Always Already Computational: Library Collections as Data

National Forum Position Statements

March 2017

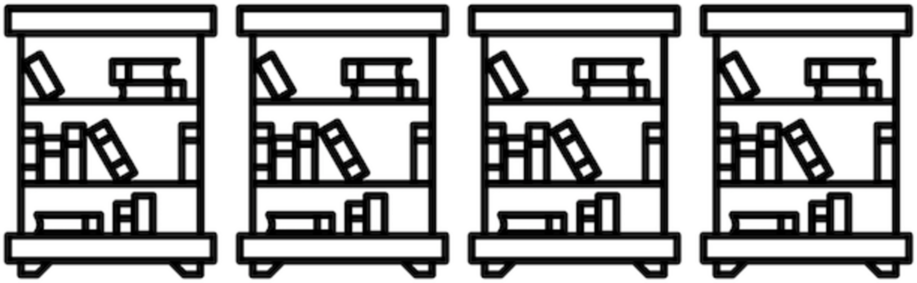


Image Credit - Library by Nikita Rozin

Jefferson Bailey
Alexandra Chassanoff
Tanya Clement
P. Gabrielle Foreman
Dan Fowler
Harriett Green
Jennifer Guiliano
Juliet L. Hardesty
Christina Harlow
Greg Jansen
Matthew Lincoln
Alan Liu
Richard Marciano
Matthew Miller
Labanya Mookerjee
Anna Neatrour
Miriam Posner
Sheila Rabun
Mia Ridge
Hannah Skates Kettler
Ben Schmidt
David Seubert
Laila Shereen Sakr
Tim Sherratt
Timothy St. Onge
Santi Thompson
Kate Zwaard

#aacdata

The position statements that follow were prepared in advance of the Institute of Museum and Library Services supported Always Already Computational: Library Collections as Data national forum.

Participants were asked to respond to the following prompt:

> *Leading up to the forum, [we] ask that you write a brief position statement derived from direct or related experience salient to the scope of work described in Always Already Computational. We welcome bridging, divergence, and provocation. Is there something concrete or conceptual we are missing? Are there projects and initiatives this work should be connected to? Are there questions and communities we aren't currently considering? This is an opportunity to highlight aspects of your experience that relate to the project and will to some extent help stage interaction at the face-to-face meeting - and beyond - as the project team works to iteratively refine forum outputs in a range of professional and disciplinary communities.*

Perspectives represented in the position statements highlight the many directions collections as data work could go. The statements will certainly inform the work of the forum, and consequently the iterative community based development of project outcomes.

Thomas Padilla
Laurie Allen
Stewart Varner
Sarah Potvin
Elizabeth Russey Roke
Hannah Frost

Pseudodoxia Data: our ends are as obscure as our beginnings

Jefferson Bailey, Internet Archive

In his meditation on oblivion and regeneration, W.G Sebald writes, "on every new thing there lies already the shadow of annihilation." Contemplating collections as data evokes a similar correlation -- one where transformation ("this as that") is less a process of alteration and more one of extraction of key, but possibly opaque, preexistent characteristics ("these from those"). When we consider the computational availability of collections, we begin from a perspective in which collections are an amalgamation of fragmentary elements -- and their decomposition is neither affordance nor flaw, but instead a natural state of flux that allows them to be contextualized anew through a continual state of reconstitution and derivation. This prevailing logic of decomposition distinguishes collections not as data but instead as pieces and processes, with attendant opportunities and entanglements -- collections and data become inseparable, commingled not in operation but instead via a type of consanguinity. Likewise, our services supporting computational access to data should match this latent consanguinity.

As a large-scale, online digital library that is also a mission-driven, nonprofit technology developer, the Internet Archive has long approached collections as data. Being fully online, with no physical reference collections other than those intended for digitization, collections and data are so intertwined as to be indivisible, either in concept, technology, or use. The Internet Archive's collections include more than 30 petabytes of unique data and has supported computational use of these collections since its beginning, from projects as wide-ranging as semantic analysis of television closed-caption transcripts to network graph study of linking behavior of hundreds of terabytes of web data. In addition, and as a self-sustaining non-profit, the Internet Archive has facilitated this type a research through a service-oriented and sustainable program development approach. Developing data-driven approaches to access and binding them to scalable, sustainable programs has elucidated many of the obstacles and potential solutions that emerge from this work. Questions that have emerged:

- How can computational research services create better pathways to interpretation through tools and methods for the smooth traversal between "reduction and abstraction" inherent in derivation and aggregation?
- How can new access models help researchers have greater comfort with technical mediation at multiple levels and with an increasing distance between the granularity and totality of the object(s) of study?
- How can programs address the challenges still inherent, even with derived datasets, of limited technical proficiency and local infrastructure?

In testing multiple models internally, and surveying and collaborating with similar efforts in the community, we developed a loose typology of program models for research services, oriented towards, but not exclusive to, very large born-digital collection such as web archives.

- **Bulk Data Model**: The totality of domain, global-scale crawl, or large born-digital collection is transferred to researchers via data shipped on drives. Analysis takes place locally, usually in a researcher's own high-performance computing environment.
- **Cyberinfrastructure Model**: A custodial/archival institution provides free/subsidized access to its own computing environment that is pre-loaded with data, VMs, and other tooling.

Researchers can do analysis in this remote environment and export results.

- **Roll Your Own Model**: Researchers receive support, generally in the form of funded or sponsored services, to create their own tools and leverage existing data platforms for candidate collection building and analysis.
- **Programming Support Model:** Researchers, generally non-technical, are given time with specialized technical support staff (engineers) to collaboratively build or aggregate datasets and perform analysis.
- **Middleware Model**: The creation of specific tools and platforms that operate between data hosted with a custodian and advanced analytics tools maintained externally.
- **Derivative Model**: Provide pre-defined datasets that contain key extracted, derived, or pre-analyzed data culled from specific resources. The derived datasets support specific research questions, are fungible, and align data and delivery with researcher need.

While the Internet Archive has pursued many of these models, the most flexible and scalable has proven to be the derivative model, in which key elements are extracted from primary resources and packaged in simple but easy-to-use datasets. This preference was the result of many lessons learned in working to support computational use of extremely large digital collections.

- Services for computational access are more successful when built on top of, or expanded from, pre-existing internal systems, processes, and infrastructure. Modular, generalized, and interoperable are preferred and boutique services don't scale.
- Research services should be flexible and, most importantly, content delivered should be disposable to the providing institution and be able to be recreated by existing, ongoing pipelines or frameworks.
- Focus on derivation (extract desired data from origin), portability (processes should work on multiple content types or in many areas of the workflow) , and access (ease of transfer of data to recipient and ease of use by the recipient).
- Focus on scalable partnerships & decentralization in research service support.
- Researcher expectations often are not aligned with available custodial resources or services and research methodologies (conceptual, practical, technical) often are not aligned with target data characteristics, acquisition methods, or management tools.
- Service models must be self-sustaining and scale. No "grant then gone."
- Continually orient towards mutually reinforcing work, be it with collaborators or researchers, and always allow for generality, in partners, technologies, and models.

Discovering how these lessons and approaches match, contest, or augment the findings of other efforts will be a particularly informative result of the "Collections as Data" forum.

Experiencing Library Collections as Data

Alexandra Chassanoff, Massachusetts Institute of Technology

Recent empirical research has confirmed that digital tools and technologies are fundamentally changing how scholars work.[1] Yet the inverse of this relationship has received little attention – how is infrastructure changing to support emergent scholarly practice?[2] As you note in your grant narrative, "Predominant digital collection development focuses on replicating traditional ways of interacting with objects in a digital space." Indeed, much of the research examining how scholars find, access, and use materials in digital collections has paid little attention to qualitative factors about the interaction between collection users and environmental aspects.[3]

My doctoral research focused on this problem – exploring how scholars were searching for, accessing, and using digitized archival photographs as forms of historical evidence. An underlying objective of my research was to explore the interpretive and evaluative practices that scholars bring to bear on non-textual objects of humanistic inquiry. The intent was to think about how digitized photographs can function as data, and to provide a perspective on what makes interactions meaningful for scholars working with digital materials.

In my role as the project manager on the BitCurator and BitCurator Access projects, I worked with scholars and archivists to develop approaches and methodologies for accessing and using born-digital materials. At the close of each project, I recall thinking that technology was hardly the difficult part of our work. Rather, the challenges we faced seemed to be conceptual in nature. How might we envision ways to access born-digital materials? Relatedly, how might we use born-digital materials in our research? What kinds of questions could be asked and answered from examination of contents of the so-called black box?

It seems that we face a similar challenge in considering library collections as data. I am grateful that this forum is explicitly seeking to address this gap, particularly through the enlistment of a diversity of players in the cultural heritage community. Technologists, librarians, museum professionals, archivists, and scholars will contribute important and unique perspectives to this conversation. Strategic approaches that facilitate access to, and preservation of, library collections as data will need to consider the constant and shifting interplay between infrastructure and emergent scholarly practices. For example, recent research has shown that scholars are using Google Image Search to locate archival photographs. Traditional archival design approaches may not accommodate the serendipitous possibilities of digital space.

In thinking about ways to facilitate use and reuse, I hope to draw on my current research as a CLIR/DLF Software Curation Postdoctoral Fellow. Since October, I have been working at the MIT Libraries to investigate and make recommendations for how institutions can manage software as complex digital objects across generations of technology. Software is another type of "data", albeit one with implicit constraints for access, use and reuse. Researchers rely on software for a variety of research activities – as a subject of research itself, a way to operationalize methods, or to reproduce and validate previous results. Institutions are increasingly tasked with activities related to the active management of software: from creation through use, dissemination, preservation and reuse. Institutional approaches to software collection development must consider software in a variety of contexts: at an intellectual level (e.g. selection and appraisal); in planning for and designing repositories, platforms, services; and in developing staff competencies.

How can we accommodate the fluid and rapidly changing practices which characterize the current scholarly landscape? The results of my dissertation research suggest that one part of the puzzle might be to develop an understanding of the factors and qualities that make experiences meaningful in different kinds of interactions. For example, what is it about the experience of (digitized) oral histories that make them accessible and usable? Rather than focusing on delivery mechanisms or crafting explicit methodological approaches, we might do well to consider the myriad ways in which specific types of materials in digital library collections can be experienced.

Works Cited

[1] Alexandra Chassanoff, "Historians and the Use of Primary Source Materials in the Digital Age," *The American Archivist* 76, no.2 (2013):458-480; Jennifer Rumer and Roger C. Schonfeld, *Supporting the Changing Research Practices of Historians, Final Report from ITHAKA S+R* (2012), 11

[2] The important relationship between infrastructure, technology, and scholarship is explored in Christine Borgman's *Scholarship in the Digital Age: Information, Infrastructure and the Internet* (Cambridge: MIT Press, 2007).

[3] Two notable exceptions in the field of Library and Information Science (LIS) are: Marcia Bates, "**T**he Cascade of Interactions in the Digital Library Interface," *Information Processing and Management* 38, no. 3, 2003; Christopher A. Lee, "Digital Curation as Communication Mediation," in *Handbook of Technical Communication*, ed. Alexander Mehler, Laurent Romary, and Dafydd Gibbon (Berlin: Mouton De Gruyter, 2012), 507-530.

Unsolved Problems in the Humanities Data Generation Workflow: Digitization Complexities, Undiscoverable Audiovisual Materials, and Limited Training for Information Professionals

Tanya Clement, University of Texas Austin

Digital Humanities has changed rapidly from a field that in which we primarily build and create access to resources in the humanities to a field in which we deploy analytics on those resources in accordance with a general move to data analytics. The Always Already Computational initiative is taking an essential step towards bridging the first activity (digitization) to the second (analytics) by focusing on how we structure, bundle, and disseminate digitized or born digital collections and metadata on such collections. This is important and much needed work, but there are three main areas of concern or "unsolved problems" that I would like to introduce into the conversation for the consideration of the group: (1) digitization workflows; (2) AV metadata; (3) and pedagogy in terms of training information professionals about data science, data analytics, and data visualization.

Digitization workflows are where much library collections "data" such as descriptive or technical metadata are born, but these workflows are complicated processes that include selecting collections; establishing performance goals based on standardized measurement protocols; developing efficient test plans; and taking corrective action to maintain quality. Even as cultural heritage institutions continue to rapidly digitize and refine these workflows, our knowledge about new approaches to digitization standards, to schemas for the semantic web, and to increasing our regard for issues of diversity and inclusivity in the digitization of cultural heritage artifacts continues to evolve. Newly issued guidelines from FADGI[1] – an initiative incorporating many entities at the Library of Congress – challenge librarians and archivists to improve image quality precisely when pressures to digitize everything including collections that embody inclusivity are building. Consequently, much of the metadata that we may use in a data framework has been generated during an evolving and complex digitization process, which is often a time of increased one-time funding for the specific digitization job. To what extent will the guidelines that we generate during Always Already Computational take digitization workflows into account? Can we advise libraries and archives on how an understanding of an eventual data framework can be integrated into these workflows such that when requests for funding are made our colleagues can anticipate generating the kinds of data that we will need for a data access environment?

Second, and a case in point for the first "unsolved" problem, Audiovisual materials are notoriously under represented in digital humanities precisely because they often lack the detailed data (or metadata) that supports their effective discovery, identification, and use by researchers, students, instructors, or collections staff. In recent years, increased concern over the longevity of physical AV formats due to issues of media degradation and obsolescence, combined with the decreasing cost of digital storage, have led libraries and archives to digitize recordings for purposes of long-term preservation and improved access. However, unlike textual materials, for which some degree of discovery may be provided through full-text indexing, AV materials that lack detailed metadata cannot be found, understood, or consumed. Most open source and commercial efforts that attempt to generate computationally-assisted metadata and to facilitate improved discovery are narrow in focus, non-scalable, developed as standalone tools, and do not address the rights and permissions that collections staff must consider for creating access. Because of the complicated morass of technical and social issues that limit AV discovery, and descriptive access to audiovisual objects at scale would require a variety of mechanisms for analysis that would need to be linked together with tasks involving human labor in a recursive and reflexive workflow platform that could eventually facilitate compiling, refining, synthesizing, and delivering metadata. Colleagues from Indiana University and AVPreserve and a team of researchers at UT including myself are in the process of developing such a workflow platform, which would allow libraries and archives to bring together and use task-appropriate tools in a production setting. This work is in direct conversation with the kind of framework that Always Already Computational is proposing, but we believe that AV needs, which include generating data about AV materials as a solitary means of providing access to materials that may never (because of privacy and copyright concerns) be publically accessible, are distinct from, though complementary with, those needs that correspond to generating data for text collections.

Third, while information literacy is today a routine goal of library instruction, data work that includes enabling data discovery and retrieval, maintaining data quality, adding value, and providing for re-use lags as a topic.[2] If the library is the laboratory of the humanities, this lag impacts how the digital collections that librarians curate are used in the humanities. Rigorous data work requires data "carpentry" knowledge that considers validity, reliability, and usability as well as critical literacies more generally such as data quality, authenticity, and lineage, but humanists and librarians are not traditionally trained on evaluating these aspects of data. The corresponding difficulty of training students and professional academic librarians lies in the ever-evolving nature of data work, which must respond to changing standards and needs in the context of increasing data in the humanities and of changing infrastructures in libraries. There is work being done in this space including the Data Science Curriculum Project, which is meeting just after the Always Already Computational meeting in Washington DC with representatives from the American Statistical Association (ASA), the ASA Business-Higher Education Forum (BHEF), the Association for Computers and the Humanities (ACH), the Association for Computing Machinery (ACM), the Association for Information Systems (AIS), the IEEE Computer Society (IEEE-CS), INFORMS, the iCaucus, EDISON, and the American Association for the Advancement of Science (AAAS). As well, many programs in Data Science have emerged in recent years at many universities and in many iSchools, but there are few programs of study that focus specifically on teaching students with concerns shaped by the humanities in the context of humanities collections. Conversations on data science pedagogy are needed to ensure the integration of up-to-date resources, theories, and practices in data work in a curriculum that will be geared towards inclusivity and teaching the next generation of our digital workforce about data preparation and analysis in the humanities. Again, this work is directly relevant to the Always Already Computational conversation since the data framework proposed requires practitioners who also have some training in data work.

Works Cited

[1] Federal Agencies Digitization Guidelines Initiative. Technical Guidelines for the Still Image Digitization of Cultural Heritage Materials. September 2016. http://www.digitizationguidelines.gov/.

[2] Association of College and Research Libraries. Working Group on Intersections of Scholarly Communication and Information Literacy. Intersections of Scholarly Communication and Information Literacy: Creating Strategic Collaborations for a Changing Academic Environment. Chicago, IL: Association of College and Research Libraries, 2013.

Computing in the Dark:
Spreadsheets, Data Collection and DH's Racist Inheritance

P. Gabrielle Foreman and Labanya Mookerjee, University of Delaware

Living in a nation of people who decided that their world view would combine agendas for individual freedom and mechanisms for devastating racial oppression presents a singular landscape.

-Toni Morrison*, Playing in the Dark*

Early on in the "Always Already Computational" abstract this assertion appears, underscoring a central assumption of the project: "predominant digital collection development focuses on replicating traditional ways of interacting with objects in a digital space. This approach does not meet the needs of the researcher, the student, the journalist, and others who would like to leverage computational methods and tools to treat digital library collections as data." Not only do the protocols and development of digital collections, of interacting with objects, not meet the needs of various users—let's call them people or communities—who interact with "objects in digital spaces," the lexicon itself reproduces particularly freighted ideas for Black communities of researchers and students, many of whose ancestors entered the West as chattel property, as people who were both called objects and "leveraged," that is bartered, mortgaged, sold and *listed* as such. In the US, this is true for the almost 250 years of municipal, census, and other records which make up collections and archives during slavery, for records that document the debt peonage that characterizes Jim Crow, and, one might argue, for ways in which Black people are accounted for in a prison industrial complex that again treats members of communities as things to be categorized, as surveilled and recorded objects.

The lexicon of digital collections extends the freighted, fretted, relation of categorization and data collection, to Black subjects and Black subjectivity. The term "item," like "object," again recalls the ways in which Black people appear/ed in public records—as items on manifests, as "losses" on insurance claims, and again as items for sale in newspapers or to be distributed in probate. "Fortune" was an 18th-century Connecticut enslaved man whose very name announces his relation to the capital production, the wealth and fortune, he was meant to produce for his enslaver, Dr. Preserved Porter (this is not a typo). When the doctor died not long after he did, Fortune appears in probate records as a skeleton the doctor made from his body, claiming him in death as in life, and literally transforming him into both material object and intellectual prop and property. Fortune's own wife, Dinah, still enslaved by the family, was worth *less* as a living, sentient, being in those records than her husband's skeleton, a skeleton she may have had to dust or clean, the bones of a husband she could not bury.

Likewise, the spreadsheet opens up complex analogies to the ledger, as Labanya Mookerjee, a former exhibits committee co-chair for the Colored Conventions Project, writes in her "Disrupting Data Viz. & the Colored Conventions Project: Interrogating Data Management Methods through Disability Studies," a piece she wrote and published on tumblr for a graduate seminar led by P. Gabrielle Foreman. Storing data in spreadsheets powered by programs such as Microsoft Excel introduces an additional layer of complications; spreadsheets, as bookkeepers of capitalism, can be traced directly to the history of slave trader ledgers. The violence of this history runs the risk of being replicated if we continue to use conventional methods of storing data. As many DH critics have now pointed out, the institutional power invested in the process of data collection—the prelude to data visualization—can be discussed alongside conversations on the power in the production of the archive. Computational activity “is contingent on the availability of collections that are tuned for computational work (Hughes 2014),” as the Always Already Computational abstract asserts. “Suitability is predicated on form, integrity, and method of access (Padilla 2016). This points us to the hegemonic logic guiding the selective operations in knowledge production that has been interrogated through studies on the archives (Trouillot) and in data visualization (Drucker). Both Trouillot and Drucker make a DH community (attuned to archive production as well as archive availability) aware of the need to name the difference between “capta” and “data” and to challenge and counter the institutional powers that authorize “credibility” or “suitability” (Padilla).

Datasets, when constructed using conventional methods of data collection and organization, run a similar risk of activating institutional power and defining “credibility,” especially when the data is procured from traditional archival sources that too often excise, anonymize and erase certain subjects, transmogrifying them in turn into (almost invisible, ghosting) “objects” and “items.” Two examples from the Colored Conventions movement obtain. First is the challenge of including Black women whose names and participation are excised when we use traditional methods of collecting and naming data (from the lists of thousands of delegates over seven decades). Curating a dataset that is reflective of the actual history of women’s involvement has prompted CCP to revisit the logic used to develop the parameters of what qualifies as “participations,” extending the definition of participation from appearing in the minutes, to attendance at the gatherings, and to hosting and curating conversations (following Psyche Williams-Forson) at boarding houses, eateries etc. where women’s presences or imprints appear. A second example is the work that Jim Casey, co-founder of CCP, has done on social network analyses and data visualization between Colored Conventions and The Underground Railroad showing a surprising lack of overlap and co-attendance. “All of this data is vexed,” asserts Casey, “shaped by centuries of decisions based on racial hierarchies about what to record, store, and reproduce.” Casey uses Siebert’s “Directory of the [3000] Names of Underground Railroad Operators” included in his Underground Railroad (1898), and Boston Public Library’s Anti-Slavery Collection Data. These sources hew to a historical imaginary that places whites at the center of the UGR and that excises Black leadership and involvement, a corrective that has just begun to appear in recent scholarship and has not produced a directory as of yet. Based on racially hegemonic raw data, the co-attendance visualizations don’t capture Black UGR involvement by default.

This leads us to this set of questions. How do we account for (new, collective) data collection that accounts for haunting imprints and outright absences in the archives upon which we depend? What are the implications of a lexicon and set of practices/tools that rely upon and reproduce a colonial language of power and entitlement in the digital humanities as we think collectively about best practices to “leverage computational methods and tools to treat digital library collections as data”.

Frictionless Collections Data

Dan Fowler, Open Knowledge Foundation

Data Package is a containerization format for all kinds of data. It provides a framework for "frictionless" data transport by specifying useful metadata that allows for greater automation in data processing workflows. The aim is to provide the minimum amount of information necessary to transfer data from one researcher to another, and, likewise, one data analysis platform to another. After several years developing these specs for general use, it is worth directly examining the extent to which library and museum collections data are amenable to this approach.

New approaches to publishing library and museum collections data are necessary. Such data, released on the Internet under open licenses, can provide an opportunity for researchers to create a new lens onto our cultural and artistic history by sparking imaginative re-use and analysis. For organizations like museums and libraries that serve the public interest, it is important that data are provided in ways that enable the maximum number of users to easily process it. Unfortunately, there are not always clear standards for publishing such data, and the diversity of publishing options can cause unnecessary overhead when researchers are not trained in data access/cleaning techniques.

One approach for publishing collections data is via an API (Application Programming Interface) on a record-by-record basis. This approach has its advantages: the data is likely structured and well described. However, these services may not map directly to the types of queries or analyses researchers need to run. Further, for both the researcher and publisher, it can be tedious and costly to provide large amounts of collections data delivered record-by-record. For certain use cases, it is preferable to publish data in bulk format in open standards like CSV or JSON. The Metropolitan Museum of Art and Tate Gallery, for instance, have released their collections data as sets of text-based files on GitHub. In this approach, associated documentation is provided via files named by convention, for example, "README" or "LICENSE". This method of publishing allows users to load data into their own tools without the overhead of programming against an API.

Documentation for data published in bulk is often ad hoc. There is often no clear or rigorous documentation of the fields (what types of data are in each column). Reading such data into data analysis programs using the built-in CSV ingest mechanisms yields data divorced from context: common date and boolean ("TRUE/FALSE") columns must be explicitly assigned as such, numeric identifiers may be incorrectly loaded as integers, etc. These datasets are often exported from in-house collections database software, and small errors in the translation of these often large datasets may go unnoticed.

Data Packages for Collections

Frictionless Data, developed in the open by Open Knowledge International and members of the open data community, is an ideal framework for publishing this type of bulk data. The Data Package format, requiring only the addition of a descriptor file called datapackage.json, provides a minimally invasive, but standardized way to provide clear and machine-readable metadata. Datasets created as Data Packages can later be easily exposed as APIs given the wealth of metadata provided.

As an example, the Carnegie Museum of Art in Pittsburgh, Pennsylvania has provided its collections data as a downloadable Data Package. Providing the data in this format yields several benefits:

1. Users are provided with useful metadata to allow for easy import into their preferred analysis tool. These explicitly defined column types and metadata can eliminate some of the tedious work involved in “wrangling” a dataset.
2. Publishers can use tooling like Good Tables to automatically validate data.
3. Basic documentation for how to use the dataset (e.g. what columns mean) can be automatically created from structured metadata.
4. Collections data can be licensed in a machine-readable manner.
5. In the absence of Data-Package-aware tooling, the original data can be read/written as usual.

Over the course of this year, with the continued support of a grant from the Sloan Foundation, we are looking to work with researchers and institutions across a variety of fields to pilot the use of the specifications. This may involve building tools and writing guides to analyse, validate, and/or visualize collections data. Through this process we hope to improve the specifications more generally while also providing useful tooling for researchers in digital humanities.

Book carts of Data:
Usability and Access of Digital Content from Library Collections

Harriett Green, University of Illinois at Urbana-Champaign

Not all of the data we create or purchase for Library collections comes in neat multi-gigabyte packages of ordered files: We recently discovered that datasets we had purchased as part of a database licensing negotiation were more shelf ready than machine ready: They currently exist as stacks of hard drives, discs, and other bewildering formats sitting on a book cart. How do we provide access to these data collections?

In my extensive work with research teams, graduate students, and faculty members to obtain, generate, and transform data derived from collections in the University of Illinois Library and far beyond, the question of access and usability consistently rises to the fore. Thus, I would ask, how can we conceptualize the full spectrum of data usability? It is not enough for us to digitize the collection materials and for the data to exist on someone's server: Usability encompasses data formats, tool interoperability to the negotiated permissions and rights for researchers to share and manipulate data as they engage in analytic workflows.

Data usability means developing data models that take into account the actions that will be performed on our data. In determining the different types of data models that we can build and implement into our collections, we must consider how humanists and social scientists effectively work with data in their research and teaching.

My work with the HathiTrust Digital Library and HathiTrust Research Center has seen this practice: The HTRC has attempted to meet various expertise levels and needs of users in enabling access to the data: On the newcomer end of the spectrum, we provide fully guided access to gathering and using data through our Workset Builder and the Portal with its pre-set algorithms. But researchers frequently express the need for larger-scale data that is more pliable and manipulatable, so the HTRC developed the Extracted Features datasets that allow researchers to generate highly customized and curated datasets. But the barriers to accessing this data can be high in terms of skillsets needed to both access and use the data.

My research explorations on scholarly research practices also have shown me that data usability is critical:

> Our research for the HTRC's Workset Creation for Scholarly Analysis project examined researcher requirements for textual corpora to be useable for research (Fenlon et al. 2015, Green et al. 2014). Our interviews with scholars revealed that the core areas of concern for researchers included the conceptualization of collections as reusable datasets and resources for scholarly communications; the ability to break apart collections into various levels of granularity to generate diverse objects of analysis; and the need for enriched metadata. We proposed building out the data model of the "workset," the HTRC-specific term for textual corpora that researchers build.
>
> Our subsequent user study for HTRC User Requirements (Green and Dickson, 2016) gave further insights on how researchers used textual corpora and their scholarly practices that shape their

needs for being able to work effectively with text collections in the HathiTrust Digital Library, as well as overall. We learned that scholarly practices and notable challenges when working with our textual collections included the ability to acquire and structure the data; the need for a space to work with various tools and generate results; the ability to share data for research collaborations; and the role of data in teaching and training.

And my recently concluded research study for Emblematica Online explored how scholars engaged with the digitized emblem books drawn from leading rare book collections at Illinois, HAB Wolfenbuettel, University of Glasgow, Duke, and the Getty Institute. In my examination of how scholars engaged with these multi-institutional collections, their metadata, and the interlinked digital content through interviews and usability testing sessions, we found that the expectations of users when exploring digital collections is complex: They range from the basic need for high-quality reproductions, which *Emblematica* was praised for by all participants; to advanced scholarly concerns such as the ability to distinguish between the types of archival content they are perusing—emblem books versus emblems themselves—and the historical particularities of this specialized genre of emblem studies. Respondents frequently expressed the need for context, annotated content, and other functionalities that would allow them to fully engage with the emblem books as an archival source and scholarly area. We considered that this may reveal the needs of interdisciplinary scholarship as researcher take advantage of easy access to vast digital collections of content: The scholarly knowledge base that users approach with digital collections varies widely, and an effective digital collection must welcome all levels and inculcate them into the scholarly domain of the collection.

These are some of the findings I have learned in my work to examine what researchers needs are as they engage with our Library collections in digital formats and make use of these materials as data. This Forum's discussion can provide critical new avenues for exploring how collections can be accessible, browseable, and extensible for addressing a diversity of emergent uses in research and teaching.

Works Cited

Fenlon K., Senseney M., Green H., Bhattacharyya S., Willis C. and Downie, J. S. (2014). Scholar-built collections: A study of user requirements for research in large-scale digital libraries. *Proceedings of the American Society for Information Science & Technology* 51(1), 1–10. doi: 10.1002/meet.2014.14505101047

Green, H. E., Fenlon, K., Senseney, M., Bhattacharyya, S., Willis, C., Organisciak, P., Downie, J.S., Cole, T., and Plale, B. (2014). Using Collections and Worksets in Large-Scale Corpora: Preliminary Findings from the Workset Creation for Scholarly Analysis Prototyping Project. Poster presented at iConference 2014, Berlin, Germany.

Green, Harriett, Eleanor Dickson, and Sayan Bhattacharyya. "Scholarly Requirements for Large Scale Text Analysis: A User Needs Assessment for the HathiTrust Research Center." Digital Humanities 2016 Proceedings, Krakow, Poland, July 11 – 15, 2016.

Green, Harriett, Mara Wade, Timothy Cole, and Myung-Ja Han. 2015. "User Engagement with Digital Archives: A Case Study of Emblematica Online." In *Creating Sustainable Community: The Proceedings of the ACRL 2015 Conference*, edited by Dawn Mueller, 177–187. Chicago, IL: Association for College and Research Libraries.

Historical Complications of/for Open Access Computational Data

Jennifer Guiliano, Indiana University–Purdue University Indianapolis

Always Already Computational seeks to support the "development of a strategic approach to developing, describing, providing access to, and encouraging reuse of library collections that support computationally--driven research and teaching." Historically, data in the digital collections sphere has most often been expressed as homogenous datasets falling into one of three primary types: textual, visual, or audio. "Scholars" or "researchers" use large scale textual information derived from digitized volumes or the extraction of text only from hypertextual and multimedia environments or they mine hundred or even thousands of hours of video or audio materials to extract and analyze subsets. Due to the dominance of datasets like those derived from the Google Books corpus or through webscraping tools that cull text,image, or audio, large or dense cultural datasets are the norm in digital humanities, and are not only homogenous in type but rarely imagine interactions as led by or with intervention from individuals not holding the role of scholar or researcher.

More simply, I am suggesting that the question of creating computationally-accessible datasets is not just the deployment of an ecosystem for development, description, access, and reuse but a recognition that there are potentially multiple ecosystems of research and teaching that *must exist simultaneously* and be treated as relational computational data. To illustrate this principle, I'll provide a brief synopsis of the work of Edward Curtis and how the open access images that are currently available as computationally-accessible data through the Library of Congress present a complicated consideration of computational data. Beginning in 1868, Edward S. Curtis embarked on a thirty-year career documenting over eighty native communities. Participating as part of scientific expeditions and anthropological excursions, he produced roughly 20 volumes of information on Native and Indigenous life that were accompanied by photographic images as part of his *The North American Indian* series. Created primarily as silver-gelatin photographic prints, this series has long held a place of prominence in historical analysis as the images are not only noted for their rarity but for the limited dissemination and reuse throughout the twentieth century as full sets of materials. Only 300 sets of the 20 volume series were sold; however, these images as individual objects have seen significant dissemination and reuse since their acquisition by the Library of Congress. More than 2,400 silver-gelatin photographic prints (of a projected total of 40,000) were acquired by the Library of Congress through copyright deposit from about 1900 through 1930. About two-thirds (1,608) of these images were not published in Curtis's multi-volume work, *The North American Indian*. The collection includes individual and group portraits, as well as photographs of indigenous housing, occupations, arts and crafts, religious and ceremonial rites, and social rituals (meals, dancing, games, etc). More than 1,000 of the photographs have been digitized and individually described and are available through the Library of Congress API as well as via manual download of both jpeg and tiff file formats.

Using strategies common to anthropologists working in indigenous communities at the turn of the 20th century, Curtis modified the images he produced to remove signs of modernity and contemporary life. This included providing specific forms of dress that were perceived as being "more traditional" as well as stronger interventionist strategies like removing objects that would signal integration with 20th century Euro-American society. When viewing an image of a Piegan lodge on the LOC website, the unretouched negative is provided to the API of an image of two Piegan men situated in their lodge with a clock centered between them. A computational dataset would expose the existence of this image, which could allow scholars to run object based visual analysis algorithms to identify the clock in the image and potentially find other images of modernity using shape-segmentation leading to some conclusions about the interventionism of technology in indigenous life---how widespread has technology embedded itself into indigenous life? But in current thinking about computationally-accessible data, what would not be revealed is that this original negative shows an alarm clock between two seated men in a Piegan lodge, not the published, retouched image that American audiences would have viewed in *The North American Indian*. Curtis physically cut the clock out of the negative. He then the retouched the image for publication in *The North American Indian*. It is important for accuracy purposes for the dataset to reflect not just the original photographic negatives but also relational data derived from what was actually published by Curtis. Otherwise, researchers might conclude that Americans were familiar with signs of modernity in indigenous life when, in fact, that conclusion is relatively recent historiographically. Other examples of this type of relational computational-data are available with Curtis: he depicted a Crow war party on horses, even though there had been no Crow war parties for years, and he used techniques of focus and duration to induce hue saturation that romanticized images.

More problematically, for our computational dataset, Curtis was also known to photograph religious rituals as part of his excursions. The [Oraibi snake dance] image depicts Hopi natives that were part of the Snake and Antelope societies participating in a communal ceremony. Performed in August to ensure abundant rainfall to help corn growth, the ritual was the most widely photographed ceremony in the Southwest Pueblos by non-native observers. In current computationally-accessible form, there are a number of issues to confront: 1) there is no notation that this image is of a religious ritual that is now prohibited from viewing by the non-Hopi public (and thus should be pulled from view for reasons of cultural sensitivity); 2) when subjected to computer vision techniques, the derivative images rely on segmentation of physical bodies---a form of disembodied violence that reflects colonial practices where Natives are treated as less than human through segmented image representation (e.g. scalps, severed limbs, etc). More holistically, this case illustrates one of the long-term challenges of computationally-enabled access: computers cannot identify culturally-sensitive data nor is there an efficient means to retrieve culturally-sensitive data once it has been distributed in computational form. While data might be displayed in an integrated manner, when it comes to the processing or analysis of our data, computational analysis has largely existed at a segmented level rather than as an integrated structural process for research and teaching purposes. A complex humanities system for data are often artificially layered representations that rely on augmentation of 'found' datasets such as traditional and web archives.

Often, human intervention is needed to verify the results of these computational processes, which have a habit of very quickly highlighting contradictions at the level of both object and corpora. An integrated data ecosystem posits that through computational analysis it is important not only for core activities of development, description, access, and reuse, but also the return of data to its originating collection through data correction and relational derivatives. More simply, what is needed is an integrated *humanities* data ecosystem that recognizes approaches to computationally-accessible data and relies on important characteristics of humanities research data and humanities research practices: 1) humanists tend to create data, not just gather data; 2) some of this data is inherently structured, but most is not; 3) the resulting data is often highly interpretative, which has implications for sharing and re-use; 4) data creation is often iterative and layered with implications for copyright, versioning and active working spaces; and 5) the process is as important as the product. And, significantly, to envision the broadest potential intervention of computationally-accessible datasets, we cannot envision that the terms “scholar” and “researcher” belong to the academic or archival communities. We must understand that the communities of origin should be the initiating point for considering development, deployment, access, etc.

Works Cited

[1] Portions of this response appeared in an earlier form in the Introduction to "The Future of Digital Methods for Complex Datasets", an *International Journal of Arts and Humanities Computing (IJHAC)* special edition and as a contribution to a Digital Library Federation panel on Humanities Data issues. Jennifer Guiliano and Mia Ridge, International Journal of Humanities and Arts Computing, Volume 10 Issue 1, Page 1-7. DOI: http://dx.doi.org/10.3366/ijhac.2016.0155.

Identifying Use Cases for Usable and Inclusive Library Collections as Data

Juliet L. Hardesty, Indiana University

A grounded, practical approach to digital projects often centers around concerns of how will the project be useful, how can the project realistically be completed, and what information is necessary to make this project (or the items in a digital project) discoverable and accessible? Based on this approach, there are two sides to making library collections useful as computational data – the collection-holding library has to be able to release the data in a way that allows for computation and researchers have to be able to find out about this data and do something with it. Putting data out there does not mean it will be used and offering a computational interface does not mean it will fit all research needs.

The grant references the HathiTrust Research Center (HTRC) as an example of a computational interface for researchers. It also references Hydra-in-a-Box as an example of an application that could benefit from computational functionality. This generated the thought of an HTRC-in-a-Box that could work for libraries to set up their own computational interface for their collections. Open government data efforts like Code for America or data.gov and ckan.org show how various groups and individuals can come together around a common goal of providing access to computational data and provide ways to access, analyze, and offer data. It would be useful to examine those models when discussing approaches to treating library collections as data.

This project is concerned with all types of digital objects. Text, images, audio, video, born-digital, 3-dimensional, all have unique aspects to them that are sometimes computationally available but often are not. Sometimes the only way to know about segments on a video or the contents of an image is to have textual description available. That requires metadata generation or metadata enhancement. This work can be manually intensive but can also be aided by software. Efforts such as AVPreserve's plan to enhance metadata in stages for Indiana University's Media Digitization and Preservation Initiative move gradually toward more advanced technologies to identify aspects such as people's faces, beats per minute, and speaker identification in video and audio for the purpose of producing metadata than can then be discovered by researchers.[1] Another project to watch will be Wikimedia Commons' Structured Data project to "develop storage information for media files in a structured way on Wikimedia Commons, so they are easier to view, translate, search, edit, curate and use."[2] This process will not always be just about putting the data out there or making it possible for researchers to access the data, it will also involve producing data about different types of objects than has traditionally been the case in digital libraries. Recommendations, tools, and workflows for metadata enhancement will be necessary to create usable computational data.

Michelle Dalmau, Head of Digital Collections Services at Indiana University, correctly points out that different use cases are needed for library collections as data.[2] At Indiana University, several digital collections are available as datasets,[3] largely based on researcher requests. Tracking use in the wild is challenging, but datasets are used in the classroom (Charles W. Cushman Photograph Collection) and for research (Wright American Fiction). Looking at how data is used for research compared to how it is used pedagogically for instruction might lead to insights on qualities of data that make collections better suited for teaching versus research. Being able to reliably trace the ways in which these data sets are used will demonstrate impact to stakeholders. Using metadata about digital collections versus using the collection items themselves for content analysis is something else to consider. The British Library offers image collections for analysis separate from bibliographic datasets about their archival holdings. Indiana University's Cushman dataset offers only the metadata about the images, not the images themselves.

A final point to bring up concerns diversity and inclusion. Not only should this project make sure the collections considered for use cases are diverse in format, content, and source, but the project itself needs to have a broad and deep representation of voices and perspectives on computational data. These are not data that are only useful in the academic realm. Access to computational data or workflows and tools to allow others to provide access to computational data will be ever more important in the world, particularly if national governments continue to trend toward populism, nationalism, and privatization.

Works Cited

[1] Rudersdorf, Amy and Juliet L. Hardesty. (2016). "AV Description with AVPreserve and IU: Strategies and tools to describe audiovisual materials at scale for Indiana University's Media Digitization and Preservation Initiative." Digital Library Federation Forum, Milwaukee, Wisconsin. https://osf.io/gfazc/

[2] Juliet L. Hardesty interviewed Michelle Dalmau regarding library collections as data in February 2017.

[3] https://commons.wikimedia.org/wiki/Commons:Structured_data

[4] British Library. Collection guides: Datasets for image analysis. http://www.bl.uk/collection-guides/datasets-for-image-analysis

Emerging Memory Institution Data Infrastructure in the Service of Computational Research

Christina Harlow, Cornell University

In my opinion, the *Always Already Computational* Forum work area rests at the intersection of the understood functionalities of memory institution's collection platforms and the needs of researchers working with large-scale or computational data analysis techniques. In thinking about this Forum's scope and my own work, I am struck by possible collaborations not leveraged or mentioned. I would like to explore if my work approach to a facet of a larger data problem could expand and, in turn, be expanded by the Forum's discussion and deliverables on computational research needs and memory institution data practices.

My position for this upcoming Forum will mostly fall along these points:

- If library collections, including but not limited to that of digital repository platforms, are considered (primarily digital repositories are targeted in the proposal), there is a wealth of data and metadata (*data) that already exists. Better yet, memory institutions already work with this *data at scale using traditional and emerging technologies that underpin and are hidden by delivery and discovery interfaces. How can this underlying ecosystem be better leveraged for computational data analysis by researchers? i.e. do we just need to make access to a Solr index publicly available? Can we plug into our library data ETL systems a public Hadoop integration point? Do we need to better document and expose to new communities our existing data APIs or data exchange protocols?

- I would like to surface the functional needs of the research areas alluded to in the proposal, then see where they overlap with existing *data operations work areas in memory institutions. A strategic partnership here means we can strengthen the cases for, collaboration on, and support of the technological, procedural, and organizational frameworks emerging. These are already being built and used to support efforts of memory institutions and their data partners.

- Computational or large-scale *data work requires transparency and agreement on a number of points to make it statistically relevant and publicly reliable. These agreement points include but are not limited to:

    - Machines should be able to understand the models or entities represented by the data;
    - This requires having shared specifications around *data representation and contextual meaning of models, datum, types, etc.;
    - We need to build and maintain consistent data exposure services, points or methods so that computational work can be reproducible, iterated, or distributed as needed (for scalability);
    - Recognize that technological frameworks for computational analysis (for example,

Hadoop) often require significant hardware, software, and maintenance to support. Stability of how data is exposed and data provenance can mitigate the technological burden by offering consistency on which multiple partners can build and coordinate efforts on the frameworks;

- And what is the responsibility of the originating memory institution to support capture of that computational data output for sake of archiving, reproducibility, discoverability, and expanded *data services?

My positions come from my own work on metadata operations within a large and well-funded academic library system. My work focuses on building an efficient and coordinated *data ecosystem among sources including but not limited to:

- A traditional MARC21 Catalog with about 9 million bibliographic records, managed in an ILS (Integrated Library System), a few Oracle databases, a Perl-based metadata reporting and management interface, and other batch job management and metadata exposure services (APIs and data exchange protocols like Z39.50 or SRU);

- A locally-developed metadata integration layer that takes multiple data representations of authority, bibliographic and other metadata retrieved via APIs, merges them, and indexes into a number of Solr indexes;

- Multiple (~8 depending on the definition) digital repository applications and services for delivery of data and metadata to user interfaces. These repositories span technology and resource types from lone Fedora 4 instances for object persistence of primarily text-focused digital surrogates to more traditional DSpace installations for user-generated scholarly output type resources;

- A locally-managed authorities and entities interface that deals with both local vocabularies and enhanced representations of currently 3 large (>1 million resources) external metadata sets;

- And *data from archives, preservation, digitization, and many other workflows and systems.

In building a coherent ecosystem for this *data, I work with enterprise data tooling and approaches that perhaps also can support the computational data analysis needs to be surfaced in the *Always Already Computational* Forum. In particular, I am leveraging ETL and distributed data management systems that then interact with (and coordinate) existing memory institution *data standards, applications, specifications, and exchange protocols. Due to the computational support of the selected distributed data systems, I run a number of processes that parallel some computational data approaches, but for different ends. I would like to outline how we could reuse or expand these existing approaches and services to support the researchers (and their respective areas) who take part in this Forum.

On the Computational Turn in Archives & Libraries and the Notion of Levels of Computational Services

Greg Jansen and Richard Marciano, University of Maryland

1. The Computational Turn in Archives & Libraries

The University of Maryland iSchool's Digital Curation Innovation Center (DCIC) is pursuing a strategic initiative to understand and contribute to the computational turn in archives and libraries. The foundational paper (with partners from UBC, KCL, TACC, and NARA) calls for re-envisioning training for MLIS students in the "Age of Big Data". See: "Archival Records and Training in the Age of Big Data". We argue for a new Computational Archival Science (CAS) inter-discipline, with motivating case studies on: (1) evolutionary prototyping and computational linguistics, (2) graph analytics, digital humanities and archival representation, (3) computational finding aids, (4) digital curation, (5) public engagement / interaction with archival content, (6) authenticity, and (7) confluences between archival theory and computational practices: cyberinfrastructure and the records continuum.

Deeper experimentation with these new cultural computational approaches is urgently needed and the DCIC is developing a CAS curriculum that brings together faculty from Computer Science, Archival & Library Science, and Data Science. We conduct experiential projects teams of students to help them: gain digital skills, conduct interdisciplinary research, and explore professional development opportunities at the intersection of archives, big data, and analytics. These projects leverage unique types of archival collections: refugee narratives, community displacement, racial zoning, movement of people, citizen internment, and cyberinfrastructure for digital curation. See "Practical Digital Curation Skills for Archivists in the 21st Century" (Lee, Kendig, Marciano, Jansen), MARAC 2016. Two workshops on the interplay of computational and archival thinking were held in April 2016 and December 2016, and a pop-up session at SAA 2016 discussed archival records in the age of big data.

Finally, the DCIC is developing new cyberinfrastructure, called *DRAS-TIC* (see Nov. 2016 CNI talk), that facilitates computational treatment of cultural data. *DRAS-TIC* stands for Digital Repository at Scale that Invites Computation (To Improve Collections), and blends hierarchical archival organization principles with the power and scalability of distributed databases.

Our position statement builds to these CAS investigations by suggesting a framework for "Levels of Computational Service" to better describe the emerging ecosystem and identify gaps and opportunities.

2. Levels of Computational Service

Journalists, researchers, planners, and other user patrons support their investigations with new methods of computational analysis. Libraries, archives, museums, and scientific data repositories hold data that will inform their disciplines. It is far easier today to analyze Twitter behavior than it is to investigate public life using public data from public institutions, such as government records, cultural heritage, and science data. We strive to make our public data and cultural memory as open to research as Twitter.

Computational analysis happens in various technical environments: on a single server; in distributed clusters; on cloud services. The tools we use have unique requirements, configurations, and hardware. It is said that a data stewardship organization cannot anticipate the uses for their data, but it is equally true that they cannot anticipate the tools used for analysis. Organizations need a service strategy that serves a range of users, from the most technically innovative, to the most time and resources constrained. We describe a range of services for collections as data without losing site of core services. This is a "maturity model" for stewardship organizations, with *levels of computational services* that show a clear progression toward full service.

2.1. Core Service Level

Shipping datasets into the researcher compute environment remains the critical use case, maximizing flexibility and allowing researchers to link many datasets into one corpus. Researchers need to *discover, scope, ship and make reference to datasets*. Though we may also move computational work across them, boundaries are an important place to define stable conditions, such as custody, provenance, security, and concise technical contracts. Even the most advanced repository must establish these boundary conditions.

- Define license terms, how can we use the data?
- Define provenance:
  - Who produced the data and why?
  - How did it arrive here?
  - Do versions exist elsewhere?
- Define dataset scope:
  - What makes the corpus complete?
  - Is it complete?
  - Is it growing? What is the update history?
- Transfer methods with integrity verification and resume from failure
- Persistently citable datasets

2.2. Protocols Service Level

- File-by-file transfer through HTTP API (instead of batch downloads, like ZIPs)
- Define citable subsets through custom queries or functions.
- Check for updates to any dataset or subset. (via HTTP API)
- HTTP API for navigation of structured collections:
  - Static site (Apache or Nginx auto-index of files)
  - Cloud Data Management Interface (CDMI)
  - Linked Data Platform (and Fedora API)
- Delivery to cloud and cloud-hosted, public datasets

2.3. Enhanced Service Level

- Derived data available as subsets:
  - plain text for documents and images
  - normalized file formats
  - tabular data for table-like sources
  - linked data for graph-like sources
- Machine-readable provenance records
- Crowd-sourcing of metadata
- Named entity indexing and subsetting (people, places, organizations, dates, events)
- Geospatial indexing and subsetting
- Consistent and citable random sample subsets (add random seeds to each observation)

2.4. Computer Room Service Level

Container technologies, such as Docker, ship a custom compute environment to the dataset location. A hosted database can be opened up for queries or distributed compute jobs. While not as flexible as the researcher environment, computer room services provide rapid and cost-effective analysis. Journalists on deadline benefit most from computer room services.

There are also growing calls, beyond the physical sciences, for analysis of big collections data in journalism and humanities scholarship. The sheer scale of big data makes transfer prohibitive, as is provisioning enough storage to host an entire corpus. At the Digital Curation Innovation Center at the University of Maryland's iSchool, we are actively developing the *DRAS-TIC* repository (Digital Repository at Scale that Invites Computation). Through *DRAS-TIC* we aim to deliver computer room-style services over heterogeneous digital collections and remove the limits of scale.

- Run an Apache Spark job on a defined dataset
- Host a compute container with a dataset mounted locally
- SPARQL query service
- Use techniques above to produce a new subset for transfer

3. Provisioning the Researcher Environment

From code notebooks to deployment scripts that provision clusters, it becomes easier to create and share compute environments. Research that aims towards publication will also need to track the research steps workflow. Through machine readable scripts and provenance, we can aim to reproduce an analysis at a different time and place, starting from the cited datasets and well described methods. The curation activities performed by a stewardship organization and the steps taken by the researcher can form an unbroken chain of events leading to a reproducible product.

Summary

For verifiable results in scholarship, or public trust in an independent press, we need to provide relevant datasets and services that make it straightforward to trace findings back to their source in the public record. We must confront a rightly skeptical reader, who faces increasingly high-flying visualizations and claims made from them. They are correct to demand links to the underlying evidence and methods. By providing these we enrich public understanding and trust. At the Digital Curation Innovation Center (DCIC) we have committed to this agenda and pursue it through our research projects, scholarly activities, and the active development of the DRAS-TIC software project, and the building of a computational archival community.

Partnership Recommended – The case of curating research data collections[1]

Lisa Johnston, University of Minnesota Libraries

Digitization alone is not enough to support large-scale computational analysis of library collections. Rather the more difficult steps of digital curation will be necessary to prepare our collections for appropriate reuse. Partnership may be the key.

Take for example the problem of analog data. The extraction of historical climate data from tables and charts and other artifacts (e.g., Zooniverse's Old Weather project) is an ambitious and important undertaking as these data are undeniably valuable and temporally unique. Yet, the digitization of data points from the written page is just the first step toward a greater integration of their meaning in modern and future research. In order for computation of these collections to be successful, the digital surrogate must be curated in a number of ways. The data may be transformed, cleaned, normalized, described, contextualized, and quality assurance measures put in place to ensure trust and track provenance of the work, to name a few. Data curation activities prepare and maintain research data in ways that make it findable, accessible, interoperable and reusable (FAIR).

In our work, the Data Curation Network project has taken steps to better understand the data curation activities mentioned above and identify ways to harness the necessary domain and file format expertise needed to curate research data across a network of partner institutions.[2] We represent academic library data repository programs that are staffed with curation experts for a range of data domains and data file formats. Our goals are to develop practical and transparent workflows and infrastructure for data curation, promote data curation practices across the profession in order to build an innovative community that enriches capacities for data curation writ large, and most importantly, develop a shared staffing model that enables institutions to better support research by collectively curating research data in ways that scale what any single institution might accomplish individually.

We are not alone in this desire to partner on data curation skills, staff, and infrastructure. National examples of data curation such as the Portage Network (https://portagenetwork.ca), developed by the Canadian Association of Research Libraries (CARL), aims to support library-based data management consultation and curation services across a broader network and the JISC-funded Research Data Management Shared Service Project aims to develop a lightweight service framework that can scale to all UK institutions and result in efficiencies by "relieving burden from institutional IT and procurement staff." In the US, partnerships on technological infrastructure are booming. The Project Hydra's Sofia platform (https://projecthydra.org), which builds in the DuraSpace Fedora framework, has been co-developed by numerous institutions that seek to build a better digital repository infrastructure for data. And the Hydra-in-a-Box project (lead in part by another partnership success story for disseminating archival materials, the Digital Public Library of America) aims to provide a networked platform for repository services that will scale for institutions big and small. Another inspiring example is the Research Data Alliance, which provides an incubator for collaboration around a range of data-related topics. RDA projects to track include the Publishing Data Workflows working group and the newly formed Research Data Repository Interoperability working group. And partnerships do not necessarily need to start at the national-level. Several smaller-scale partnerships underway for sharing curation staff expertise across institutions include the Digital Liberal Arts Exchange, which facilitates data-related problem solving and communication amongst peers as well as providing hosting services that allows digital humanities projects to be run on shared infrastructure. And the DataQ Project, which provides a virtual online forum for expert data staff to discuss and provide solutions for data issues in a collaborative way.

By partnering on data curation efforts like these we may move beyond individualized digital curation strategies toward what I hope will become a robust "network" of digital collections that are computational, but also trusted. And as partners in this effort we may continue a shared dialogue and collectively develop new and improved processes for curating research data and other digital objects. Finally, our networked research collections will demonstrate our continuing and important role that libraries and archives have to play in the broader scholarly process.

Works Cited

[1] Portions of this statement were also published in "Concluding Remarks" by Lisa R. Johnston in *Curating Research Data Volume 2: A Handbook of Current Practice* (ACRL, 2017) available as an open access ebook at http://www.ala.org/acrl/publications/booksanddigitalresources/booksmonographs/catalog/publications.

[2] Currently in our planning phase, the Data Curation Network aims expand into a sustainable entity that grows beyond our initial six partner institutions, lead by the University of Minnesota, and are the University of Illinois, Cornell University, the University of Michigan, Penn State University, and Washington University in St. Louis.

Ways of Forgetting: The Librarian, The Historian, and the Machine

Matthew Lincoln, Getty Research Institute

Jorge Luis Borges tells us of Funes, the Memorious: a man distinguished by his extraordinary recall. So precise and complete were Funes' memories, though, that it was impossible for him to abstract from the near-infinity of recalled specifics he possessed, to general principles for understanding the world:

> *Locke, in the seventeenth century, postulated (and rejected) an impossible idiom in which each individual object, each stone, each bird and branch had an individual name. Funes had once projected an analogous idiom, but he had renounced it as being too general, too ambiguous. In effect, Funes not only remembered every leaf on every tree of every wood, but even every one of the times he had perceived or imagined it... He was, let us not forget, almost incapable of general, platonic ideas... he was not very capable of thought. To think is to forget a difference, to generalize, to abstract. In the overly replete world of Funes there were nothing but details, almost contiguous details. (Borges 1962, 27)*

Attending to Drucker's admonition that all "data" are properly understood as "capata", the story of Funes is a potent reminder that it is not only inevitable that we will be selective when capturing datasets from our collections, but that it is actually *necessary* to be selective.(Drucker 2014) A data set that aims for perfect specificity does so at the expense of allowing any generalizations to be made though grouping, aggregating, or linking to other datasets. For our data to be useful in drawing broad conclusions, it is an *imperative* to forget.

However, in considering library and museum collections as data, we must grapple with several different frameworks of remembering, forgetting, and abstracting: that of the librarian, the historian, and the machine. These frameworks will often be at cross-purposes:

- The librarian favors data that is **standard**: forgetting enough specifics about the collection in order to produce data that references the same vocabularies and thesauri as other collection datasets. The librarian's generalization aims to support access by many different communities of practice.

- The historian favors data that is **rich**: replete with enough specifics that they may operationalize that data in pursuit of their research goals, while forgetting anything irrelevant to those goals. The historian's generalization aims to identify guiding principles or exceptional cases within a historical context. (No two historians, of course, will agree on what that context should be.)

- The machine favors data that is **structured**: amenable to computation because it is produced in a regularized format (whether as a documented corpus of text, a series of relational tables, a semantic graph, or a store of image files with metadata.) In a statistical learning context, the machine seeks generalizations that reduce error in a given classification task, forgetting enough to be able to perform well on new data without over-fitting to the training set.

At the Getty Research Institute, our project to remodel the Getty Provenance Index® as Linked Open

Data is compelling us to balance each of these perspectives against the labor required to support them. Our legacy data is filled with a mix of transcriptions of sales catalogs, archival inventories, and dealer stock books, paired with editorial annotations that index some of those fields against authorities or other controlled vocabularies. Originally designed to support the generation of printed volumes, and then later a web-based interface for lookup of individual records, these legacy data speak mostly about *documents* of provenance events, and do so for an audience of human readers. To make these data linkable to museums that are producing their own Linked Open Data (following the general CIDOC-CRM principles of defining objects, people, places, and concepts through their event-based relationships), we are transforming these data in to statements about those provenance events themselves. In so doing, we are **standardizing** the terms referenced, **enriching** fields by turning them from transcribed strings into URIs of things, and explicitly **structuring** the relationships between these data as an RDF graph.

All this work requires dedicated labor. This leads to hard questions about priorities.

To what extent do we preserve the literal content of these documents, versus standardizing the way that we express the ideas those documents communicate (in so far as we, as modern-day interpreters, can correctly identify those ideas)? To maintain (to remember) plain text notes about, say, an object's materials as recorded by an art dealer, is to grant the possibility of perfect specificity about what our documents. But not aligning descriptions with authoritative terms for different types of materials and processes forecloses the possibility of generalizing about the history of those materials and processes across hundreds of thousands of objects. Remember too much, in other words, and we become Funes: incapable of synthetic thought.

Capacious collections data must remember enough *and* forget enough to be useful. For which terms will we expend the effort to do this reconciliation? Which edge cases will we try to capture in an ever-more-complex data model? Opinions on how to draw that line will frequently set the librarian, the historian, and the machine at cross purposes. Outlining the necessary competencies a collections data production team needs, and the key questions, in order to navigate perspectives must therefore be a crucial output of this forum.

Works Cited

Borges, Jorge Luis. 1962. “Funes, the Memorious.” In *Ficciones*, edited by Anthony Kerrigan, 107–15. New York: Grove Press.

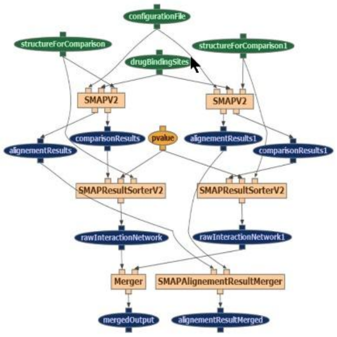
Drucker, Johanna. 2014. *Graphesis: Performative Approaches to Graphical Forms of Knowledge Production in the Humanities.* Cambridge: Harvard University Press.

Assessing Data Workflows for Common Data 'Moves' Across Disciplines

Alan Liu, University of California Santa Barbara

In considering how library collections can serve as data for a variety of data ingest, transformation, analysis, replication, presentation, and circulation purposes, it may be useful to compare examples of data workflows across disciplines to identify common data "moves" as well as points in the data trajectory that are especially in need of library support because they are for a variety of reasons brittle.

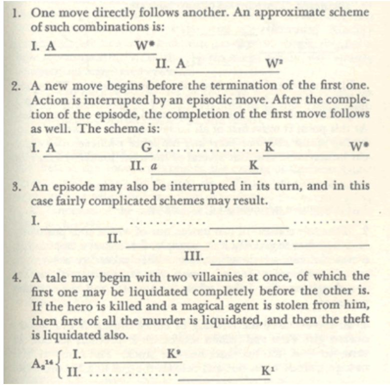
We might take a page from current research on scientific workflows in conjunction with research on data provenance in such workflows. *Scientific workflow management* is now a whole ecosystem that includes integrated systems and tools for creating, visualizing, manipulating, and sharing workflows (e.g., Wings, Apache Taverna, Kepler, etc.). At the front end, such systems typically model workflows as directed, acyclic network graphs whose nodes represent entities (including data sets and results), activities, processes, algorithms, etc. at many levels of granularity, and whose edges represent causal or logical dependencies (e.g., source, output, derivation, generation, transformation, etc.) *(see fig. 1)*. *Data provenance* (or "data lineage" as it has also been called in relation to workflows) complements that ecosystem through standards, frameworks, and tools--including the Open Provenance Model (OPM) the W3C's PROV model, ProvONE, etc. Linked-data provenance models have also been proposed for understanding data-creation and -access histories of relations between "actors, executions, and artifacts."[1] In the digital humanities, the in-progress "Manifest" workflow management system combines workflow management and provenance systems.[2]



The most advanced research on scientific workflow and provenance now goes beyond the mission of practical implementation to meta-level *analyses* of workflow and provenance. The most interesting instance I am aware of is a study by Daniel Garijo et al. that analyzes 177 workflows recorded in the Wings and Taverna systems to identify high-level, abstract patterns in the workflows.[3] The study catalogs these patterns as *data-oriented motifs* (common steps or designs of data retrieval, preparation, movement, cleaning/curation, analysis, visualization, etc.) and *workflow-oriented motifs* (common steps or designs of "stateful/asynchronous" and "stateless/synchronous" processes, "internal macros," "human interactions versus computational steps," "composite workflows," etc.). Then, the study quantitatively compares the proportions of these motifs in the workflows of different scientific disciplines. For instance, data sorting is much more prevalent in drug discovery research than in other fields, whereas data-input augmentation is overwhelmingly important in astronomy.

Since this usage of the word *motifs* is unfamiliar, we might use the more common, etymologically related word *moves* to speak of "data moves" or "workflow moves." A *move* connotes a combination of *step* and *design*. That is, it is a step implemented not just in any way but in some common way or form. In this regard, the Russian word *mov* for "motif," used by the Russian Formalists and Vladimir Propp, nicely backs up the choice of the word *move* to mean a commonplace data step/design. Indeed, Propp's diagrammatic analyses of folk narratives *(see fig. 2)* look a lot like scientific workflows. We might even generalize the idea of "workflows" in an interdisciplinary way and say, in the spirit of Propp, that they are actually *narratives*. Scientists, social scientists, and humanists do not just process data; they are telling data stories, some of which influence the shape of their final narrative (argument, interpretation, conclusion).

The takeaway from all the above is that a comparative study of data workflow and provenance across disciplines (including sciences, social sciences, humanities, arts) conducted using workflow modeling tools could help identify high-priority "data moves" (nodes in the workflow graphs) for a library-based "always already computational" framework.



1. One move directly follows another. An approximate scheme of such combinations is:

I. A W*
II. A W²

2. A new move begins before the termination of the first one. Action is interrupted by an episodic move. After the completion of the episode, the completion of the first move follows as well. The scheme is:

I. A G .... K W*
II. *a* K

3. An episode may also be interrupted in its turn, and in this case fairly complicated schemes may result.

I.
II.
III.

4. A tale may begin with two villainies at once, of which the first one may be liquidated completely before the other is. If the hero is killed and a magical agent is stolen from him, then first of all the murder is liquidated, and then the theft is liquidated also.

A₂¹⁴ { I. K⁹
II. .... K¹

One kind of high priority is likely to be very common data moves. For example, imagine that a comparative study showed that in a sample of *in silico* or data analysis projects across several disciplines over 40% of the data moves involved R-based or Python-based processing using common packages in similar sequences (perhaps concatenated in Jupyter notebooks); and, moreover, that among this number 60% were common across disciplinary sectors (e.g., science, social science, digital humanities). Then these are clearly data moves to prioritize in planning "always already computational" frameworks and standards.

Another kind of high priority may be data moves that involve a lot of friction in projects or in the movement of data between projects. One simple example pertains to researchers at different universities ingesting data from the "same" proprietary database who are prevented from standardizing live references to the original data because links generated through their different institutions' access to the databases are different. Friction points of this kind identified through a comparative workflow study are also high value targets for "always already computational" frameworks and standards.

Finally, one other kind of high priority data move deserves attention for a combination of practical and sensitive issues. Many scenarios of data research involve the generation of transient data products (i.e., data that has been transformed at one or more steps of remove from the original data set). A comparative workflow study would identify common kinds of transient data forms that require holding for reasons of replication or as supporting evidence for research publications. In addition, because some data sets cannot safely be held because of intellectual property or IRB issues, transformed datasets (e.g., converted into "bags of words," extracted features, anonymized, aggregated, etc.) take on special importance as holdings. A comparative workflow study could help identify high-value kinds of such holdings that could be supported by "always already computational" frameworks and standards.

Works Cited

[1] Hartig, Olaf. "Provenance Information in the Web of Data." In *Proceedings of the Linked Data on the Web Workshop at WWW*, edited by Christian Bizer, Tom Heath, Tim Berners-Lee, and Kingsley Idehen, April 20, 2009. http://ceur-ws.org/Vol-538/ldow2009_paper18.pdf.

[2] Kleinman, Scott. Draft Manifest schema. WhatEvery1Says (WE1S) Project, 4Humanities.org.

[3] Garijo, Daniel, Pinar Alper, Khalid Belhajjamey, Oscar Corcho, Yolanda Gil, and Carole Goble. "Common Motifs in Scientific Workflows: An Empirical Analysis." *2012 IEEE 8th International Conference on E-Science (e-Science)*, 2012: 1–8. doi: 10.1109/eScience.2012.6404427.

At the intersection of institution and data

Matthew Miller, New York Public Library

Libraries are awash in data, from the large reservoirs of bibliographic metadata that power discovery and access systems, to boutique datasets created from the documents themselves and even the ephemeral data exhaust produced by staff and patrons conducting research. Emerging from practical day-to-day working with this type of data below are some proposed observations and questions around description, distribution and access that are potentially useful and could benefit from closer examination.

The most potentially kinetic computationally amenable data comes from the conversion and processing of documents themselves. Transforming documents into data at the New York Public Library took the form of small projects that converted special collection materials into datasets through the power of algorithms, staff and the crowd. The results were a domain specific dataset often with a necessarily unique data model. Taking stock of the growing number these datasets we theorized about their possible integration with our traditional metadata systems. Would it be possible to go beyond simply linking to the dataset as a digital asset? If we were to build a RDF metadata system from the ground up could we begin thinking of it as an open-world assumption system where the contents of these datasets could exist alongside traditional bibliographic metadata? As more cultural heritage organizations continue to produce similar datasets we need to consider how they shape the next generation of our metadata and discovery platforms.

Stepping back from this larger question, when thinking about these resources as discrete datasets, what work could be done to improve their use and interoperability? WC3 standards such the VoID Vocabulary provided the means to describe the metadata about datasets. Leveraging such standards and establishing best practices and preferred authorities could we increase access across humanities datasets? How much work and what sort of resources are required to accomplish this at the dataset level and perhaps at the data level as well. For example using common non-bibliographic authorities such as Wikidata URIs in the data to facilitate interoperability across datasets and even institutions.

When publishing data for others it is a balance between providing access to the data in a format that provides the least friction for adoption and use versus how knowledge organization systems work within a cultural heritage institution. This often requires preprocessing of library metadata turning it into a more accessible form that does not require extensive domain knowledge. For example, when releasing the metadata for NYPL's public domain images we did not publish the MODs XML metadata, the format that it is inherently stored in our systems. Instead we opted to publish it as JSON and also as simple CSV files along with extensive documentation. Reducing the complexity of the format reduced the complexity of the tools and skills needed to work with it.

Another example taking this approach a step further is in Linked Jazz project in which we provided access to the data in the form of a SPARQL endpoint. The data, which is stored as RDF statements, represent a social network of Jazz musicians. This dataset lends itself to network analysis using popular tools such as Gephi. To make the application of such a tool as simple as possible we added a Gephi file export API allowing anyone to quickly download a gexf file of part of or the whole network to import into the software. This sort of scholarly API is geared for delivering the resources needed to begin utilizing the data immediately as opposed to just providing access to the underlying data store.

The topic of preprocessing introduces the question of best practices and standards that could be followed to ensure the broadest access to our datasets. What are some additional use cases that could drive shared best practices or tools for releasing cultural heritage data? Are there more advanced preprocessing that could be done to some of the common archetypical data formats found in libraries, archives and museums? And what sort of resources are required in an organization to process datasets for public consumption?

As institutions increasingly produce and release datasets, establishing some best practices around description, distribution and access can facilitate collaboration between organizations and ensure productive use of these resources by patrons.

Metadata and Digital Repository Accessibility Issues for Library Collections as Data

Anna Neatrour, University of Utah

In thinking of ways to use library collections as data, I was struck with the theme of accessibility. Are researchers genuinely invited to engage with library collections as data? I'm going to focus on this narrowly, looking mainly at aspects of metadata and technical infrastructure in digital repositories.

Metadata as invitation to computation

Encouraging usage of library collections as data could be embedded in digital collections metadata by including a statement that metadata is free to reuse, providing a CC0 license, or stating that metadata is open as a policy. One example of this is seen in the Harvard policy on open metadata. Many institutions have agreed that their metadata is in the public domain, which is a condition for harvest by DPLA, but there is often no metadata reuse statement available at the item or collection level in the source digital repositories for these shared collections. Making it clear that we expect metadata to be reused and repurposed improves the accessibility of digital library collections as data. Providing an easy way for researchers to download metadata in addition to a digital image might also encourage more research engagement with digital collections metadata. An example of this can be found in the University of Hull's repository, where records are easily downloaded in Mods or Dublin Core. In addition, highlighting investigations undertaken by repurposing library metadata within the digital repository itself could spark additional ideas for research from people who might be encountering this possibility for the first time.

Make digital repositories more welcoming

While offering access to digital collections via an API may be an effective way of showing that computation is possible with digital collections, it doesn't provide a welcoming environment for students or researchers who are at the initial stages of their research and who might not yet have the technical expertise to utilize an API. Providing a portal to a suite of sample apps created with an API, as DPLA does along with the search interface for a digital repository creates a signal that application development and computation utilizing a digital library is both possible and desired.

With libraries everywhere continually being asked to do more with less, curating all digital collections for computational purposes may be impossible. However, developing easy ways of bulk download for both images and metadata outside of an API may open up windows for researchers. Providing clear methods to download digital objects across different collections, or interact with images across repositories through a framework like IIIF could be yet another method for enabling researchers to interact with library collections as data.

Digital collection managers may be able to curate new local or regional corpora by thinking creatively about digital items they already own. For example, in my own library at the University of Utah, I've wondered about the possibility of making our typewritten oral history transcripts available to researchers. These oral histories were scanned as PDFs, and I expect the OCR would be decent enough to support text based topic modeling. Figuring out how to make these resources accessible to researchers by packaging them in a way that would encourage computational use is a goal of mine.

What does a digital collections as data repository look like?

Providing additional layers and portals that leverage computational exploration to existing collections might serve as an intermediate step. Imagine if text based digital collections also had a Voyant-like layer built into the digital repository itself that researchers could use, along with pre populated queries and visualizations so people at the beginning stages of inquiry could see examples of text analysis. This could support an introductory approach to exploring collections as data in the classroom. Many digital library repositories leverage visual possibilities for geospatial visualization and browsing, as in the Open Parks Network Map that shows thumbnail images of digital items along with map locations. Could an interface be built into a digital repository that would enable researchers to easily mash up digital items into a personalized portal that would support geospatial visualization without the need to download metadata, enhance information with coordinate data, and then create a more static map in an external system from that exported data? Could our digital repositories provide a mechanism for researchers to curate their own research collections, providing a space where digital library objects could be combined with researcher supplied data? Any approach have to blend what is pragmatically possible along with support for experimentation with the existing infrastructure for our digital repositories. Keeping in mind the idea of accessibility for researchers and library users at all stages of inquiry will hopefully result in an effective blend of solutions for interacting with library collections as data.

I'd like to thank Jeremy Myntti and Jim McGrath for providing feedback on a draft of this position statement.

Actually Useful Collection Data: Some Infrastructure Suggestions

Miriam Posner

Libraries and archives are increasingly making their materials available online, but, as a general rule, these materials aren't of much use for computational purposes. For the most part, institutions have sought to replicate as closely as possible the experience of being in a reading room with an individual object. We see this in artifacts like skeumorphic "swishes" on digital page-turns, mammoth lists of browsable topics, and, what concerns me most here, the inability to download large quantities of object metadata. Many of us have learned the basics of webscraping precisely to get around this problem, laboriously writing scripts to harvest metadata that we know must already exist somewhere, as data, in a repository.

There are many good reasons cultural institutions impose these limitations on their metadata. For one thing, it's not at all clear how many people actually *want* to treat collections as data. Most patrons aren't accustomed to encountering data in a cultural institution. So perhaps archives are just being good stewards of limited resources by focusing their attention on simply making digital facsimiles available. But the lack of collection data also limits other people's imaginations about what they might do with collections' materials.

I've also been told by various institutions that they don't have the right metadata for researchers to work with -- that their descriptive information is often schematic, high-level, and meant for search and discovery, not for visualization and analysis. I agree that this is a concern that we need to take seriously, but I contend that even the most basic metadata is often more useful for understanding a collection than many librarians imagine. Simply having author or creator information, or language information, can be very helpful. My impression is that many institutions are holding onto their data tightly, with the hope of cleaning and improving it in the future. But researchers can work with imperfect data, if its limitations are discussed frankly. We can also contribute improved data back to the institution.

Going forward, I imagine multiple pieces of infrastructure that could help make the data of cultural institutions as widely usable -- and widely *used* -- as possible:

A workable humanities data repository or registry. A good many open data repositories already exist. Most of them are designed to hold scientific data, although this need not disqualify them for humanities data. Humanists are actively contributing data (albeit on a relatively small scale) to general-use data repositories such as FigShare and Zenodo. The more troublesome problem is that a) consensus hasn't built around one particular repository; and b) absent a central repository, no substitute, such as a data registry, gathers lists of cultural data in one place. What cultural data exists is stored, for the most part, on GitHub — fine for downloading, versioning, and contributing data, but a terrible way to discover new datasets. We need a better way to find cultural data.

Consideration of APIs versus "data dumps." Many cultural institutions, reasonably enough, offer APIs as a means of accessing their data. This makes sense for a lot of different reasons, including access to the most recent data and the ability to retrieve institutions' data in many different ways. The problem here is that many humanists can work with structured data, but *not with APIs*. Many common visualization tools require no programming, and so it's possible for humanists to work with data, even in sophisticated, thoughtful ways, without necessarily knowing how to program. Developers at cultural institutions may feel that learning an API is trivial, but for many people, the availability of simple flat files can be the difference between using and not using a dataset. I therefore hope that cultural institutions will consider the possibility of providing unglamorous flat files, in addition to API access to their data.

Really lowbrow thought about data formats. Very simply, my students can work with CSVs, but not XML or JSON. Visualizing and analyzing the latter two formats takes programming knowledge, while even non-coders can import CSVs into Excel and create graphs and charts. Obviously, one can convert XML and JSON to CSVs, but doing this requires some knowledge of these formats, and sometimes some programming (or at least command-line) ability.

Case studies. It may seem unlikely, given the recent proliferation of digital humanities journals, but it's relatively difficult to find vetted, A-to-Z, soup-to-nuts examples of how to build visualizations and analysis from datasets. The aggregation of a number of fairly simple examples would, I believe, go far in demonstrating how people might use datasets in their own work, and would certainly be of great utility in the classroom. The key here would be to keep the examples quite simple, so that people can replicate and build on them with relative ease.

Interoperability and Community Building

Sheila Rabun, International Image Interoperability Framework (IIIF) Consortium

I am coming from a non-traditional background, with a Master's in interdisciplinary folklore studies, having gained the majority of my experience in libraries as the digital project manager and subsequently the interim director of the University of Oregon (UO) Libraries' Digital Scholarship Center. Among many digital projects, I was responsible for the Oregon Digital Newspaper Program, where we made large sets of newspaper OCR data and images available to the public online, following the Library of Congress' Chronicling America site and open API. While digital newspaper data has been used to create visualizations and other computational projects (for example, the Mapping Texts collaboration between the University of North Texas and Stanford University), the learning curve for scholars to find, harvest, and use the data provided remains a challenge. Students and faculty from all subject areas are increasingly looking to library and information professionals for guidance on where to find accessible data resources, how to use them, and recommendations on platforms for sharing their work. In addition to determining best practices for making collections available as data, comprehensive training materials and documentation for end users will be key to lowering the barrier of entry to make it easier for researchers to get started working with data on their own, encouraging wider re-use and experimentation.

Over the past 7 months I have shifted my focus slightly, as the Community and Communications Officer for the International Image Interoperability Framework (IIIF) Consortium, to improve digital image repository maintenance and sustainability as well as access and functionality for end users. As a community-driven initiative including national and state libraries, museums, research institutions, software firms, and other organizations across the globe, IIIF provides specifications for publishing digital image collection data to allow for interoperability across repositories. IIIF specifically addresses the "data silo" problem that has been plaguing the digital repository community, particularly by using existing standards and models such as JSON-LD and Web Annotation that make sharing and re-use easy. A growing number of digital image repositories are by adopting IIIF, and the IIIF Consortium has grown to include 40 institutional members since it was formed in 2015.

The IIIF community and specifications are especially relevant to the goals of the Always Already Computational (AAC) work, especially regarding digital images. IIIF has laid a groundwork for creation of a library collections as data as an internationally agreed-upon best practice for making digital image data shareable and more usable for study. IIIF utilizes JSON-LD manifests (representations of a physical object such as a book, as described in the IIIF Presentation API), to encourage sharing, parsing, and re-use of data regardless of differing metadata schemas across collections and repositories. The IIIF community has built the specifications specifically around use cases to solve real problems, so far primarily focusing on the needs of those both using and making available digitized manuscripts, newspapers, and museum collections.

We are currently working on extending the IIIF specifications to include interoperability for Audio and/or Visual materials (with 3D materials further along the roadmap), as well as improved discovery of IIIF-compatible resources on the web. Collaboration with the existing community that has formed around IIIF will be essential for the work of AAC and we welcome new interested parties to get involved, inform and provide feedback on approaches for discovery and stay informed with new innovations. Libraries and museums have been the primary adopters so far, but we have plans to do more outreach to scholars and researchers in all disciplines, STEM imaging providers, publishers, and the commercial sector. Vendors like CONTENTdm and LUNA have incorporated IIIF into their products, and IIIF is gaining speed in open source efforts like the Hydra-in-a-box repository product, which is IIIF-compatible. The goals of IIIF and AAC are in alignment, and there is an exciting potential to work more closely together, leveraging the existing IIIF community network and technical framework to create and build upon best practices.

From libraries as patchwork to datasets as assemblages?

Mia Ridge, British Library

The British Library's collections are vast, and vastly varied, with 180-200 million items in most known languages. Within that, there are important, growing collections of manuscript and sound archives, printed materials and websites, each with its own collecting history and cataloguing practices. Perhaps 1-2% of these collections have been digitised, a process spanning many years and many distinct digitisation projects, and an ensuing patchwork of imaging and cataloguing standards and licences. This paper represents my own perspective on the challenges of providing access to these collections and others I've worked with over the years.

Many of the challenges relate to the volume and variety of the collections. The BL is working to rationalise the patchwork of legacy metadata systems into a smaller number of strategic systems.[1] Other projects are ingesting masses of previously digitised items into a central system, from which they can be displayed in IIIF-compatible players.[2]

The BL has had an 'open metadata' strategy since 2010, and published a significant collection of metadata, the British National Bibliography, as linked open data in 2011.[3] Some digitised items have been posted to Wikimedia Commons,[4] and individual items can be downloaded from the new IIIF player (where rights statements allow). The BL launched a data portal, https://data.bl.uk/, in 2016. It's work-in-progress - many more collections are still to be loaded, the descriptions and site navigation could be improved - but it represents a significant milestone many years in the making. The BL has particularly benefitted from the work of the BL Labs team in finding digitised collections and undertaking the paperwork required to make the freely available. The BL Labs Awards have helped gather examples for creative, scholarly and entrepreneurial uses of digitised collections collection re-use, and BL Labs Competitions have led to individual case studies in digital scholarship while helping the BL understand the needs of potential users.[5] Most recently, the BL has been working with the BBC's Research and Education Space project,[6] adding linked open data descriptions about articles to its website so they can be indexed and shared by the RES project.

In various guises, the BL has spent centuries optimising the process of delivering collection items on request to the reading room. Digitisation projects are challenging for systems designed around the 'deliverable item', but the digital user may wish to access or annotate a specific region of a page of a particular item, but the manuscript itself may be catalogued (and therefore addressable) only at the archive box or bound volume level. The visibility of research activities with items in the reading rooms is not easily achieved for offsite research with digitised collections. Staff often respond better to discussions of the transformational effect of digital scholarship in terms of scale (e.g. it's faster and easier to access resources) than to discussions of newer methods like distant reading and data science.

The challenges the BL faces are not unique. The cultural heritage technology community has been discussing the issues around publishing open cultural data for years,[7] in part because making collections usable as 'data' requires cooperation, resources and knowledge from many departments within an institution. Some tensions are unavoidable in enhancing records for use externally - for example curators may be reluctant or short of the time required to pin down their 'probable' provenance or date range, let alone guess at the intentions of an earlier cataloguer or learn how to apply modern ontologies in order to assign an external identifier to a person or date field.

While publishing data 'as is' in CSV files exported from a collections management system might have very little overhead, the results may not be easily comprehensible, or may require so much cleaning to remove missing, undocumented or fuzzy values that the resulting dataset barely resembles the original. Publishing data benefits from workflows that allow suitably cleaned or enhanced records to be re-ingested, and export processes that can regularly update published datasets (allowing errors to be corrected and enhancements shared), but these are all too rare. Dataset documentation may mention the technical protocols required but fail to describe how the collection came to be formed, what was excluded from digitisation or from the publishing process, let alone mention the backlog of items without digital catalogue records, let alone digitised images. Finally, users who expect beautifully described datasets with high quality images may be disappointed when their download contains digitised microfiche images and sparse metadata.

Rendering collections as datasets benefits from an understanding of the intangible and uncertain benefits of releasing collections as data and of the barriers to uptake, ideally grounded in conversations with or prototypes for potential users. Libraries not used to thinking of developers as 'users' or lacking the technical understanding to translate their work into benefits for more traditional audiences may find this challenging. My hope is that events like this will help us deal with these shared challenges.

Works Cited

[1] The British Library, ‘Unlocking The Value: The British Library’s Collection Metadata Strategy 2015 - 2018’.

[2] The International Image Interoperability Framework (IIIF) standard supports interoperability between image repositories. Ridge, ‘There’s a New Viewer for Digitised Items in the British Library’s Collections’.

[3] Deloit et al., ‘The British National Bibliography: Who Uses Our Linked Data?’

[4] https://commons.wikimedia.org/wiki/Commons:British_Library

[5] http://www.bl.uk/projects/british-library-labs, http://labs.bl.uk/Ideas+for+Labs

[6] https://bbcarchdev.github.io/res/

[7] For example, the 'Museum API' wiki page listing machine-readable sources of open cultural data was begun in 2009 http://museum-api.pbworks.com/w/page/21933420/Museum%C2%A0APIs following discussion at museum technology events and on mailing lists.

Maintaining the ‘why’ in Data:
Consider user interaction and consumption of library collections

Hannah Skates Kettler, University of Iowa

Always Already Computational represents the next hurdle for libraries, archives and museums. Now that the profession is comfortable with the notion of digitization, and have reaped the rewards of greater and broader impact (Proffitt and Schaffner, 2008), it has now turned its focus towards born digital materials. It's not that born digital materials, in 2017, is a new notion but it is definitely a concept the profession has been aware of, but has been hesitant to tackle. As a Digital Humanities professional, I deal with the use and creation of born digital materials every day and adapt to the multiplicitous ways library collections are created and made available, especially in the Humanities.

I therefore approach the questions in Always Already Computational with these concepts in mind:

Relational Datasets:

No library collection is an island. Library collections are not simply a list of ones and zeros that wait to be consumed and reused, then spat out again as something different. At least, not when we want to be able to cite them. Data (which henceforth will be a stand in for 'library collections') must be persistent in order to be effectively accessible and reused for research. In order to amalgamate various datasets, immense amount of time is spent standardizing the data into something that can be cross referenced and used computationally. Understanding that our data are unique, it does not necessarily follow that access should be as unique and idiosyncratic. What that Linked Data has provided is a framework to link disparate ideas to each other relationally. I am particularly interested in the possibilities of the Linked Data at it applies to datasets that would allow one to describe contextual relationships between the data, relationships which typically are entirely use and user based. By generalizing data in a way that is useful in multiple contexts by creating a framework that is flexible enough to accommodate data's multiplicity.

Association of Paradata:

Pulling from experience with 3D collections, functioning without standards of how to make born digital materials more usable makes interfacing with other datasets much more difficult than other more traditional data. For example, visual materials are much more reliant on supplemental contextual data than text. That is not to say there is no context within textual data, but the aforementioned data could include context within it. Visual data, usually lacks this packaged approach. Visuals are associated with text in order to provide that context. Beyond catalogues, visual data's supplemental material is separated from and unintentionally disassociated from the visual (think a search result in an image database). Few image datasets are accompanied with *why* the image was created. True, one can inference based on the basic metadata included with the object, but without intent, it is much more difficult to make judgement about why the dataset (as generated by an API for instance) is included and why others were not. It also makes it easier to fake, or misrepresent library data/collections.

Cultural Constructs of Data:

Compounding the narrowed context of textual and numerical datasets, problematic visual datasets, and even mixed data sets, you have the social constructs that support data. This aligns very well with the work I, and a group of librarian and museum professionals are doing in association with the Digital Library Federation. As was mentioned in the October 2004 Information Bulletin from the Library of Congress, "Because there is no analog (physical) version of materials created solely in digital formats, these so-called 'born-digital' materials are at much greater risk of either being lost and no longer available as historical resources, or of being altered, preventing future researchers from studying them in their original form." Their particular focus for this remark was the preservation of born-digital data. Now that the profession, to some extent, has the ability and focus for preservation of born-digital, it is time to turn our eye to interoperability (like Always Already Computational) and the cultural context of the data itself. Consider the book *The Intersectional Internet: Race, Sex, and Culture Online* by Safiya Noble and Brendesha Tynes (2016) which underscores "how representation to hardware, software, computer code, and infrastructures might be implicated in global economic, political, and social systems of control." Data without context is meaningless. Data with context but without social awareness is deceptively meaningless. With that deception comes, in the worst case, the use and articulation of argument founded on a lack of understanding and awareness of perpetuating ideas that are intrinsically linked to the creation and curation of said data. A question for this group would be; how do we attempt to preserve that context without overwhelming the user?

The Always Already Computational group can hopefully come together to attempt to solve this and other concerns regarding digital aggregate data.

References

"Born Digital': Eight institutions and their partners received awards totaling almost $15 million from the Library to collect and preserve digital materials as part of the National Digital Information Infrastructure and Preservation Program". 2004. *Library of Congress Information Bulletin.* 63 (10): 202-203.

Noble, Safiya Umoja, and Brendesha M. Tynes. 2016. The intersectional Internet: race, sex, class and culture online. ISBN: 978-1-4331-3000-7.

Proffitt and Schaffner. 2008. The Impact of Digitizing Special Collections on Teaching and Scholarship: Reflections on a Symposium about Digitization and the Humanities. Report produced by OCLC Programs and Research. Published online at: www.oclc.org/programs/reports/200804.pdf

People and machines both need new ways to access digitized artifacts nonconsumptively

Ben Schmidt, Northeastern University

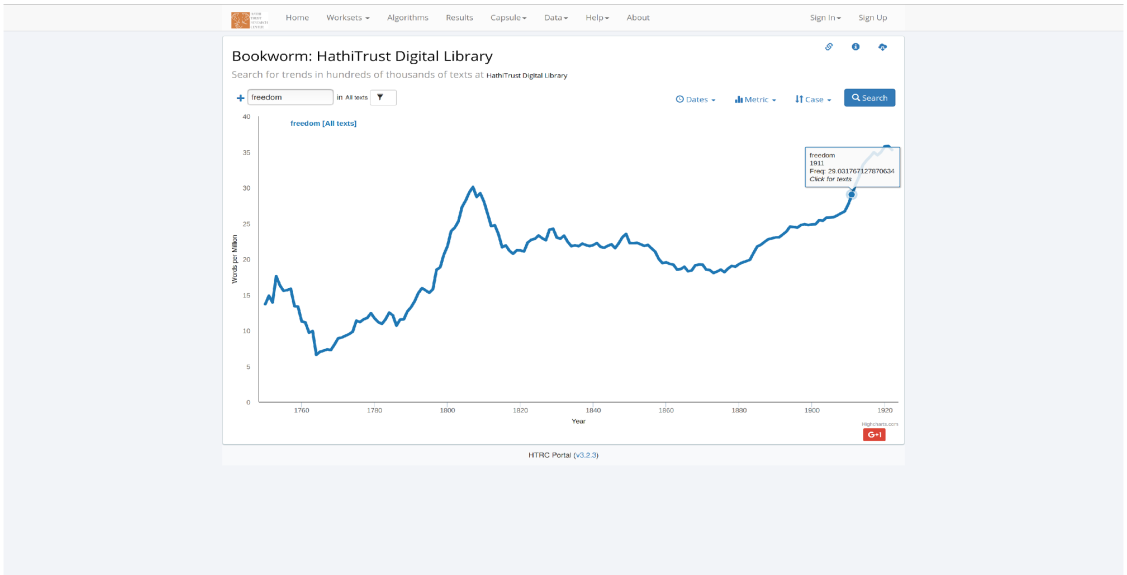
How can we integrate generations of high-quality, professionally-created metadata with electronic versions of the object itself? Particularly when copyright comes into play, we can't simply hope for openness; and there's a steep trade-off between the thoroughness of a well-thought-out standard and a simplicity of conception that makes a digital resource useful for (for instance) a graduate student just beginning to get interested in working with large collections.

When we digital humanities researchers say that we're working with the "full text" of a scanned book, it's usually more posturing than truth. In fact, what datasets like the Hathitrust Research Center's Extracted Features really do is just radically transform the amount of metadata we have; instead of knowing 10 or 20 things from a MARC record (eg: the language, four or five subject headings, the author, the publisher), we just add on an additional several thousand ("How many times does it use the word "aardvark?" "aardvarks?" "abacus?"...). All the rest of the information (even simple stuff like syntax, word order, negation) is thrown out. It's great that organizations like JStor and Hathi are starting to release this computationally-derived metadata. But there's no clear way to incorporate this computational metadata into a traditional library catalog. The technical demands of even *downloading* something like the HTRC EF set exceed both the technical competencies and computing infrastructure of most humanists--I've literally spent several weeks recently, restarting downloads and identifying missing files as I try to fill up a RAID array with several terabytes of data. Processing these files into the raw material of research is even harder.

So how do we make collections accessible for work? There are two ways that libraries can take more of the burden onto themselves, and distribute (non-copyright-violating) distillations of texts that provide an onramp for digital analysis within the reach of mere mortals.

Visual Exploration

One useful and important way to work with this metadata and full text is by exposing through visualization; this is what projects like the Google Ngrams viewer and the Hathi+Bookworm project I've helped work on under an NEH grant. Patrons are able to use this combination of full text and catalog metadata to explore the shapes and contours of vast digital libraries. Since they know (sort of!) what any given word means, they can use it to understand how vocabulary changes; find anomalous, interesting, or misclassified items; or understand the limits and constraints of an entire collection, a sorely-needed form of information literacy. We've built the Bookworm platform so the advances we're making with Hathi can be used on any smaller (or larger) library, and we hope others will be interested in using to explore their texts in the context of their metadata.



Bookworm: HathiTrust Digital Library

Search for trends in hundreds of thousands of texts at HathiTrust Digital Library

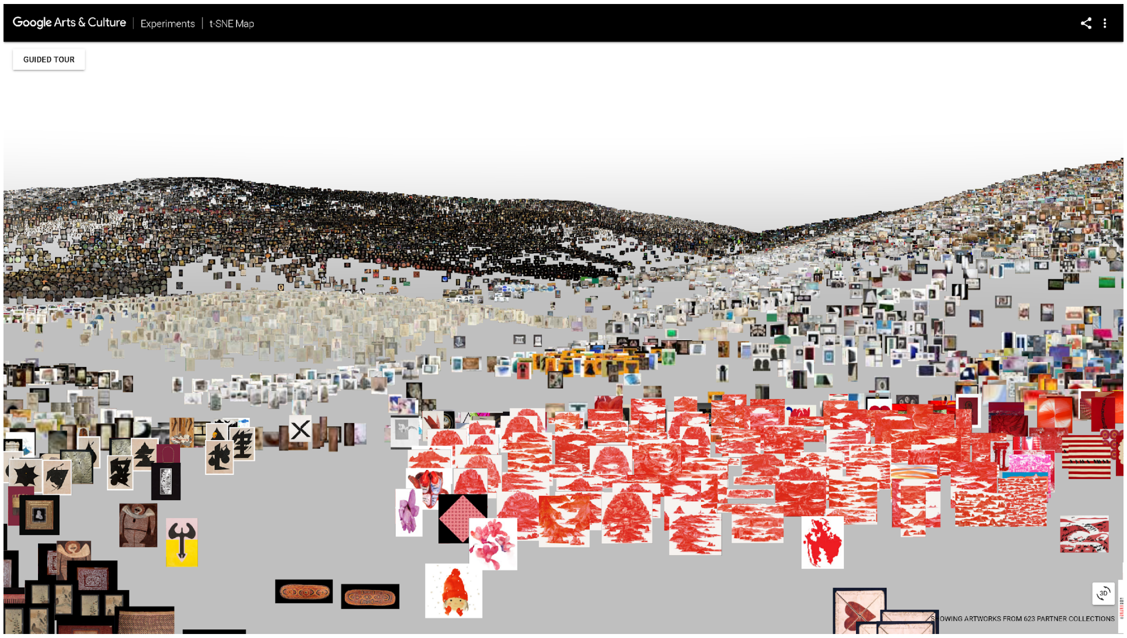
Hathi Trust Bookworm browser

Low-dimensional embeddings

I'd also like to put on the radar a farther out-there idea that extrapolates from the current trends in the world of machine learning: the idea of a *shared embedding* for digital items that would allow machines to compare items across various collections, times, and artifacts. The basic idea of an embedding is to associate a long list of numbers (maybe a few hundred) with a digital object so that items that are similar have similar lists of numbers. These are sort of the inverse of the checksums that libraries frequently associate with digital artifacts now, which are designed so that even the slightest change makes a file get a completely different number. A good embedding will do the opposite; allow users and software to find *similar* items. In a single collection like Hathi, this practice I've found with even a simple embedding that it's possible to, for instance, look in the neighborhood of a book like "Huckleberry Finn" and find, in the immediate neighborhood, dozens of titles like "Collected Works of Mark Twain, vol. 8" that lack proper titles that would identify them; and in the extended neighborhood other novels about American boys on riverboats.

Inside a collection, this makes it possible to find works with improbable metadata. (It's sadly common for the wrong scan to be associated with metadata, and this can be extremely hard to catch.) Across collections, this makes it possible to engage in the work of comparison, duplicate detection,

Perhaps the most interesting things about embeddings of digital files is that they're *not* restricted to textual features. Image embeddings are just as possible as textual embeddings, as in this landscape visualization of artworks that Google recently produced.



When Google recently released half a million hours of video, they did it not as image stills but as vectorized features read by a neural network.

These features--essentially, a computer's rough summary of an artifact into a few hundred numbers--could make it possible to researchers and students to immediately engage in computational analysis without having to wade through the preparatory steps. If done according to shared standards, they could make collections interoperable in striking ways *even when texts or images can't be distributed*. It's probably a few years too early to set a specific embedding for different types of documents, but it is time now to contemplate what it would mean to distribute not documents themselves, but a useful digital shadow of them.

Repurposing Discographic Metadata and Digitized Sound Recordings as Data for Analysis

David Seubert, University of California Santa Barbara

Use of sound recordings for research has been slow to develop due to bias against sound recordings as historical documents by textual scholars, lack of descriptive data (discography), and lack of access because of restrictive copyright laws that make it difficult to digitize and provide access to collections. The use of digitized sound recordings or the discographic metadata about sound recordings as data to study is underdeveloped. The UCSB Library wants to encourage scholarship of this kind using the data from the American Discography Project.

The American Discography Project that is presently based at the UCSB Library with funding from the Packard Humanities Institute was originally conceived as the Encyclopedic Discography of Victor Recordings by two record collectors in the early 1960s. They began a project to document every classical recording by the Victor Talking Machine Company, but eventually broadened their goal to include every Victor recording session for 78rpm discs. In 1966 they were granted liberal access to the recording files held by RCA Victor Records (now Sony Music Entertainment) and devoted many thousands of hours to compiling lists of the tens of thousands of Victor master recording sessions from around the world.

The American Discography Project and its principal product, the Discography of American Historical Recordings (DAHR) is now a research, publication, and digitization program based at the UCSB Library with a goal of documenting disc recordings made during the standard groove era (1900-1950s) by American record companies and to digitize as many as possible for online access. Much of the data about a recording (who, what, where, when) is not documented on the recordings themselves, and only can be determined by consulting a published discography or primary source documents like company recording ledgers.

Now in its fifth decade, the project has expanded beyond Victor to incorporate other published discographies and includes data on recordings made by five early 20th century record companies (Berliner, Victor, Zonophone, Columbia and Okeh) with three more large labels (Brunswick, Decca, Edison) and several smaller ones in the pipeline.

The sheer amount data documented in the online database is significant. DAHR currently contains over 6.5 million data points documenting systematically and comprehensively the first 45 years of American recording history including:

- 146,524 recording sessions
- 417,428 recording events (takes)
- 107,784 physical manifestations (discs)
- 36,767 names of performers, authors, composers
- 90 languages
- 393 recording locations

The initial project design was to document these recordings in a systematic fashion for the purposes of identification, cataloging by libraries and archives, collectors, and others. A bibliography of sound recordings. One of the further goals of the project is to encourage use of sound recordings as primary source documents by scholars in fields beyond the study of music and as the project has grown, we have growing success in this area. Systematically adding audio to the database has allowed scholars to study the recordings, in context with authoritative data about their creation.

Sound recordings and the metadata associated with them have not been mined and analyzed the way textual archives have. As the Discography of American Historical Recordings grows in size, it is a prime candidate for manipulation and analysis as data, as it contains standardized elements including language, dates, geographic information (recording locations), genres, names, and titles.

Since the project was designed from the outset to be structured data, including authority control and standardized vocabularies for many elements, a potential and as yet unrealized reuse of the metadata as data, is now possible. As a participant in the National Forum, we hope to be able to further conceptualize how this can be best realized.

The Library as Virtual Reality: A Worldbuilding Approach

Laila Shereen Sakr, University of California Santa Barbara

The process of considering digital library collections as data points relies on similar logics foundational to the development of virtual reality (VR). Imagine the library as a VR film or as a computer --- temporally and spatially. If the goal of the "Always Already Computational: Library Collections as Data" project is to find a common framework among librarians, curators, and researchers that makes digitally-born scholarship possible, I would like to suggest considering speculative design methodologies, or what Alex McDowell has described as worldbuilding.

Alex McDowell, a deeply influential designer has shifted how we think about design by fundamentally changing the role design plays in the creative process, potentially altering audiences' expectations of creative work that ranges from architecture to computer games. Drawing on the literary metaphor "worldbuilding" to explain his approach to design, McDowell's methods represent a cultural shift in his industry's production process. Speculating about what the world "might" look like in the future is easy. More challenging, though, is realizing that speculative vision through the design process. McDowell's work realizing a future-world inspired by Philip K. Dick's novella in the 2002 film *Minority Report* is emblematic of a transformation in design process that is made possible through the use of computational media. On *Minority Report*, McDowell led his production design team, which began as a largely analog art department, through a transition in which they became the first fully-digital art department in the film industry — an example that many other design departments would soon follow and that foreshadowed a broader cultural shift in creative process.

Most of the film's audience will probably remember the gestural interface of the 3D screens used by the agents in the department — speculative designs that, in turn, have influenced actual technologies ranging from Apple's iPad to Microsoft's Kinect. However, *Minority Report*'s influence in design reached an even wider array of design cultures, including biometrics (particularly retinal scanning), through other imagined technologies woven throughout the film's environment and plot.

In other words, McDowell's world building integrates interdisciplinary humanistic, scientific, and design inquiry with emerging forms of computational media to fundamentally alter the film production process, blurring boundaries between physical and virtual environments and the distinctions between film and other media forms. In the digitally designed world of *Minority Report*, props could be modeled first as two-dimensional images and later as three-dimensional physical objects. Then, through computer-controlled milling, those models could be used to create final props by sculpting and mold-making. Bringing direction, cinematography, and design together in the virtual space of the pre-visualization stage, props, actors, and the created world interacted throughout the production process. As a result, *Minority Report* and McDowell's world building process signaled a transformation in design culture that has not yet fully played out.

One approach to worldbuilding builds upon a procedure of information design that moves from archiving, to visualizing, to rationalizing, and then to governing. This process must take into account matters of scale. Taking from both information design and game design, worldbuilding relies on several distinct way visual perspectives: drawing a complete world map and filling in as much information as possible, then running the game and letting the players explore that world. This visual perspective operates on a large scale. Another perspective begins within specific town/city/place/room...and as they explore more and more of the world is revealed. These are some basic guidelines to consider as one conceptualizes building a virtual word of data.

Applying this theoretical framework to a process of speculative design for future library collections, could yield interesting results. The practice and ideas of worldbuilding, in McDowell's definition, are a clear example of interdisciplinary work connecting the arts, design, media-focused computer science, and elements of the humanities and social sciences. Worldbuilding is both the creation of media and a design research practice, and in neither case is its interdisciplinarity a luxury, because the work simply must engage multiple disciplines in order to achieve a coherent vision and to push many fields forward.

The struggle for access

Tim Sherratt, University of Canberra

For me, exposing cultural heritage collections to computational methods raises difficult, important, and interesting questions about the nature of 'access' itself. So while we can and should develop best-practice guidelines, I think we should also admit that we will never be, should never be, satisfied with what cultural institutions deliver. We will always want something more. And that's a good thing.

I've spent far too much of my life hacking the web interfaces of libraries and archives in the pursuit of useful data. But while I would gladly take the time back, I recognise the value of the struggle. Processes such as screen-scraping and normalisation are often frustrating, but they do at least make you think about the processes by which the data was created, managed, and shared.

So for me, one of the key questions is how we expose data to facilitate the use of computational methods while preserving some of the difficulties and irregularities – the chisel marks in the smooth worked surface – that remind us of its history and humanity.

I'm not sure whether this is a metadata question, or a matter of how we frame the relationship between researcher and institution. If we think of machine-actionable data as a product or service delivered by institutions, then researchers are cast as clients or consumers. But if each dataset is not a product, but a problem, then we open up new spaces for collaboration and critique.

I've started to realise that I have very little interest in statistics, or even data visualisation as I understand it. I use computational methods to manipulate the contexts of cultural heritage collections. Sometimes this results in useful tools or interfaces, sometimes it's more akin to art. I'm motivated by the simple desire to see things differently – to poke at the boundaries and limits of systems in the hope that something interesting happens.

What seems to happen fairly regularly is that I find where the systems are broken. For example, while harvesting debates from the Australian parliament's online database, I discovered about 100 sitting days were missing. This sort of thing happens with complex systems, and the staff at the Parliamentary Library have now fixed the problems. For me, it's an example of the fact that we can never simply accept what we're given – search interfaces lie, and datasets have holes. But it's also shows that once you open up channels for the transmission of data, information flows both ways.

We can't talk about the need for institutions to provide computation-ready data without considering what they might get in return. The struggle for access might not always be comfortable, but it can be productive. If data is a problem to be engaged with, rather than a service to be consumed, then we can see how researchers might help institutions to see their own structures differently. On a practical level, how might we make it easier for institutions to re-ingest the features and derivative structures identified through use.

I'm also a bit suspicious of scale. Big solutions aren't always best. Large data dumps are great for researchers with adequate computing power and resources, but APIs support rapid experimentation and light-weight interventions. Similarly, while articulating best-practice for computation-ready data we shouldn't lose sight of other ways data can be exposed. I want hackable websites as well as downloadable CSVs – all that basic stuff like persistent urls, semantic html, and maybe a sprinkle of RDFa or JSON-LD, enables data to be discovered everywhere, not just in a designated repository.

As I said, we will always want more. Access will never be open and the job will never be done. We need systems, protocols, guidelines, and collaborations that remind us there is always more to do, and offer the support to continue.

Implications for the Map in a 'Collections as Data' Framework

Tim St. Onge, Library of Congress

I am arriving of the challenge of developing computationally amenable digital library collections from the perspective of a digital cartographer and geospatial analyst. My work for the Library of Congress as a cartographer primarily involves digital map-making and the analysis of born-digital and made-digital geographic information and maps to serve Congressional research requests. My academic and professional backgrounds are based in geographic information science (GIS) rather than in library science. However, I am often thinking about how the Library of Congress can best serve our collections to meet the research and access needs of geographers in a digital age.

All of this is to say that my initial thoughts on developing a “library collections as data” framework are largely shaped by the implications for one type of collection material in particular: the map.

There is enormous potential for the computational analysis of historic maps en masse, with methods that are both text-based (e.g. extracting written text to create gazetteers of place names from certain time periods, cultures, languages, etc.) and image-based (e.g. extracting map features based on groupings of image pixel values of similar color) (Chiang, Leyk & Knoblock 2014). For the full integration of historic maps into Geographic Information Systems, processes like georeferencing and feature digitization, which have achieved varying levels of automation potential, must be completed. It is my view that georeferenced versions of scanned maps in library collections are highly appreciated among researchers and should be more standard “collections as data” offerings from libraries. The georeferenced map viewer created by the National Library of Scotland (2017) demonstrates the tremendous value of this type of data offering.

Given the unique challenges of offering historic maps as computationally amenable collections, I admire the objective of the Always Already Computational to conceive of a “collections as data” framework that is multimedia in scope and not only concerned with text analysis of written works (as critically important and valuable as this is).

In my reading of the “Statement of Need” from the Always Already Computational scope of work document, I interpret four major current problems of computationally amenable collections to be (1) the lack of a common collections-transformation framework across institutions, (2) a lack of solutions for non-text media, (3) technical inadequacies in providing collections in large scale, and (4) no data reuse paradigm for collections.

In addressing the first and second problems, I look forward to hearing more on the needs of computational researchers who are working with image-based collections, including, but not exclusively, scanned and digitized maps. In this needs assessment more broadly, in an abstract way, I imagine a hierarchy of use cases and analysis tools. Towards the top are elements that are most readily shared among all kinds of library collections (e.g. all collection items have metadata files in standard format; all text-based, text-extracted items could undergo analyses like frequency visualization or topic modeling). Towards the bottom are more medium-specific (e.g. only scanned maps are concerned with georeferencing and geographic projections). In laying out the strongest commonalities among researcher needs in working with library collections, perhaps a framework can be developed that addresses the greatest, unifying needs of collection patrons across diverse uses in the digital humanities and other disciplines. Furthermore, I hope that this framework highlights the unique and worthy challenges of devising solutions for researchers of non-text media.

The third problem of providing collections on a large scale is certainly a critical concern to computational research. If access to collection items is limited to one-by-one downloads or deliveries of physical DVDs of data, simply the "data acquisition" phase can be sufficiently burdensome to slow or stop computational analyses before they even begin. The challenges of large-scale collection access appear to be technological and, as is often the case for libraries and the digital humanities, budgetary. The methods of access detailed in the Always Already Computational scope of work document demonstrate the wide variability among different institutions. I am interested to hear from project participants on the merits of these methods from their experience and what technical and budgetary considerations should be made in the process of developing best practices on this issue.

On the fourth problem of the data reuse paradigm, I believe this issue involves not only technological hurdles, but policy ones as well. Simply put, when researchers or patrons more broadly want to give back to libraries, libraries should trust them. For example, this can take the form of an online-based crowdsourced georeferencing tool that allows users to georeference scanned maps from a library collection and share them back to the library, which thereby shares that resource universally as a GIS-ready raster image (Fleet, Kowal, & Přidal 2012). Another example would be for libraries to host hackathons and other events that invite researchers to interrogate their collections as data and present on their findings, thereby allowing libraries learn lessons of the kinds of computational research that can (or cannot) work with their collections. I believe the Archives Unleashed series, which focuses on web archive research, is a great model for this kind of project (Weber 2016). Any frameworks arising from the Always Already Computational should encourage these kinds of "data sandbox" projects that allow for experimentation that reveal new insights into the computational analysis of collections as data and provide derived content and research directly back to libraries.

I look forward to learning from the diverse array of participants and contributing my insights to the Always Ready Computational initiative.

Works Cited

Chiang, Y., Leyk, S., & Knoblock, C. A. (2014) A survey of digital map processing techniques. *ACM Computing Surveys*, 47 (1), Article 1 (April 2014), 44 pages. Retrieved from http://usc-isi-i2.github.io/papers/chiang14-acm.pdf.

Fleet, C., Kowal, K. C., & Přidal, P. (2012) Georeferencer: Crowdsourced Georeferencing for Map Library Collections. *D-Lib Magazine*, 18 (11/12). Retrieved from http://www.dlib.org/dlib/november12/fleet/11fleet.html.

National Library of Scotland (2017) *View maps overlaid on a modern map / satellite image*. Retrieved from http://maps.nls.uk/geo/explore/.

Weber, M. S. (2016) Archives Unleashed! *Collections as Data | September 27, 2016 | Library of Congress*. Retrieved from http://digitalpreservation.gov/meetings/documents/dcs16/3_Weber_ArchivesUnleashed.pdf.

Considering the user

Santi Thompson, University of Houston

As the forum unfolds, I would encourage participants to question and expand our assumptions of those who (re-)use computational library collection data. In my mind, the identities of users and their motivations for coming to the digital library are just as important to understand as the technical requirements needed to re-use data in interoperable and collaborative ways. Knowing your users helps cultural heritage professionals, among other things, to better select content for the future, market the resources and collections available to them, and understand how to describe and make content available to others.[1]

I was pleased to see that the proposal for *Always Already Computational* acknowledges the user to some degree, noting that current digital library infrastructure and digital collection paradigms do "not meet the needs of the researcher, the student, the journalist, and others who would like to leverage computational methods and tools to treat digital library collections as data." As such, part of our forum objectives will be to draft potential user stories and “to apply [data definitions and concepts] to a range of potential user communities.” I find this to be incredibly important because libraries (and most likely other cultural heritage organization types) have not spent a vast amount of time asking and publishing on “who is a digital library user.”

My own research has focused in some narrow ways on better understanding digital library users. My collaboration with other members of the DLF Assessment Interest Group’s User Studies Working Group has found that the assessment of digital library reuse is complicated for a whole host of reasons, including the profession’s inability to systematically identify and understand digital library users.[2] Additional research I have done with a co-author suggests that digital library users (note: **NOT** users of computational data) are more frequently (1) from outside of academia and (2) reusing digital library content for a wide array of non-scholarly pursuits.[3]

I find *Always Already Computational* to be an exciting opportunity to address major gaps in our current understanding of what is a digital library collection and how is it being used by targeted audiences. While I recognize that demystifying the digital library user is not the primary pursuit of this national forum, I look forward to discussing this as well as other important aspects of the grant with a deeply knowledgeable and inspiring group of participants. I appreciate the opportunity to contribute to such a discussion.

Works Cited

[1] For more on how understanding users and reuses can inform digital library management, see my work with Michele Reilly: “Understanding Ultimate Use Data and Its Implication for Digital Library Management: A Case Study,” *The Journal of Web Librarianship 8* (2) (2014): 196-213. DOI: http://dx.doi.org/10.1080/19322909.2014.901211.

[2] In 2015 the User Studies Working Group drafted a white paper, “Surveying the Landscape: Use and Usability Assessment of Digital Libraries,” that explored the state of research around three assessment topics: user/usability studies, return on investment, and content reuse. A copy can be found here: https://osf.io/uc8b3/.

[3] See Reilly and Thompson, “Understanding Ultimate Use,” and Michele Reilly and Santi Thompson, “Reverse Image Lookup: Assessing Digital Library Users and Reuses,” *The Journal of Web Librarianship* (2016): 1-13. DOI: http://dx.doi.org/10.1080/19322909.2016.1223573.

Building Institutional and National Capacity for Collections as Data

Kate Zwaard, Library of Congress

About a year ago, the Library of Congress created a new division, National Digital Initiatives, which I am proud to lead. Our mission is to maximize the benefit of the digital collection, to incubate innovation, and to encourage national capacity for digital cultural memory.

In a recent New Yorker article, the Librarian of Congress said she wants The Library of Congress “to get to the point where there’ll still be a specialness, but I don’t want it to be an exclusiveness. It should feel very special because it *is* very special. But it should be very familiar [1]” We in NDI take that message to heart. We believe that an important step in getting users to engage with the Library’s digital material and staff is to provoke, explore, tell stories, and invite.

EVENTS
OUTREACH
PROGRAMS
EXPERIMENTS
INVESTIGATIONS
PARTNERSHIPS

```
0110000101000100000011101000101010101001
0110000101000100000011101000101010101001
0110000101000100000011101000101010101001
0110000101000100000011101000101010101001
0110000101000100000011101000101010101001
0110000101000100000011101000101010101001
0110000101000100000011101000101010101001
0110000101000100000011101000101010101001
0110000101000100000011101000101010101001
```

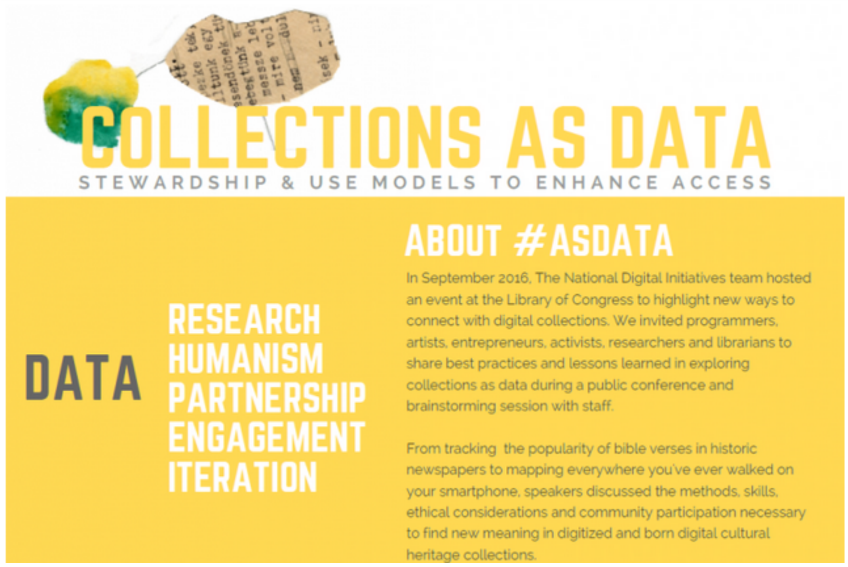
Our vision is for NDI to help libraries and patrons explore the edges of possibility. To try things ourselves and share with the profession. To help highlight the treasures we have -- here at the Library of Congress and in our nation’s cultural heritage institutions – and spark people’s imagination around the potential uses of digitized or born digital collection objects. To encourage the curious and help them get answers.

To help people understand what a library is.

Upon our founding, the director of National and International Outreach said “It’s not enough anymore to just open the doors of this building and invite people in. We have to open the knowledge itself for people explore and use. [2]”

A few things we’ve been working on:

- We organized “**Collections as Data**,” [2] a conference devoted to exploring what’s possible using computation with digital collections.
- We hosted an **Archives Unleashed hackathon**, bringing together programmers, librarians, and scholars looking at computational analysis of web archives collections [4]
- We performed a **digital lab proof of concept** along with a report exploring how to deliver Library of Congress digital collections as data to on-site researchers [5]
- We hosted a **Software Carpentry Workshop** [6] to help teach Library of Congress librarians and others in the neighborhood how to use code to manage and analyze digital collections.
- We’ve started a series of **sample code notebooks** to help people work with Library of Congress data [7]



COLLECTIONS AS DATA

STEWARDSHIP & USE MODELS TO ENHANCE ACCESS

DATA

RESEARCH
HUMANISM
PARTNERSHIP
ENGAGEMENT
ITERATION

ABOUT #ASDATA

In September 2016, The National Digital Initiatives team hosted an event at the Library of Congress to highlight new ways to connect with digital collections. We invited programmers, artists, entrepreneurs, activists, researchers and librarians to share best practices and lessons learned in exploring collections as data during a public conference and brainstorming session with staff.

From tracking the popularity of bible verses in historic newspapers to mapping everywhere you've ever walked on your smartphone, speakers discussed the methods, skills, ethical considerations and community participation necessary to find new meaning in digitized and born digital cultural heritage collections.

My background is in software development. Before this job, I ran the Repository Development group [8] at the Library of Congress and before that I worked on creating digital preservation software solutions for the Government Publishing Office. My perspective is on the very practical. Institutions have spent a lot of time, effort, and money on digitizing collections and establishing policies and infrastructures around the model of access that mimics analog models. Transforming the technology, staff, and practice to accommodate data analysis is a second paradigm shift that will be just as difficult. For many knowledge institutions, funding is decreasing and becoming less secure while the volume and complexity of digital information is multiplying and the commitment to analog collections remains. In my view, the only way forward is together:

- Leverage connections with physical sciences, social sciences, and journalism. Work together on tooling and training.
- Highlight digital scholarship projects with easy to understand outcomes to make the case beyond academia.
- Support distributed fellowship models (NDSR) for building digital stewardship curation skills and building skills for doing digital research.
- Create train-the-trainer programs to help scholars understand what’s possible using

computation

- Get content, methodologies, and tools to K-12 educational audiences.
- Explore legal, cultural and privacy review models to guide researchers using novel digital content, like a light-weight IRB.
- Provide space and time for experimentation.

The Library of Congress “preserves and provides access to a rich, diverse and enduring source of knowledge to inform, inspire and engage you in your intellectual and creative endeavors.” [9] We are thrilled to be a part of this exciting conversation, and look forward to working together.

Works Cited

[1] “The Librarian of Congress and the Greatness of Humility” by Sarah Larson. *The New Yorker*. February 19, 2017 http://www.newyorker.com/culture/sarah-larson/the-librarian-of-congress-and-the-greatness-of-humility

[2] “Data and Humanism Shape Library of Congress Conference” by Mike Ashenfelder. *The Signal*. October 21, 2016 http://blogs.loc.gov/thesignal/2016/10/data-and-humanism-shape-library-of-congress-conference/

[3] “Collections as Data Report Summary” by Jaime Mears. *The Signal*. February 15, 2017 http://blogs.loc.gov/thesignal/2017/02/read-collections-as-data-report-summary/

[4] “Co-Hosting a Datathon at the Library of Congress” by Jaime Mears. *The Signal.* July 21, 2015 http://blogs.loc.gov/thesignal/2016/07/co-hosting-a-datathon-at-the-library-of-congress/?loclr=blogsig

[5] “Library of Congress Lab: Library of Congress Digital Scholars Lab Pilot Project Report” by Michelle Gallinger and Daniel Chudnov. December 21, 2016 http://digitalpreservation.gov/meetings/dcs16/DChudnov-MGallinger_LCLabReport.pdf

[6] Software Carpentry at the Library of Congress https://oulib-swc.github.io/2017-02-15-loc/

[7] data-exploration Github page https://github.com/LibraryOfCongress/data-exploration

[8] “Yes, the Library of Congress Develops Lots of Software Tools” by Leslie Johnston. August 16, 2011 https://blogs.loc.gov/thesignal/2011/08/yes-the-library-of-congress-develops-lots-of-software-tools/

[9] “About the Library” https://www.loc.gov/about/