posting on social media, engaging with student organizations, hosting presentation events such as poster sessions, and offering research awards for projects using Collections as Data are just a few of the methods that might be employed. So who is it that will be performing these activities?

Possible staffing models

At McGill, we plan to consider as many staffing models as possible for digital scholarship support, of which Collections as Data is a key part. To date, I have identified the following for further exploration.

Subject Liaisons. Currently the library has a large number of librarians assigned to specific departments to meet faculty research and teaching needs. In a Collections as Data context, these individuals would be knowledgeable about the data sets available in the field and where to look for them, able to perform basic tasks in software tools commonly used with data sets in the field, and be familiar with the type of research done with Collections as Data for the discipline.

Research Support Librarians. While the subject liaisons are our primary user-facing librarian staff at McGill, we also employ a small number of functional liaisons, for scholarly communication, copyright, and research data management. One or more research support librarians would focus on digital data, methods, and tools and how they might be used across disciplines. They would understand the current state of the art in Collections as Data, along with its major projects, people, and trends. Digital research support librarians can be found placed either with other functional liaisons in digital scholarship units or as a group in a public services unit as distinct from librarians who focus on teaching and learning.

Student Employees. Both graduate students and undergraduate students are commonly used in digital scholarship centres. They perform a wide variety of tasks from assisting with physical hardware to providing support for digital tools. In a Collections as Data context, they would likely be knowledgeable about several specialized software packages offered by the library, such as NVIVO, ArcGIS, SPSS, or Photoshop, and be on hand to assist users with data analysis and use of these tools.

Fellows & PostDocs. This model is used by academic libraries to bring in credentialed researchers for a dual role of advancing their own research and providing library services to users. They have the benefit of deep knowledge of tools, methods, and data sources in their fields, but not necessarily in other knowledge areas. This model is best used in an environment where one general area of study is prominent or receiving special attention in a given time frame. Researchers may give more weight to the counsel of a research fellow or PostDoc than to a librarian given the former's academic credentials.

Departmental "Champions". Perhaps the most effective advocates for library services to researchers are the researchers themselves. Finding the researchers who are already doing the activity a library service supports, be it open access publishing or text mining, and collaborating with them to get the word about their research, has proven effective in many fields, and would likely be so for Collections as Data.

Artists. Similar to Fellows and PostDocs, libraries occasionally bring in talented professionals to support cutting-edge services. Artists of any type—writers, visual artists, musicians, stage performers—could create highly innovative products from Collections as Data which would catch the eye of researchers, make them aware of the underlying data and library services, and spark their own ideas.

McGill will likely use a mixture of staff in our Fiat Lux service model. We believe Collections as Data is a key part of our mission to build collections and make knowledge available to researchers and citizens alike. We must act to ensure our campus researchers can make the best use of our efforts.

"Collections as Data" as Simile: Poetic Effects and Contextual Corollaries of the Project

Devin Becker, University of Idaho

My friend Zanni introduced me to the writer Quan Barry's book *Controvertibles* after one of our poetry workshops at UC Irvine many years ago.[1] I was astonished by the poems, each of which connects, via its title, two subjects using an "as" simile that it then draws and extrapolates from in varying degrees. These titular juxtapositions range from the pop cultural–"Nick Drake's "Pink Moon" as Infatuation," "Rage Against the Machine as Plate Tectonics" – to the political – "Richard Nixon's 1972 Christmas Bombing Campaign as Gospel" – to the obscure – "Pernkopf's Atlas as Interregnum," each poem triangulating a reader's experience among the dual subjects of the title and the poem's subject matter. This makes for a particularly personal experience of the poems, with a reader's relationships (or lack thereof) with each of the title's subjects brought into concert with the poem itself and extending as one progresses throughout the book.

Barry's poems work this way because our minds are conditioned to make meaning through comparison and, due to the routine experience of this process, to draw particular pleasure from new and unusual connections and subjects, from, as another poet puts it, "[a]ll things counter, original, spare, strange."[2] The simile at the heart of Collections as Data works in the same way. That is to say it works poetically, playing off personal and historical understandings of "collections" and "data" to create possibilities for both individual interpretations and communal extensions of its meaning. Further, the "collections as data" simile achieves this poetic resonance because it presents a juxtaposition that is arresting and unique, while seeming at the same time inevitable, true.

As a librarian concerned with building digital collections and exhibits, I connected with the "collections as data" phrase immediately, intuiting the project's aims from its title even before reading many of its outputs. My colleagues Evan Wililamson and Olivia Wikle, with whom I develop the digital exhibit framework CollectionBuilder, had similar reactions. We all just *got* the idea behind the simile and project, and we quickly made it our own, using, in the same way scholars have used poetic concepts for centuries, the phrase's connections and connotations to promote and prod our own work and ideas: We titled the data output section featured on the homepage of CollectionBuilder templates "Collections as Data," and we shamelessly borrowed the simile when naming our related oral history platform, Oral History as Data.

A Collections as Data section from the CollectionBuilder digital exhibit framework. This collection of multiple data outputs from a collection is featured as a default on all CollectionBuilder websites.

Using the static site generator, Jekyll, CollectionBuilder creates templates for expressing collections through their data as web exhibits, allowing a user to create visualizations, browsing/searching mechanisms, and multiple data outputs of their collection by uploading a spreadsheet (CSV) and a folder of files into a cloned GitHub repository.[3] Through this process, the collection, both the objects of which it consists and the metadata describing those objects, is leveraged as data in a variety of ways to create both the underlying HTML, CSS, and JavaScript files that enable its interactive presentation and the multiple data outputs (CSV, JSON, GEOJSON) that express various numerical, geographical, temporal, and subject-based features inherent within the information it records. Further, each of these data outputs is found in the same location for each CollectionBuilder collection and described by a Frictionless Standards' Data Package template,[4] this being a standard architecture and description that enables machine reading and data harvesting across similarly constructed collections. Finally, each item page's meta markup is imbued with the item data as Dublin Core, Open Graph, Schema.org and JSON-LD formats, allowing for machine reading, harvesting and better indexing and search engine optimization for item pages.

In this way, CollectionBuilder embodies many of the Collections as Data principles that inspired its development. In aggregate, this focus can enable the computational leveraging many of those connected who contributed to the drafting of those principles envisioned from the movements outset, but we have increasingly been concerned and encouraged by the principles' application to individual collections themselves. While creating open data outputs for collections facilitates the preservation and reuse of collections and is thus a solid goal in itself,[5] CollectionBuilder also demonstrates how embodied Collections as Data principles can also lead to increased contextual possibilities for individual collections.

Some of what gets lost in the Collections as Data project's inevitable emphasis on the "data" side of its phrasing is the power of collections themselves. As those who work with archival collections often notice, collections can come to mean more than the sum of their objects (or their data, if you will), revealing stories and insights about their subject matter, creators, and stewards through the proximal relations of their items. What is more, as we've seen via collections developed by CollectionBuilder users, these collections need not possess grand historical significance to generate this effect; a well-described collection of family recipes, for instance, can tell a resonant story about family, regional cuisine, and spice tolerances. The facilitation of this effect has figured prominently in recent CollectionBuilder development and writing on the project. Evan, Olivia and I have paired our focus on "collections as data" with another complementary development principle, "collections in context," to ensure we emphasize CollectionBuilder's interpretive possibilities as much as its computational ones.

We have found that these two principles work well together both in rhetoric and in practice. CollectionBuilder's interactive features visually reward the metadata work of stewards and creators and often lead to their uncovering new perspectives on the collections with which they work. Our interpretive features then allow them to communicate these discoveries multimodally, writing with and into the collection in ways that hopefully, together with the data-driven features, inspire future

scholars, readers, and browsers to investigate and interpret the collection further. In sum, this, too, this discovery-driven cycle of expression, interpretation, and presentation we intend CollectionBuilder to enable, is an effect and desired outcome of the Collections as Data project and idea. My great hope for CollectionBuilder then is that it might have a similarly resonant effect for others as Collections as Data has had for us, and that the work we have done and the digital exhibits developed because of that work inspire those designing the technical and intellectual frameworks of digital collections in the future to embrace both sides (i.e. the "negative capability") of the simile that drives the Collections as Data movement.

Notes and Works Cited

[1] Barry, Quan. Controvertibles. University of Pittsburgh Press, 2004.
[2] Hopkins, Gerard Manley. "Pied Beauty." *Poems of Gerard Manley Hopkins*. Humphrey Milford,1918.
[3] For more on how to build CollectionBuilder exhibits, see our documentation.
[4] This "containerization format" is described in a Forum 1 Collections as Data position paper entitled titled "Frictionless Collections Data" by Dan Fowler.
[5] This itself is a tremendous gain for digital collections as we have seen many digital projects and collections created on database driven platforms during the past 25 years lost to institutional neglect, security concerns, and/or software updates.

Sovereignty as Data? Indigenous Knowledges and Library Collections

Kayla Lar-Son, X̲wi7x̲wa Library, University of British Columbia

Under the colonial influence of intellectual heritage, non-western thought was devalued and misinterpreted. Indigenous Knowledges and knowledge transmission were deemed to be inferior when compared to western traditions. Educational institutions, guided by a western bias, transformed the plurality of knowledge systems into a hierarchy of knowledge systems. When knowledge plurality mutated into knowledge hierarchy, the horizontal ordering of diverse but equally valid systems was converted into a vertical ordering of unequal systems, and the epistemological foundations of western knowledge were imposed on non-western knowledge systems with the result that the latter were invalidated (Shiva, 2000).

In western epistemologies and ontologies, knowledge is patriarchal, racialized, possessive, and extractive. As libraries sit in a position where we steward intellectual heritage we must consider and address the following;

- Knowledge and the control of knowledge = **Power.**
- Libraries and librarians sit in the position to continue the misinterpretation and dismissal of knowledges while being guided by Western bias and education.
- What unlearning must take place, and where does this unlearning start?

Indigenous Data, Cultural, and Sovereignty

The right for Indigenous communities to define their kin systems, recognize legal traditions, maintain self-governance, control their own education is important for Indigenous communities to remain as sovereign entities. Indigenous conceptions of Knowledge do not follow the same trajectory of western understandings of knowledge and knowledge transmission. Indigenous Peoples do not focus their Knowledge systems in binaries. Understandings of relationality, kinship, and governance are rarely learned from within institutional walls, and seldom from academic texts. It is learned on the Land, listening to stories, and from close connections to one's community. According to Tenille Marley when “Indigenous sovereignty is threatened when cultural sovereignty is violated through exploitation, theft and misrepresentation of culture (2020).” Cultural sovereignty is essential for the survival of Indigenous Nations culture, including language, songs, oral histories/stories, dress, ceremonies, dances, etc., as well as land and community are at the core of Indigenous identity and way of life.

Debates about ‘data sovereignty’ have been dominated primarily by non-Indigenous governments and corporations for legal reasons. Until recently, Indigenous peoples have long been left out of the conversation relating to the collection, ownership and application of data about their people, lifeways and territories (Kukutai and Taylor, 2016). Indigenous peoples have now begun to recognize that data is a resource that has value. Currently, the main issues with Indigenous data, pertain to control, ownership, and power dynamics.

Indigenous data sovereignty centers Indigenous peoples collective rights to data about people, territories, lifeways and natural resources, and is supported by Indigenous peoples' inherent rights of self-determination and governance (Walter & Sunia, 2019). In essence Indigenous data sovereignty puts forward the following principles;

- Indigenous nations have the right to ownership and governance over data about them, regardless of where it is held and by whom;
- Indigenous nations have the right to govern data in a way that aligns with their own data protocols and laws;
- Indigenous Peoples also have the right to access data that supports nation re-building. This often includes access to government documents both historic and contemporary and archival documents.

Transforming library futures to uphold Indigenous Sovereignty

As more Universities are focusing on incorporating Indigenous Knowledges and perspectives into research, curriculum development, and strategic initiatives. Libraries are being positioned to evaluate their collections for any misinterpretation of Indigenous Knowledges and legalized theft, condoned by copyright laws, by non-Indigenous researchers. In addition, libraries are also being called upon to steward community produced collections that include community protocols, Traditional Knowledges, and align with community specific Worldviews. Most libraries are not fully equipt to steward these collections as there is a lack of understanding and training on what Indigenous sovereignty looks like within a library. In able to transform libraries to be accountable to Indigenous communities the following needs to be considered and discussed;

- How do libraries uphold the principles of Indigenous Data and Knowledge Sovereignty?
- What do responsive library policies look like in regards to Indigenous Knowledge misinterpretation, dissemination, protection, and consultation?
- How do we build culturally relevant and responsible collections, by, for, and with Indigenous peoples? What does this look like in terms of relationship building?
- How do libraries engage with the concept of ownership from a non-western understanding?

Lastly, I challenge libraries to think about and reflect on the following questions. **Why are and should libraries be acting as stewards of Indigenous Knowledges?**

Works Cited

First Nations Information Governance Connect. 2016. "Pathways to First Nations Data' and Information Sovereignty." In *Indigenous Data Sovereignty Towards and Agenda,* eds. Tahu Kukutai and John Taylor. Australian National University Press, 130:155.

Kukutai, Tahu & Taylor, John. 2016. "Data Sovereignty for Indigenous Peoples: Current Practices and Future Needs." In *Indigenous Data Sovereignty Towards and Agenda,* eds. Tahu Kukutai and John Taylor. Australian National University Press, 1:11.

Marley, Tenille. 2020. “Indigenous Data Sovereignty and the role of universities.” In Indigenous Data Sovereignty and Policy 1st ed. Walter, M., Kukutai, T., Carroll, S.R., & Rodriguez-Lonebear, D. (Eds.), New York: Routledge, 157:168.

Shiva, Vandana. 2000. "Foreword: Cultural Diversity and the Politics of Knowledge," forward in *Indigenous Knowledges in Global Contexts: Multiple Readings of Our World*, ed. George J. Sefa Dei, Budd L. Hall, and Dorothy Goldin Rosenberg. Toronto: University of Toronto Press, vii-x.

Walter, Maggie & Suina, Michele. 2019. “Indigenous data, indigenous methodologies and indigenous data sovereignty,” *International Journal of Social Research Methodology*, 22:3.

Strategic Pessimism and the Sustainability of Digital Collections as Data

Miguel Escobar Varela, National University of Singapore

Scholars of religion often talk of a "methodological agnosticism" (Porpora 2006). Even if a scholar herself is a religious person, assuming an agnostic position is a productive stance. The digital humanities are, in my experience, characterized by a contagious and inspiring sense of optimism. This attitude often drives innovation, fosters community building, and encourages researchers to explore new methods and technologies in their work. But I have found that this optimism often creates problems for research teams at the beginning of a research project. For example, overly confident assumptions about funding, resources, and project timelines can lead to difficulties down the line. As The Ending Projects (2023) notes, sustainability needs to be planned for from the beginning, rather than as an afterthought. My suggestion is simple: assume things will fail. Even if you, like me, consider yourself an optimist, you can deploy a stance of strategic pessimism to make projects more accessible and more sustainable. This approach acknowledges the realities of research projects, such as funding limitations and unexpected challenges, and it helps to ensure that cultural collections remain available as data in the future.

I am writing here with small-scale, researcher-led and grant-funded projects in mind. In this contexts, it is prudent to assume that institutional funding may not be available in the future. By adopting a pessimistic stance, we can prioritize the preservation and accessibility of our digital collections, ensuring that they remain valuable and usable for future research. This approach helps to safeguard the investment of the time, effort, and resources that have been made in creating these collections. In this context, strategic pessimism encourages us to plan for the worst-case scenario and explore alternative solutions to preserve our digital collections.

One way to implement strategic pessimism is to develop sustainability tiers, which prioritize the preservation of the most important data. These tiers can be based on factors such as the uniqueness of the data, its potential value to future research, and the feasibility of preserving it. By identifying these tiers, we can allocate resources and adopt preservation strategies that ensure the future viability of our digital collections. In the face of uncertainty, this tiered approach allows us to focus on what is most essential, safeguarding our research and scholarly endeavors. We can then maximize the impact of our preservation efforts, while also minimizing the potential loss of valuable data. In other words, some aspects of a project are essential and must be preserved even if everything else is lost. In my book *Theatre as Data* (whose title is a direct reference to this community's work) I suggest that we should think of this essential data as the "preservation core" of a project (Escobar Varela, 2021). Having grown up in an earthquake-prone city, I learned to keep crucial documents in a sealed plastic bag by the door of my house. In case of an emergency, I could take these documents and run. The same principle applies to the design of a digital preservation core. A second basket will include digital artifacts that can be kept in case we have a bit more time to run away before the building (or the funding) comes crumbling down. This tiered system of preservation provides a flexible and adaptable framework for addressing the

challenges and uncertainties associated with digital humanities research, while following the recommendations of the Digital Curation Centre (Higgins 2008).

We often don't want to believe that disaster will strike and when it does, we are caught off guard. I know of at least one project where the digital architecture collapsed overnight, when there was a mandatory update from the base system they were using. Since there was no multitiered approach to preservation it was very difficult for the team to recover the functionality quickly. More importantly, this sudden collapse imperiled the long-term preservation of the project. Data marked for long-term preservation should be stored according to explicit data models and ideally made openly accessible. Both preservation and openness depend on each other. Open sharing requires sustainability practices. But in turn, resources that are openly shared are more likely to be preserved. This is one of the core principles of LOCKSS or "Lots of Copies Keeps Stuff Safe" (Maniatis et al. 2005).

To enhance the reusability and reproducibility of our digital collections, we should strive to share them in easily sustainable flat formats, such as CSV files. These formats are widely used and understood, making it easier for other researchers to access and work with the data. Flat formats promote accessibility and interoperability, allowing researchers to reuse and build upon existing data. This fosters collaboration and the exchange of ideas within the digital humanities community, supports effective teaching, and ultimately leads to more innovative and impactful research. In previous *Collections as Data* workshops, the importance of flat formats has been widely recognized. For example, Miriam Posner (2017) remarks that these formats are preferable to APIs since this is what many humanists are comfortable with, and "the availability of simple flat files can be the difference between using and not using a dataset". I would emphasize that simple flat files are also more sustainable and easier to maintain than APIs.

In this position paper, I want to strengthen the bridge between the need for accessible data formats and the importance of sustainability in digital humanities research. I argue that their adoption is not only beneficial for facilitating access but also crucial for ensuring the sustainability of our digital collections. By making data more accessible and easier to work with, we can increase the likelihood that it will be reused and preserved, ultimately contributing to the long-term success of the digital humanities field.

This position paper does not argue against more ambitious and optimistic projects, such as the development of live APIs that enable a wider range of use cases. Instead, it suggests that strategic pessimism can work in conjunction with these approaches, providing a safety net for our digital collections in the event of funding shortfalls or other unforeseen circumstances. By balancing our optimism with a healthy dose of strategic pessimism, we can ensure that our digital humanities projects are not only ambitious and innovative, but also resilient and sustainable in the face of the many challenges that cultural projects inevitably encounter.

References

Higgins, Sarah. 2008. "The DCC Curation Lifecycle Model." *International Journal of Digital Curation 3*, No. 1.

Ending Projects Team. 2023. *Endings Principles for Digital Longevity*, Version 2.2.1. Last modified March 3, 2023. https://endings.uvic.ca/principles.html.
Escobar Varela, Miguel. 2021. *Theater as Data: Computational Journeys into Theater Research*. Ann Arbor: University of Michigan Press.
Maniatis, Petros, Mema Roussopoulos, Thomas J. Giuli, David SH Rosenthal, Mary Baker, and Mary Baker. 2005. "The LOCKSS Peer-to-Peer Digital Preservation System." *ACM Transactions on Computer Systems (TOCS)* 23, No. 1: 2-50.
Porpora, D. V. 2006. "Methodological Atheism, Methodological Agnosticism and Religious Experience." *Journal for the Theory of Social Behaviour 36*, No. 1: 57-75.
https://doi.org/10.1111/j.1468-5914.2006.00296.x.
Posner, Miriam. 2017. "Actually Useful Collection Data: Some Infrastructure Suggestions." In Always Already Computational: Library Collections as Data: National Forum Position Statements. https://github.com/collectionsasdata/collectionsasdata.github.io/raw/master/aac_positionstatements.pdf.

Transforming "collections as data" into a force for good in Latin America and the Caribbean

Wouter Schallier, Chief of the Hernán Santa Cruz Library, ECLAC and Linn Enger Leigland, Associate Librarian, ECLAC

Disclaimer: The views expressed in this article are the personal views of the authors, and do not necessarily reflect the views of ECLAC/United Nations.

Collections as data represent opportunities for making traditional collections available in new formats, reaching a broader audience, offering new research insights, reducing traditional biases, enriching analytics, and facilitating new collaborations, amongst others. As such, we think it is vital for the Latin America and Caribbean region that it can reap the full benefits of the collections as data innovation. However, we cannot assume this will automatically be so.

We believe some actions need to be taken first, to make sure the data revolution benefits Latin America and the Caribbean and indeed leads to less inequality amongst regions globally as established in the UN Sustainable Development Goals (United Nations DESA, n.d.).

First, **we believe collections as data initiatives should aim to systematically build partnerships between the Global North and the Global South.** The inclusion of partners from the Global South could contribute to more collaboration across regions and the exchange of good practices and lessons learned. Broader diversity in terms of geography, project partners and gender can lead to more inclusiveness and a fairer allocation of resources. Collections with a broader set of perspectives provide a more comprehensive resource base for researchers and can contribute to reducing biases and generating new ideas. Opening the space for new partners, including those with limited experience, will gradually expand the circle of potential project partners and turn collections as data into a tool for innovation, inclusion, and capacity building, for all, and not just for a few.

Second, **digitization presents new opportunities for collaboration between libraries and publishers**. The technological revolution rapidly changes how we access, manage, and consume information. While publishers in Latin America and the Caribbean do not have the same visibility, negotiating power and technical infrastructure as the large, international publishers, they can capitalize on the data revolution and expand their reach in partnership with libraries in the region. It is urgent to radically modernize access to information by achieving maximum availability in digital formats and open access. We encourage publishers from the region to embrace these new opportunities for collaboration with libraries and explore new business models.

Third, **to empower people to become digital citizens (instead of just remaining digital consumers), fostering information and data literacy is key.** According to the UN Secretary-General, nurturing capabilities in analytics and data management is a success criterion to unlock the full potential of data to secure better decisions and support for people and the planet (United Nations, 2020). Data literacy

requires information, statistical and technical literacy. Fostering data literacy as a fundamental skill of digital citizenship will help citizens to separate facts from fiction. Imagine a reality where everyone can, in principle, critically and actively source information instead of being held at the whim of non-transparent algorithms and biases developed by a few providers. Capabilities to read, manage, analyze, and argue with data can facilitate better decision-making and transparency, expose bias and democratize the use of data. Data-literate citizens are better equipped to use collections as data for participation in society and innovation.

Fourth, **securing reliable access to the internet and digital technology is fundamental**. According to ECLAC, Latin America and the Caribbean are among the most unequal regions in the world (ECLAC, 2020), and the digital divide between the rural and urban populations is alarming. “While 68% of urban homes had internet access in 2018, the same held true for only 23% in the countryside” (ECLAC, 2023). Digital inequality is also gendered. ECLAC estimates that 40% of women in the region lack internet access (Ibid.). Poverty and lack of digital infrastructure, basic digital skills or a device keep people off the grid. Transforming collections as data into a force for good and leaving no one behind means, in effect, leaving no one offline.

Fifth, **data must comply with international standards** such as the FAIR Guiding principles for scientific data management and stewardship. Digitization alone is not sufficient. Data must be Findable, Accessible, Interoperable and Reusable (GoFair, n.d.), both for computers and humans. A challenge in Latin America is that much data does not fully comply with these principles, which limits its usability. For the Caribbean, the availability of data itself is a challenge. Supporting libraries in the region to design and implement institutional data management plans that consider these, and other international standards is fundamental to harnessing the transformation of collections as data.

As a conclusion, we are eager to collaborate with partners across the globe to further digitize traditional collections whenever possible, and contribute to improving access, availability, and the quality of collections as data. However, in order to make this transformation a force for good that benefits the region, we need to make sure these collections as data initiatives resolve some of the specific challenges in Latin America and the Caribbean instead of perpetuating them.

Sources:

GoFair (n.d.), “FAIR Principles”. Available at: FAIR Principles - GO FAIR (go-fair.org)

United Nations DESA (n.d.), “The 17 Goals”. Available at: THE 17 GOALS | Sustainable Development (un.org)

United Nations ECLAC (2023), “ECLAC Seeks to Bring More Women into STEM, Close the Digital Gap and Eradicate Gender Cyberviolence”. Available at: ECLAC Seeks to Bring More Women into STEM, Close the Digital Gap and Eradicate Gender Cyberviolence | Economic Commission for Latin America and the Caribbean (cepal.org)

United Nations ECLAC (2020), “New Document by ECLAC Examines the Structural Gaps that Characterize the Region”. Available at: New Document by ECLAC Examines the Structural Gaps that Characterize the Region | Economic Commission for Latin America and the Caribbean (cepal.org)

United Nations (2020), “UN Secretary-General Data Strategy”, Available at: UN Secretary-General’s Data Strategy

Unlocking the potential of biodiversity collections as data

David Iggulden, Royal Botanic Gardens, Kew & Biodiversity Heritage Library (BHL)

My work at a leading botanic garden library in the UK has shown me the significant potential for next steps in working with collections as data. More specifically, in my role as current Chair of the Biodiversity Heritage Library (BHL) Executive Committee, it has become increasingly clear that unlocking the data held within resources such as BHL will be key to fulfilling this potential, by facilitating data reuse and linked data.

The Biodiversity Heritage Library (BHL) is a global consortium of botanical and natural history institutions working together to digitise their collections and make them freely available at biodiversitylibrary.org. Simply put, BHL is an open access digital library for biodiversity literature and archives originally conceived in 2006, which provides open access to collections from the 15th-21st centuries. Most of the collections are out-of-copyright, but permissions have also been agreed with rights holders for an increasing collection of in-copyright content to be included.

At the Biodiversity Next conference in Leiden in 2019, there was a call for better data, better science and better policies with a clear target to reduce data silos. The aims were simple: to significantly enhance the efficiency of research and to review the landscape overall with a view to new collaborations and linkages allowing for enriched resources. The appetite and potential for more innovative work with collections as data was also clear throughout the conference.

None of us could have foreseen the approaching pandemic nor the impacts this would have on work with collections over the coming years. However, the loss of access to physical collections shone a spotlight on the collections data available online and, more significantly, its limitations, fragmented nature and omissions. For BHL, it also highlighted the value of the resource, how far we had come but also how far we still needed to go.

BHL's model as a global consortium that operates almost entirely via virtual collaboration was well suited to the pandemic scenario. The wide variety of partners from across the globe provided a collective resiliency to the challenges being encountered, and many scientists identified how BHL was key in simply allowing work to continue during this period.

The pandemic additionally highlighted how much more could be done to integrate BHL data with global communities. There remains a clear need to connect BHL data with museum deposit records, geographic collection site info and species descriptions. The existing geographic and author data held within BHL also needs to be further indexed and made discoverable at a granular level. Basically, BHL data needs to flow seamlessly to other organisations to enrich the worldwide network of biodiversity data.

Yet there are many other possibilities still to be considered. For example, how can we more effectively accommodate original materials within BHL to maximise their value for collections as data? A recent project at Kew in partnership with the UK's National Archives saw volunteers transcribing a volume and

then applying TEI (Text Encoding Initiative) XML markup to their transcriptions. BHL cannot currently display this markup which limits the value and potential reuse of this data, so this must be considered in plans for the future.

How can we further unlock data overall to support collections as data? A recent hackathon at Kew focussed on the digitised correspondence and specimen and illustration collections of the eminent 19th century botanist Joseph Hooker. The event showed just some of the many opportunities there could be for innovation in working with collections as data. Outputs included a search engine developed to find natural remedies for medical issues through plants based on the collections data and a virtual reality environment which allowed the user to step inside and explore a landscape from Hooker's painting of the Himalayas.

The potential for further work with biodiversity collections as data is undoubtedly significant, but this will require further breaking down of data silos, a concerted effort to catalogue and digitise collections, increased collaboration and integration between existing resources. As a high-profile resource now frequently identified as a key component in supporting biodiversity research, BHL is ideally placed to be an important partner in unlocking the potential of biodiversity collections as data.

Collections as data – questions which keep us up at night

Margaret Warren, Director, Digital Delivery, State Library of Queensland

Is 'Collections as data' in danger of being seen by some as this decade's 'linked open data' - seen outside the community involved and invested in it as arcane, and with limited tangible outcomes beyond some amazing case studies and examples of where it is being used by researchers in academic institutions? How can we change this? What language can we use to talk about the incredible usefulness and power of collections as data for research, new knowledge, creativity and the telling of new stories?

What is the role of collecting organisations inside and outside academic institutions in the next steps with collections as data? How can we further progress the ideas of digitised collections as being more than, or beyond, the replication of the physical experience in digital form? How do we make room for born-digital collections, and what have we learned from digitisation that will help us make born-digital collections useful as data?

How can we influence or advocate with LMS (Library Management System) vendors for tools and systems that support collections as data? I do not see this as the current state of play. LMS vendors are very invested in the search, find, view paradigm. The conversation needs to change if collections as data is to be amplified in our institutions.

How can we better understand and acknowledge bias, cultural sensitivities, Indigenous Cultural and Intellectual property (ICIP), and the data that is missing because in general digitised collections come from institutions with colonial histories and collecting patterns?

How do we keep the integrity of data that should be investigated and used for research and new knowledge, with context as paramount? Is this something that we are currently doing well? Or is it a point of concern?

Does collections as data have a place outside academic institutions? What might those places be? How can we facilitate that when most people working in this space are within the academic institutions?

How does the collections as data community move beyond a ‘cottage industry’ mindset to scale? Should it? Is there a tipping point beyond which data loses utility for progressing meaningful research and outcomes?

What is the significance of the growth in concerns in multiple spheres about artificial intelligence, including machine learning, on collections as data? How does the collections as data community

promote and demonstrate in an ongoing way ethically grounded approaches to computational work with collections?

Use and Communication of Collections as Data

Kayla Abner, University of Delaware (UD)

My introduction to collections as data came from two sources: discussing a potential position at UD tasked with creating a digital scholarship project from start to finish: selection and digitization of materials, creation of data for the project (collections as data), through to publication; and the work at the University of Utah on the Kennecott Miner Records [1] project. Throughout my exploration of collections as data, I have two unresolved questions:

1. How do we encourage use of our datasets for research and teaching? How can we teach the data literacy skills necessary to successfully use the data in a project?
2. When communicating the value of collections as data, how do we clarify what is metadata *about* the collection, and what is "data data," *derived from* the materials themselves?

Strategies to encourage use of collections as data

As Harriett Green suggested at the 2017 Collections as Data Forum, throwing the data up on a server (or on hard drives on a book cart) is not enough to get people to use it [2].

My current collections as data project involves creating network data derived from the letters of Alice Dunbar-Nelson, a 20th century African American writer, activist, and wife of prominent 19th century novelist Paul Laurence Dunbar. Because Alice was active in Delaware politics and social justice, this collection is often used in teaching and research at UD. The dataset itself is essentially item-level metadata, recording basic information on each letter. To expand the dataset, a graduate assistant from the English department is recording Alice's relationship to the sender, the topic of the letter (a controlled field), and a brief description of what is discussed. The expanded data could be used to illustrate Alice's personal and professional social network and understand her writing and activism work more broadly, both in network form and through other data visualizations.

Our teaching support at UD has explored an "assignment package" approach to facilitate integration of digital projects into courses [3]. Once the Alice Dunbar-Nelson data is complete, I will create an assignment package, including the final dataset, to guide students and instructors through creating a network graph. The packages include items like rubrics for assignment grading and a schedule for any library support sessions. The packages illustrate learning outcomes for the assignment and provide scaffolded support for completion; they are not meant to be taken and integrated in a course without library support. Paving the way for frictionless integration of data-driven scholarship is an attempt to rectify what Mia Ridge described at the 2017 Collections as Data Forum as a greater interest in *access* rather
than distant reading and data science [4]. We can generate interest in data-driven scholarship by highlighting pedagogical outcomes and providing scaffolded support.

Because my current work with collections as data is mostly teaching-focused, I prioritize collections that are heavily used in courses, while considering which are suitable for "datification." Materials

such as ledgers, ships' logs, typed political documents, and directories are more easily transformed into a dataset *derived from* the collection.

Metadata or "data data"?

When we say "library collections as data," I often think of projects like the Kennecott Miners Records and my own correspondence data from Alice Dunbar-Nelson. However, "library collections as data" can refer to metadata *about* the collection, or data *derived from* the collection. This is an important distinction that affects both how we as practitioners understand the goal, and how we communicate the value to our patrons. Even the term "metadata" can be illegible to someone outside the information professions, or even practiced differently among these professions.

I have attempted to analyze our museums' general art collection to clearly identify where the collection has strengths and gaps, especially in regards to racial and gender representation. However, my colleagues in museums do not use a controlled "subject" field (such as the CDWA "subject matter" [5]) to describe the work itself. While I could endeavor to identify which percentage of our works were created by women, for example, I could not identify which works by women are about the experience of women or women's issues. In some ways, the description that we as librarians do is at cross-purposes with museum professionals' work, both because art is up to interpretation and because scholars often do not expect subject headings while conducting research with museum collections.

I would encourage the community to think about ways to distinguish metadata and "data data." Should this data be packaged and promoted together? Who is most negatively affected if we do not define different data sources? And, how can we use a clear distinction to our advantage when communicating with researchers and even colleagues?

References

[1] Marriott Library, University of Utah. "Transcribed Data from the Kennecott Miner Records Collection." GitHub, 2020. https://github.com/marriott-library/ KennecottMinerRecords.
[2] Green, Harriett. *Book carts of Data: Usability and Access of Digital Content from Library Collections.* Always Already Computational: Collections as Data National Forum, 2017. https://zenodo.org/record/3066161#.Y9FmT-zMI-Q

[3] https://dlfteach.pubpub.org/pub/vol3-abner-etal-timelinejs/release/2 Abner, Kayla, Morgan, Paige, & McCollom, Amanda. *Creating Digital Stories with TimelineJS.* #DLFteach Publications, 2022. https://doi.org/10.21428/65a6243c.5e643ac8

[4] Ridge, Mia. *From libraries as patchwork to datasets as assemblages?* Always Already

Computational: Collections as Data National Forum, 2017. https://zenodo.org/record/3066161#.Y9FmT-zMI-Q

[5] Getty Institute. *Categories for Descriptions of Works of Art. 16. Subject Matter.* Getty, 2019. https://www.getty.edu/research/publications/electronic_publications/ cdwa/18subject.html.

Web Collections as Training Data: Bringing together Archival Provenance and Ethical AI Approaches to Dataset Documentation

Emily Maemura, University of Illinois Urbana-Champaign

My research has studied the processes of creating and using web archives, or historical snapshots of web pages and websites, like those browsable through the Internet Archive Wayback Machine. I've conducted ethnographic fieldwork focusing on the development of research infrastructures to support data-centered uses of web materials collected through web archiving programs at a national library and university libraries. My work focuses on born-digital materials which are collected with the aid of crawlers and bots; in contrast, many of the previous cohorts of the *Collections as Data* project have focused on digitization and other processes used to make collections of print and analog multimedia materials machine-actionable. As a provocation here, I invite the community to consider in greater depth and detail the under-studied material transformations and investment of resources that turn born-digital and born-networked data into "collections."

Engaging with born-networked data requires attention to the ways that the "always already computational" nature of these materials can become invisible and taken-for-granted. I ground my work in the precedents set by Geoff Bowker and Susan Leigh Star (2000) whose approach to Infrastructure Studies aims to bring attention to the standards, processes, and labour that often recede to the background of information infrastructures. For my work with web archives, this discussion of the taken-for-granted centres on the ways in which web materials are inherently tied to the web and Internet's protocols and standards, and can only exist within a finite, predetermined set of classification schemes. Web materials enter the archive pre-arranged according to these classifications which allow for easy sorting of data by HTTP status code, web domain, as text which adheres to a standardized character encoding scheme that allows for search, filtering, and other analysis of manipulation. Yet, as addressed in Maemura (2023), significant labour and resources are also expended to rearrange web data into new material forms, such as text corpora which are ordered according to new classification schemes that align better with researcher wants and needs. Taking an infrastructural approach, I emphasize the sociotechnical nature of web archives creation and use, which requires both an understanding of how data are created in the context of technical systems that impose or manipulate data structures, as well as the influences of organizational policies and legal restrictions upon data transformation, movement, and analysis. For researchers studying web archives as data, it's important to understand where and when a decision or circumstance determines that certain data would be included or excluded. Maemura et al. (2018) begins to develop a framework for tracing the provenance of these collections in sociotechnical terms.

Beyond web archives, the provenance, origins of, and decisions for curating data have recently become the focus of discussion for another kind of text corpus or dataset of web materials. While the web archives I study focus on preserving the web as an essential part of cultural heritage, similar methods of web crawling serve as the basis for generating datasets of web materials for work in data science,

machine learning and AI. For instance the Large Language Model (LLM) GPT-2 uses a training dataset of "8 million web pages" (Radford et al., 2019). The creators from OpenAI (whose subsequent work developed the recently released ChatGPT) describe a selection process which includes "only pages which have been curated/filtered by humans—specifically, we used outbound links from Reddit which received at least 3 karma" (Radford et al., 2019). Scholars in Critical Data Studies have taken issue with the way these web-derived datasets are used in machine learning since the lack of diversity of this data can "encode the dominant/hegemonic view, which further harms people at the margins," perpetuating biases with real world effects (Bender et al., 2021 p. 613). Although the digital collections of cultural heritage aren't instrumentalized in the same way as LLM training data, I believe we share an interest in curation decisions made upon data, and need to similarly consider the biases and dominant views embedded in datasets from the web.

Critical data scholars have therefore proposed approaches to document and trace the processes of dataset creation, manipulation and use for work in machine learning, with approaches like Datasheets for Datasets (Gebru et al., 2021). Given that web archiving and machine learning share similar concerns around documenting datasets (and work like Jo and Gebru, 2020, already identifies possibilities for bridging concepts between these fields) I ask how the Collections as Data community can contribute to these conversations on approaches to collection development and documentation that consider the biases and harms embedded in born-networked machine-actionable datasets. In particular, this attention to dataset ethics can lead us to ask: what are the politics of classification schemes that we take for granted? What hegemonic views are embedded in both data artifacts and curation practices? While the field of critical data studies has begun to address the key concerns and human impacts of a data-driven future, I believe the collecting work of libraries and archives must consider the impacts of a data-driven understanding of the past, and a society whose history is shaped by the legacies of our current digital curation practices.

References

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Bowker, G. C., & Star, S. L. (2000). *Sorting Things Out: Classification and Its Consequences*. MIT Press.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., III, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM, 64*(12), 86–92. https://doi.org/10.1145/3458723

Jo, E. S., & Gebru, T. (2020). Lessons from archives: Strategies for collecting sociocultural data in machine learning. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 306–316. https://doi.org/10.1145/3351095.3372829

Maemura, E. (2023). All WARC and no playback: The materialities of data-centered web archives research. *Big Data & Society*, *10*(1), 20539517231163172. https://doi.org/10.1177/20539517231163172

Maemura, E., Worby, N., Milligan, I., & Becker, C. (2018). If These Crawls Could Talk: Studying and Documenting Web Archives Provenance. *Journal of the Association for Information Science and Technology, 69*(10), 1223–1233. https://doi.org/10.1002/asi.24048

Radford, A., Wu, J., Amodei, D., Amodei, D., Clark, J., Brundage, M., & Sutskever, I. (2019, February 14). *Better Language Models and Their Implications*. OpenAI. https://openai.com/blog/better-language-models/

Towards a Praxis of Radical Empathic Design

Summer Hamilton, Pennsylvania State University

For my first job out of undergrad, I worked as a systems analyst at a now defunct technology outsourcing company. One of my major projects involved transitioning the department to a new IT service management system and providing end-user training and ongoing support. This meant learning the business processes as well as the out-of-the-box product functionality to effectively communicate customizations to the developers. To learn the business processes, I decided to conduct interviews with the employees who would be using the new system.

I spent hours with people, glancing at cubicle walls decorated with photos of children and grandchildren, as they discussed purchasing, inventory, or sometimes their grandchildren. The non-technical employees were often nervous–-concerns of job security frequently surfaced. Would they be able to use the new system? Would they be replaced by the new system? After two years, I left the outsourcing company to work in the financial aid systems department at a university. There, I performed the same sort of work with a new product and new processes. And during my interviews, I encountered a nearly identical set of concerns and fears.

Empathic design is the name given to the practice of listening to and factoring in the user's feelings during the preliminary stages of a project. It's an approach that has only recently gained traction. Indeed, it ran counter to the culture at my first job. An article in the *Journal of Engineering Design* from 2009 introduces one of the first "framework[s] for applying empathy in design." At that time, empathic design was simply a way to better sell products; more recently, it has taken a humanistic turn. A 2021 *Sustainability* article recognized empathy as "play[ing] a central role" in crafting designs that improve the well-being of communities. And though empathy wasn't discussed in any of my computer science classes, recent reports indicate that this, too, is changing.

This new direction supports my findings from those interviews. By listening to people, I not only could learn their processes but also their frustrations, hopes, and fears. As a result, I could design a system better suited to the user's style which increased their comfort, minimizing both anxiety and disruption. Trust was built and the work became communal when people felt integral to the design process. What I heard also informed how I crafted my documentation and structured the training.

I left the technology industry to pursue my PhD in English and am now a Digital Projects Designer within the Digital Liberal Arts Research Initiative at Pennsylvania State University. In this role, I support the Center for Black Digital Research, and I conduct my own research on African American literature. Thus, I find myself in the role of both designer *and* user. As I prepare to attend this collective gathering regarding the transition to a new system—I use "system" interchangeably hereon to refer to both a digital tool and a set way of working—I draw upon previous design experience while being cognizant of my own anxiety as a user. Pondering how we might employ a radical empathic design approach, I pose the following questions:

How do we select and design a system that fits the style and comfort of the user? If the "key stakeholders" present—as referenced in the Collections as Data abstract—cannot physically include everyone, what are ways to expand the voices included during the design stage? How do we avoid being like so-called allies who speak for others, assuming that we know what users know and what they need?

How do we ease the anxiety (and the precarity) of transitioning to collections as data? When practicing radical empathic design, we are motivated to increase well-being and reduce the threat of loss—loss of time, job, opportunities, joy, etc. Consider graduate students or researchers in precarious academic employment or with finite amounts of time or resources to complete their research projects. Consider scholars with marginalized identities. Though much progress has been made, digital projects are still not as widely accepted as traditional scholarship. What can we do to ensure all who desire can obtain the skills needed to engage in some level of computational research in a reasonable amount of time?

Tangentially, what of the precarity of the research subjects affected as traditionally underrepresented narratives are transformed into data? I am struck by the words of Ann Petry whose reason for writing *The Street* was to show what the data could not about "the high crime rate, a high death rate…. There are no statistics in *The Street*, though they are present in the background, not as columns of figures but in terms of what life is like for people who live in overcrowded tenements." Might an empathic approach also reduce what is lost when people and their stories are once again flattened into data?

How do we provide training in ways that make the system accessible to all? The Santa Barbara Statement on Collections as Data discusses making "a range of accessible instructional materials and documentation." This is encouraging and vital, and yet, I wonder if hearing from end-users earlier in the process might improve the range and accessibility of the instruction. Furthermore, how might early engagement allow us to deliver training in ways that meet various learning styles and needs?

Having reviewed the input/output of the previous *Collections as Data* meeting, I feel optimistic in bringing these questions forward because related concerns regarding user comfort and access are already a part of the conversation. I am excited about the direction of Collections as Data in developing and adopting radical best practices to ensure we are not creating divides while making advances.

Bibliography

Afroogh, Saleh, et al. "Empathic Design in Engineering Education and Practice: An Approach for Achieving Inclusive and Effective Community Resilience." *Sustainability*, vol. 13, no. 7, 7, Jan. 2021, p. 4060. www.mdpi.com, https://doi.org/10.3390/su13074060.

Current Biography, 1946: Who's News And Why. "Ann Petry." New York : H.W. Wilson, 1947. *Internet Archive*, http://archive.org/details/currentbiography1946unse.

Kouprie, Merlijn, and Froukje Sleeswijk Visser. “A Framework for Empathy in Design: Stepping into and out of the User’s Life.” *Journal of Engineering Design - J ENGINEERING DESIGN*, vol. 20, Oct. 2009, pp. 437–48. *ResearchGate*, https://doi.org/10.1080/09544820902875033.

Toddlers to teenagers: AI and libraries in 2023

Mia Ridge, British Library

My favourite analogy for AI / machine learning-based tools[38] is that they're like working with a child. They can spin a great story, but you wouldn't bet your job on it being accurate. They can do tasks like sorting and labeling images, but as they absorb models of the world from the adults around them you'd want to check that they haven't mistakenly learnt things like 'nurses are women and doctors are men'.

Libraries and other GLAMs have been working with machine learning-based tools for a number of years, cumulatively gathering evidence for what works, what doesn't, and what it might mean for our work. AI can scale up tasks like transcription, translation, classification, entity recognition and summarisation quickly - but it shouldn't be used without supervision if the answer to the question 'does it matter if the output is true?' is 'yes'.[39] Training a model and checking the results of an external model both require resources and expertise that may be scarce in GLAMs.

But the thing about toddlers is that they're cute and fun to play with. By the start of 2023, 'generative AI' tools like the text-to-image tool DALL·E 2 and large language models (LLMs) like ChatGPT captured the public imagination. You've probably heard examples of people using LLMs as everything from an oracle ('give me arguments for and against remodeling our kitchen') to a tutor ('explain this concept to me') to a creative spark for getting started with writing code or a piece of text. If you don't have an AI strategy already, you're going to need one soon.

The other thing about toddlers is that they grow up fast. GLAMs have an opportunity to help influence the types of teenagers then adults they become – but we need to be proactive if we want AI that produces trustworthy results and doesn't create further biases. Improving AI literacy within the GLAM sector is an important part of being able to make good choices about the technologies we give our money and attention to. (The same is also true for our societies as a whole, of course).

Since the 2017 summit, I've found myself thinking about 'collections as data' in two ways.[40] One is the digitised collections records (from metadata through to full page or object scans) that we share with researchers interested in studying particular topics, formats or methods; the other is the data that GLAMs themselves could generate about their collections to make them more discoverable and better connected to other collections. The development of specialist methods within computer vision and

[38] I'm going to use 'AI' as a shorthand for 'AI and machine learning' throughout, as machine learning models are the most practical applications of AI-type technologies at present. I'm excluding 'artificial general intelligence' for now.
[39] Tiulkanov, "Is It Safe to Use ChatGPT for Your Task?"
[40] Much of this thinking is informed by the *Living with Machines* project, a mere twinkle in the eye during the first summit. Launched in late 2018, the project aims to devise new methods, tools and software in data science and artificial intelligence that can be applied to historical resources. A key goal for the Library was to understand and develop some solutions for the practical, intellectual, logistical and copyright challenges in collaborative research with digitised collections at scale. As the project draws to an end five and a half years later, I've been reflecting on lessons learnt from our work with AI, and on the dramatic improvements in machine learning tools and methods since the project began.

natural language processing has promise for both sorts of 'collections as data',[41] but we still have much to learn about the logistical, legal, cultural and training challenges in aligning the needs of researchers and GLAMs.

The buzz around AI and the hunger for more material to feed into models has introduced a third – collections as training data. Libraries hold vast repositories of historical and contemporary collections that reflect both the best thinking and the worst biases of the society that produced them. What is their role in responsibly and ethically stewarding those collections into training data (or not)?

As we learn more about the different 'modes of interaction' with AI-based tools, from the 'text-grounded', 'knowledge-seeking' and 'creative',[42] and collect examples of researchers and institutions using tools like large language models to create structured data from text,[43] we're better able to understand and advocate for the role that AI might play in library work. Through collaborations within the *Living with Machines* project, I've seen how we could combine crowdsourcing and machine learning to clear copyright for orphan works at scale; improve metadata and full text searches with word vectors that help people match keywords to concepts rather than literal strings; disambiguate historical place names and turn symbols on maps into computational information.

Our challenge now is to work together with the Silicon Valley companies that shape so much of what AI 'knows' about the world, with the communities and individuals that created the collections we care for, and with the wider GLAM sector to ensure that we get the best AI tools possible.

[41] See for example *Living with Machines* work with data science and digital humanities methods documented at https://livingwithmachines.ac.uk/achievements

[42] Goldberg, "Reinforcement Learning for Language Models."

[43] For example, tools like Annif https://annif.org, and the work of librarian/developers like Matt Miller and genealogists. Little, "AI Genealogy Use Cases, How-to Guides." Miller, "Using GPT on Library Collections."