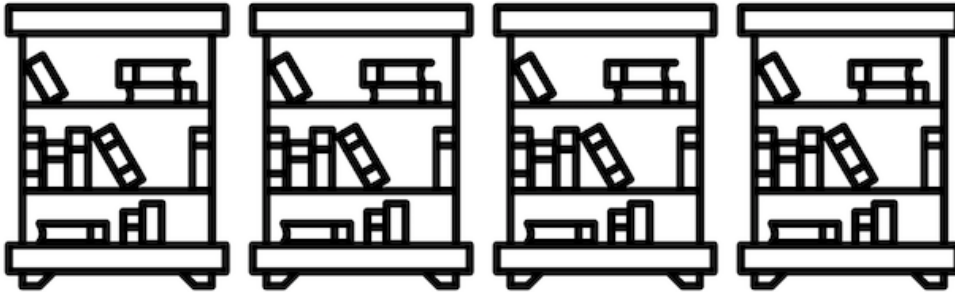


Always Already Computational: Collections as Data

Final Report



Thomas Padilla (PI)
Laurie Allen (Co-PI)
Hannah Frost (Co-I)
Sarah Potvin (Co-I)
Elizabeth Russey Roke (Co-I)
Stewart Varner (Co-I)



This project was made possible by the Institute of Museum and Library Services (LG-73-16-0096-16).
The views, findings, conclusions, or recommendations expressed in this publication do not necessarily represent those of the Institute of Museum and Library Services or author host institutions.

Acknowledgements

The project team would like to acknowledge the Institute of Museum and Library Services, whose support made this project possible. Program Officers Trevor Owens and Emily Reynolds provided essential guidance throughout. Patricia Hswe, formerly at Penn State University Libraries, now at the Andrew W. Mellon Foundation, helped spark the idea that became reality. We thank Stanford University, Texas A&M University, Emory University, and the University of Pennsylvania for their contributions to the project. Project home institutions - the University of California, Santa Barbara, and later the University of Nevada, Las Vegas - provided crucial support to the project. Special thanks to Amy Gros Louis, Kee Choi, Lonnie Marshall, Maggie Farrell, and Michelle Light.

Individuals listed below authored and edited project resources, participated in national forums, and presented or served on program committees for project-initiated events. Many others beyond the list below contributed - we are grateful to them all.

Ajao, John	Foreman, Gabrielle P.	Liu, Alan
Almas, Bridget	Fowler, Daniel	Locke, Brandon
Anderson, Clifford	Galarza, Alex	Lynch, Katherine
Arroyo Ramirez, Elvia	Gniady, Tassie	Mannheimer, Sara
Averkamp, Shawn	Gradeck, Bob	Marciano, Richard
Bailey, Helen	Green, Harriett	Marcus, Cecily
Bailey, Jefferson	Guiliano, Jennifer	Martinez, Alberto
Baumgardt, Frederik	Hardesty, Julie	Mason, Ingrid
Becker, Devin	Harlow, Christina	Matienzo, María
Butterhof, Robin	Higgins, Devin	McLaughlin, Steve
Capell, Laura	Horowitz, Sarah M.	Meredith-Lobay, Megan
Chassanoff, Alex	Hswe, Patricia	Miller, Matthew
Claeyssens, Steven	Ikeshoji-Orlati, Veronica	Milligan, Ian
Clement, Tanya	Jansen, Greg	Mookerjee, Labanya
Coates, Heather	Johnston, Lisa	Morgan, Paige
Coble, Zach	Jordan, Mark	Neatrou, Anna
Collard, Scott	Jules, Bergis	Newbury, David
Craig, Kalani	Kashyap, Nabil	Nunes, Charlotte
Cram, Greg	Kaufman, Micki	Orlowitz, Jake
Del Rio Riande, Gimena	Kerchner, Dan	Patterson, Sarah
Di Cresce, Rachel	Kizhner, Inna	Phillips, Cheryl
Dickson, Eleanor	Kouper, Inna	Pollock, Caitlin
Dombrowski, Quinn	Leem, Deborah	Porter, Dot
Elings, Mary	Lill, Jonathan	Posner, Miriam
Enderle, Scott	Lillehaugen, Brook	Powell, Chaitra
Escobar Varela, Miguel	Lincoln, Matthew	Rabun, Sheila
Ferriter, Meghan	Littman, Justin	Ridge, Mia

Rodgers, Richard
Romary, Laurent
Ross, Denice
Roued-Cunliffe, Henriette
Sakr, Laila
Scates Kattler, Hannah
Schmidt, Ben
Schwartz, Daniel L.
Scott Weber, Chela
Senseney, Megan
Seubert, David
Severson, Sarah

Sherratt, Tim
Simpson, John
Souther, Mary
Sutton Koeser, Rebecca
St. Onge, Timothy
Terras, Melissa
Thomas, Deborah
Thompson, Santi
Tomasek, Kathryn
Tracy, Daniel G.
Van Tine, Lindsay
Vejvoda, Berenica

Vogeler, Georg
Weigel, Tobias
Weingart, Scott
Whitmire, Amanda
Williams, Elliot
Wolf, Nick
Wrubel, Laura
Yarasavage, Nathan
Zarafonetis, Michael
Zastrow, Thomas
Ziegler, Scott
Zwaard, Kate

Scope Note	8
Introduction	9
Activities	11
About our approach	11
Collections as Data Framework (v1)	12
The Santa Barbara Statement on Collections as Data	12
Collections as Data Facets	12
Collections as Data Personas	12
50 Things	12
Methods Profiles	13
Collections as Data Position Statements	13
Additional Resources	13
Impacts	14
Findings	16
Collections as data development requires critical engagement with the ethical implications of cultural heritage organization work	16
Collections as data development is possible at a wide range of organizations	16
Collections as data development benefits collection users and stewards	17
Challenges to collections as data development are more organizational than technical	17
Collections as data development benefits from engaging specific community needs	18
Collections as data development benefits from collaboration across multiple communities of practice	18
Areas for Further Investigation	19
Moving from ethical consideration to action	19
Conducting more community-specific user studies to inform workflow development	19
Developing functional requirements in service to user and collection steward needs	19
Publicly charting and sharing the terms of relationships with commercial entities	19
Enabling widespread collections as data discovery	20
Addressing collections as data preservation needs and obstacles	20
Exploring post-custodial approaches to collections as data	20
Appendices	21
Appendix 1: The Santa Barbara Statement on Collections as Data	21
Appendix 2: Collections as Data Facets	24
Appendix 3: Collections as Data Personas	80
Appendix 4: 50 Things	93
Appendix 5: Collections as Data Methods Profiles	98
Appendix 6: National Forum Position Statements	103

Appendix 7: Forum Summaries	161
Appendix 8: Conference engagements, 2017-2018	168
Appendix 9: Digital Humanities 2017 preconference: Shaping Humanities Data	171

Scope Note

From 2016-2018 *Always Already Computational: Collections as Data* documented, iterated on, and shared current and potential approaches to developing cultural heritage collections that support computationally-driven research and teaching. With funding from the Institute of Museum and Library Services, *Always Already Computational* held two national forums, organized multiple workshops, shared project outcomes in disciplinary and professional conferences, and generated nearly a dozen deliverables meant to guide institutions as they consider development of collections as data.

This report documents the activities and impacts of the Always Already Computational project, delineates findings, and identifies areas for further inquiry.

Introduction

Always Already Computational: Collections as Data arose from practical need and a desire to build upon decades of digital collection practice. While cultural heritage practitioners have broad experience replicating the analog experience of watching, viewing, and reading in a digital environment, they less commonly share the experience of supporting users who want to work with collections as data - a conceptual orientation to collections that renders them as ordered information, stored digitally, that are inherently amenable to computation. These users come from many disciplines and professions, they act within and outside of the university, and they share in common a desire to leverage computational methods like machine learning, computer vision, text mining, visualization, and network analysis. Meeting their needs is contingent on the availability of collections, infrastructure, and services that are tuned for computational work.

At the time *Always Already Computational* formed, existing experience in this space was difficult to discern beyond relatively well-resourced efforts like the HathiTrust Research Center and the British Library. Without diversification of examples and corresponding paths to doing the work, the viability of collections as data efforts ran the risk of being perceived as an elite activity - smaller actors need not apply. It became clear that a broader field of participation was needed. Ideally, this field would exhibit variation in institutional resources, collection types, and community responsibilities. All of the above would critically contend with the ethical implications of producing and making use of collections as data. From 2016-2018, *Always Already Computational* sought to cultivate this field by openly documenting, iterating on, and sharing current and potential approaches to developing cultural heritage collections that support computationally-driven research and teaching.

At inception, anticipated project outcomes were as follows: gather key stakeholders to craft a strategic direction that leads to **(1)** creation of a collections as data framework that supports pragmatic collection transformation and documentation, **(2)** development of computationally amenable collection use cases and user stories **(3)** identification of methods for making computationally amenable library collections more discoverable through aggregation and other means, and **(4)** guidance, in the form of functional requirements that support development decisions relative to technical feature integrations with repository infrastructures.

As synchronous and asynchronous engagements began in earnest, project scope and the shape of deliverables morphed accordingly. The tension between creation of particular solutions and universal solutions was persistent. Given its nature as a broadly conceived community project, *Always Already Computational* was not positioned to make overly specific technical recommendations. Preference was ultimately given to the creation of malleable deliverables that could be shaped to guide engagement with particular community needs. We determined that collections as data discoverability and the development of specific functional requirements were projects that required independent investigation. Ideally these investigations will be tied to specific contexts, a framing distinct from a project like *Always Already Computational*, which sought cultural heritage community-wide engagement.

Always Already Computational deliverables constitute version 1 of the collections as data framework. This framework includes a range of resources, expressed in different forms, providing multiple points of engagement throughout the process of considering collections as data efforts. For example, *The Santa Barbara Statement on Collections as Data* is a set of principles developed with community feedback designed to help guide practitioners through the practical, theoretical, and ethical dimensions of collections as data work. This deliverable does not advance solutions, rather it raises core questions to be resolved in local contexts. The *Collections as Data Facets* describe a range of institutional approaches to implementing collections as data. This resource aims to help practitioners see multiple paths into doing the work. The *Collections as Data Personas* represent high level role types associated with collections as data development and use. Together, the personas, derived from *Always Already Computational* community engagements and project team experience, aim to surface needs, motivations, and goals in context. Compiled at the end of two years of project engagements, the *50 Things* provide examples of things a practitioner can do to initiate collections as data at their institution.

Throughout the course of the project, *Always Already Computational* was inspired and humbled by the active interest and ingenuity shown by librarians, archivists, museum professionals, researchers, educators, and more as they engaged with collections as data challenges and opportunities. By emphasizing diverse community engagement and documentation over prescriptive recommendations, we hope that we have cultivated, encouraged, and questioned in ways that a wide range of communities find to be useful.

Thomas Padilla

Laurie Allen

Hannah Frost

Sarah Potvin

Elizabeth Russey Roke

Stewart Varner

Activities

About our approach

From the beginning, *Always Already Computational* held an expansive view of collections as data work. The project sought to document implications of collections as data work across cultural heritage organization functions, practices, and roles. A National Forum with participants representing a broad spectrum of perspectives kicked off project activity. Two years of synchronous and asynchronous community engagements spanning a range of professional and disciplinary contexts followed.

Project activity was designed to serve three near-term goals: **(1)** identify cross-cutting issues and bring common themes into focus, **(2)** scaffold project activity with those issues and themes **(3)** identify special concerns or less clear areas that required deeper investigation. Discussions at the first National Forum informed overall project goals and direction. Project deliverables were iterated on over the course of the project activity. Iteration was by design, given the need to engage, respond to, and incorporate diverse community input. Deliverables were shared across a range of venues including but not limited to the Digital Library Federation, American Historical Association, Society of American Archivists, the Coalition for Networked Information, Association of College and Research Libraries, NICAR, and Open Repositories.

Always Already Computational community engagements drew inspiration from human-centered design methods. The LUMA Institute *Handbook of Human-Centered Design Methods* and the *Liberating Structures* toolkit provided a series of generative activities¹:

- **Round Robin** - generate fresh ideas by providing a format for group authorship.²
- **Concept Poster** - promote an idea and rally support for its development.³
- **Affinity Clustering** - teams sort items based on perceived similarity, defining commonalities that are inherent but not necessarily obvious.⁴
- **Importance/Difficulty Matrix** - establish priorities by plotting relative importance and difficulty.⁵
- **1-2-4-All** - generate ideas that open with self-reflection in response to a prompt and expand into larger group discussion.⁶

¹ LUMA Institute, *Innovating for People: Handbook of Human-Centered Design Methods* (Pittsburgh, PA: LUMA Institute, 2012); <http://www.liberatingstructures.com/>

² *Innovating for People*, 64.

³ *Ibid.*, 76.

⁴ *Ibid.*, 40.

⁵ *Ibid.*, 44.

⁶ <http://www.liberatingstructures.com/1-1-2-4-all/>

Individual and group perspectives gathered through these activities directly informed the framework described below.

Collections as Data Framework (v1)

The Santa Barbara Statement on Collections as Data

The Santa Barbara Statement on Collections as Data is a set of principles developed with community feedback designed to help guide practitioners through the practical, theoretical, and ethical dimensions of collections as data work. This deliverable does not advance solutions, rather it raises core questions to be resolved in local contexts. The first version of the Santa Barbara Statement was inspired by the first collections as data national forum (UC Santa Barbara, March 1-3 2017). After its release, the team asynchronously gathered comments on the web via open annotation and sought synchronous feedback across a series of conversations and workshops. The second version of the statement was revised and released based on community feedback.

Permanent link: <https://doi.org/10.5281/zenodo.3066209>

Collections as Data Facets

Collections as Data Facets, authored by community contributors, document collections as data implementations. An implementation consists of the people, services, practices, technologies, and infrastructure that aim to encourage computational use of cultural heritage collections. The fifteen facets represent collections as data efforts at museums, academic libraries, societies, and institutions like the Library of Congress.

Permanent link: <https://doi.org/10.5281/zenodo.3066240>

Collections as Data Personas

Collections as Data Personas represent high level role types associated with the development *and* use of collections as data. The personas aim to surface needs, motivations, and goals in context.

Permanent link: <https://doi.org/10.5281/zenodo.3066515>

50 Things

50 Things is designed for practitioners who are seeking to get started with collections as data. 50 Things provides an impetus for exploring, learning from colleagues, deepening knowledge and understanding, and taking that first step. Participants at our second National Forum (University of Nevada Las Vegas, May 7-8, 2018) provided the bulk of recommendations.

Permanent link: <https://doi.org/10.5281/zenodo.3066237>

Methods Profiles

Methods Profiles characterize common research methods in relation to the process of collections as data development. They are designed to help collection stewards bridge the gap between research methods and design of workflows that support creation of machine actionable collections.

Permanent link: <https://doi.org/10.5281/zenodo.3146756>

Collections as Data Position Statements (Forum 1)

Prepared by invited participants in advance of the first [collections as data national forum](#) (UC Santa Barbara, March 1-3 2017), the twenty-six position statements describe challenges, opportunities, connections, and gaps in the work of collections as data. Perspectives subsequently informed project activity.

Permanent link: <https://doi.org/10.5281/zenodo.3066161>

Additional Resources

- [National Forum 2 Livestream Recording](#)
- [Collections as Data Google Group](#) - As of May 2019, the Google Group includes 57 topics and 413 members⁷
- [Collections as Data Group Library](#) - As of May 2019, this Zotero group includes 266 items and 73 members⁸
- [Serendipitous Collections as Data](#)

⁷ <https://groups.google.com/forum/#!forum/collectionsasdata>

⁸ https://www.zotero.org/groups/2171423/collections_as_data_-_projects_initiatives_readings_tools_datasets

Impacts

Always Already Computational's primary role, as expressed in the framework, was to highlight existing work, foster conversations, identify gaps, collect feedback, and spark further conversation and adoption in the context of specific community needs. The impact of *Always Already Computational* is likely best measured by its potential to motivate further development.

Over two years of project activity, *Always Already Computational* saw collections as data:

- ... taken up as a strategic priority within the University of California's Shared Content Leadership Group's *Plans & Priorities for 2017/2018 Based on the University of California Library Collection: Content for the 21st Century and Beyond*
- ... incorporated as a feature of the OCLC *Research and Learning Agenda for Archives, Special, and Distinctive Collections in Research Libraries*
- ... inform the creation of permanent positions like the Digital Collections as Data Manager position at Johns Hopkins University Libraries
- ... inform the creation of postdoctoral positions like the British National Archives' FTNA Postdoctoral Research Fellowship, focused on unlocking "archival collections as data"
- ... identified as a core driver for an international, future of archival science curriculum effort
- ... presented as a component of the Digital Library Federation eResearch Network
- ... inform Software Preservation Network outreach
- ... delivered as a week-long collections as data course at the Humanities Intensive Learning and Teaching Institute
- ... inspire reading groups, international hackathons, workshops, and conference sessions that span disciplinary, library, archives, and museum communities.⁹

⁹ "2017/2018 SCLG Plans & Priorities for 2017/2018 Based on the University of California Library Collection: Content for the 21st Century and Beyond," University of California, last modified September 29, 2017, http://libraries.universityofcalifornia.edu/groups/files/sclg/docs/SCLG_2017_2018%20Plan.pdf; Chela Scott Weber. "Research and Learning Agenda for Archives, Special, and Distinctive Collections and Research Libraries." OCLC Research, 2017. <https://doi.org/10.25333/C3C34F>; "Manager of Digital Collections as Data." <https://jobs.jhu.edu/job/Baltimore-Manager-of-Digital-Collections-as-Data-MD-21218/546941200/>; "Developing a Computational Framework for Library and Archival Education." Developing a Computational Framework for Library and Archival Education. <https://dcicblog.umd.edu/ComputationalFrameworkForArchivalEducation/>; "FTNA Postdoctoral Research Fellowship (Datafication) at The National Archives," February 9, 2018. <https://web.archive.org/web/20180209203649/http://www.jobs.ac.uk/job/BHO511/ftna-postdoctoral-research-fellowship-data-fication/>; "EResearch Network - DLF Wiki." Accessed January 21, 2019. https://wiki.diglib.org/EResearch_Network#webinars; "Events | The Software Preservation Network." Accessed May 14, 2018. <http://www.softwarepreservationnetwork.org/events/>; Padilla, Thomas, and Mia Ridge. "Collections as Data." *HILT* (blog). Accessed January 21, 2019. <https://dhtraining.org/hilt/course/collections-as-data-2018/>; September 12, Natalia Ermolaeva. "CDH Reading Group: Collections as Data." Center for Digital Humanities @ Princeton University, September 12, 2018. <https://cdh.princeton.edu/updates/2018/09/12/cdh-reading-group-collections-data/>; Moore Institute. "Collections as Data - Hackathon / Collaborative Workshop - Moore Institute." Text. *NUI Galway* (blog). Accessed January 21, 2019. <http://mooreinstitute.ie/event/collections-data-hackathon-collaborative-workshop/>; Dalmau, Michelle. "Collections as Data at Indiana University and Beyond," November 16, 2018. <https://libraries.indiana.edu/collections-data-indiana-university-and-beyond>; Menendez, Rebecca, Cheryl Miller, Andrzej Rutkowski, and Stacy R. Williams. "ARLIS/NA 47th Annual Conference: Getting Started with Collections as Data." Accessed January 21, 2019

- ... directly inform *Collections as Data: Part to Whole*, awarded \$750,000 by the Andrew W. Mellon Foundation to foster the development of broadly viable models that support implementation *and* use of collections as data.

In addition to tracking the various examples of impact above, the *Always Already Computational* team simply asked through an open survey, “Have you used this project?”. We include below a sampling of responses:

More than the resources, which I've referenced and read and looked at off an on during the projects run, we (the digital library folk at Idaho) have used the idea(s) promoted through the project to stimulate our own thinking, development, and conversations. I've had other librarians I don't work with that closely with bring up the project to me, and that's led to some really interesting conversations.

Devin Becker, University of Idaho

(1) I'm leading data curation work package in a national research and data infrastructure project for humanities, arts and social sciences. I drew on the Collections As Data facets to augment the advice given to a colleague new to data curation, to help them think about how to make data available e.g. via API or as static snapshots in a sustainable manner, and to think about their collection "as data". (2) The facets have also informed the development of a data curation framework for data sharing and interoperability across multiple platforms (discovery, access, research and archiving).

Ingrid Mason, Australia's Academic and Research Network (AARNet)

So far, we have developed one research project exploring the use of oral histories as collections as data. Collections as Data has also strongly influenced our thinking of how best to digitize and make available a collection of mining records from the early 1900s, which would be best expressed more as a database set up for computational use by researchers in addition to a traditional digital collection.

Anna Neatrou, University of Utah

Use of the [collections as data] facets were instrumental in explaining the widespread practice of working with collections as data. Before this list of examples, it was a constant struggle to explain the idea and justify the work. I frequently cite the Santa Barbara Statement when writing about the the use of data in special collection libraries. I've used the Personas somewhat less

<https://arlisna2019.sched.com/event/ITtA/getting-started-with-collections-as-data?iframe=no&w=100%&sidebar=yes&bg=no>; Neely, Liz, Anne Luther, and Chad Weinard. “Cultural Collections as Data: Aiming for Digital Data Literacy and Tool Development – MW19 | Boston.” Accessed January 21, 2019.

<https://mw19.mwconf.org/proposal/cultural-collections-as-data-aiming-for-digital-data-literacy-and-tool-development/>; Padilla, Thomas, Hannah Scates Kettler, Laurie Allen, and Stewart Varner. “Collections as Data: Part to Whole.” Collections as Data - Part to Whole. Accessed January 21, 2019. <https://collectionsasdata.github.io/part2whole/>.

regularly, but I have referenced them to offer examples for what types of researchers might be interested in different types of data.

Scott Ziegler, Louisiana State University

I think it helps people draw the connection between digital archives and progressive values. I also think it's a helpful, positive avenue into discussing what resources are necessary in terms of storage, repository infrastructure, etc. in order to archive collections digitally, and why institutions should earmark funds and other resources to support collections as data.

Charlotte Nunes, Lafayette College

Findings

3

Appendices

Appendix 1: The Santa Barbara Statement on Collections as Data

May 2018

The Santa Barbara Statement on Collections as Data was written by the Institute of Museum and Library Services supported Always Already Computational: Collections as Data project team. The first version was based on the collaborative work of participants at the first Collections as Data National Forum (UC Santa Barbara, March 1-3 2017). After its release, the team gathered comments from the Hypothesis web annotation tool and sought additional feedback across a series of conversations and workshops (April 2017 - April 2018). The current version of the statement was revised based on that community feedback, especially the close, directed feedback provided by workshop participants at the Digital Library Federation Forum 2017.

What are “collections as data”? Who are they for? Why are they needed? What values guide their development? The Santa Barbara Statement on Collections as Data poses these questions and suggests a set of principles for thinking through them, as part of a community effort to empower cultural heritage institutions to think of collections as data and consequently to explore what might be possible if cultural heritage seen in this light was more readily open to computation.

The concept of collections as data emerges at – and is grounded by – a particular moment in the recent history of cultural heritage institutions. For decades, cultural heritage institutions have been building digital collections. Simultaneously, researchers have drawn upon computational means to ask questions and look for patterns. This work goes under a wide variety of names including but not limited to text mining, data visualization, mapping, image analysis, audio analysis, and network analysis. With notable exceptions like the Hathitrust Research Center, the National Library of the Netherlands Data Services & APIs, the Library of Congress’ *Chronicling America*, and the British Library, cultural heritage institutions have rarely built digital collections or designed access with the aim to support computational use. Thinking about collections as data signals an intention to change that, and efforts like the Library of Congress’ *Collections as Data: Stewardship and Use Models to Enhance Access* and the multinational *Digging into Data* suggest that a broader community shift intentionally scoped to institutions large and small comes at an opportune time.

While the specifics of how to develop and provide access to collections as data will vary, any digital material can potentially be made available as data that are amenable to computational use. Use and reuse is encouraged by openly licensed data in non-proprietary formats made accessible via a range of access mechanisms that are designed to meet specific community needs.

Ethical concerns are integral to collections as data. Collections as data should make a commitment to openness. At the same time, care must be taken to comply with legal requirements, cultural norms, and the values of vulnerable groups. The scale of some collections may also obfuscate what is hidden or missing in the histories they are perceived to represent. Cultural heritage institutions must be mindful of these absences and plan to work against their repetition. Documentation should be informed by archival principles and emergent reproducibility practice to ensure that users have the information they need to work with collections responsibly.

Principles

1. **Collections as data development aims to encourage computational use of digitized and born digital collections.** By conceiving of, packaging, and making collections available as data, cultural heritage institutions work to expand the set of possible opportunities for engaging with collections.
2. **Collections as data stewards are guided by ongoing ethical commitments.** These commitments work against historic and contemporary inequities represented in collection scope, description, access, and use. Commitments should be formally documented and made publicly available. Commitment details will vary across communities served by collections but will share common cause in seeking to address the needs of the vulnerable. Collection stewards aim to respect the rights and needs of the communities who create content that constitute collections, those who are represented in collections, as well as the communities that use them.

3. **Collections as data stewards aim to lower barriers to use.** A range of accessible instructional materials and documentation should be developed to support collections as data use. These materials should be scoped to varying levels of technical expertise. Materials should also be scoped to a range of disciplinary, professional, creative, artistic, and educational contexts. Furthermore the community should be motivated and encouraged to build and share tools and infrastructure to facilitate use of collections as data.
4. **Collections as data designed for everyone serve no one.** Specific needs inform collections as data development. These needs may be commonly held by particular user communities. Rather than assuming these needs or imagining these communities, stewards should be intentional about who their collections are designed for, work to lower the barriers to use for the people in those communities, and continue to assess these needs over time. Where resources permit, multiple approaches to data development and access are encouraged.
5. **Shared documentation helps others find a path to doing the work.** For example, collections as data work can entail decisions about selection, description, conversion cleaning, formatting, and delivery mechanisms or platforms that enable discovery and provide access. In order for a range of individuals and institutions to engage collections as data work, it must be possible to locate documentation that demonstrates how and why the work is done. Documentation must also attest to the history of how the collection has been treated over time. While no documentation can be fully comprehensive, incomplete or in-progress documentation is better than no documentation. Examples of documentation include human and machine readable metadata schemas, data sheets, workflows, application profiles, deeds of gift, and codebooks. Documentation should be publicly accessible by default.
6. **Collections as Data should be made openly accessible by default, except in cases where ethical or legal obligations preclude it.** Terms of use for collections as data must be made explicit and should align with community-based practices such as RightsStatements.org and standard licenses such as Creative Commons, Open Data Commons, and Traditional Knowledge licenses.
7. **Collections as data development values interoperability.** Interoperability entails alignment with emerging and/or established community standards and infrastructure and eases integration with centralized as well as distributed infrastructure. This approach facilitates collections as data discovery, access, use and preservation.
8. **Collections as data stewards work transparently in order to develop trustworthy, long-lived collections.** Trustworthiness depends upon efforts to ensure and publicly document the technical integrity of the data as well as its provenance. It also requires that data stewards acknowledge absences and areas of uncertainty within the collection as data. Trustworthy collections as data should include open, robust metadata, and should be under

the care of stewards and institutions committed to their preservation.

9. **Data as well as the data that describe those data are considered in scope.** For example, images and the metadata, finding aids, and/or catalogs that describe them are equally in scope. Data resulting from the analysis of those data are also in scope.
10. **The development of collections as data is an ongoing process and does not necessarily conclude with a final version.** Work in progress status can be seen as a virtue when iteration is geared toward developing productive collaborations and integrations between new and existing technologies, workflows, and service models. The ongoing development of collections as data can impact staffing models, workflows, and technical infrastructure.

Appendix 2: Collections as Data Facets

August 2017 - August 2018

Collections as Data Facets document collections as data implementations. An implementation consists of the people, services, practices, technologies, and infrastructure that aim to encourage computational use of cultural heritage collections.

Facet 1: MIT Libraries Text and Data Mining

Richard Rodgers, Massachusetts Institute of Technology

1. Why do it

MIT Libraries collect, curate, and provide access to numerous digital collections that comprise important research outputs and contributions to the scholarly record. Access is typically provided via traditional web applications designed for individual users in browsers. In assessing the patterns of use of these collections, it became apparent that a significant amount of traffic was due to various automated processes that ‘scraped’ the sites, but did not identify themselves

as indexing services. At the same time, we began to receive more and more direct requests from individual scholars on campus (and beyond) for bulk delivery of textual corpora in our collections, in order to perform text-mining on them. It was clear that these 'alternative' uses of collections were not well served by existing access methods and systems.

2. Making the Case

We saw that we needed to explore how better to provide access for these kinds of use, and this need dovetailed with a broader agenda that the Libraries were pursuing of reconceiving library services as a 'platform': a notion articulated in recommendation 6 of the Future of the Libraries Task Force Report, which specifically mentions text and data mining as important 'non-consumptive' uses of library-stewarded material. The platform model emphasizes empowering users to create their own discovery/access/consumption tools by providing open, standards-based, and performant APIs or other services that such tooling can leverage. So the case was made by arguing that an experiment to expose collection data via a new API designed for bulk access would teach us how to build a library platform that would increase the value of all collections.

3. How you did it

Based on the analytics, we selected MIT's Electronic Theses and Dissertations as the initial collection to work with: it was highly sought after, fairly extensive (close to 50K theses, with plans to digitize the entire historical run), and already under management in our institutional repository (DSpace@MIT). We wrote a formal proposal for a project to design and build a prototype of a new discovery and access service for this collection to enable text and data mining (or other non-consumptive uses).

The project team consisted of:

- a project manager, who oversaw the scrum-agile process used to manage the development
- three software developers, who took responsibility for content accession, repository management, and API design and development, respectively
- an analyst, who surveyed the field of existing text and data mining services, and who worked with potential users of the system to understand their needs
- a UI/UX expert, who helped in designing intuitive and effective user interfaces (which complemented and documented the API).
- The development project ran for 10-11 months, and a functional prototype was built that exposed an API for discovery and bulk access of theses. The user could request any (or all) of 3 content representations: the metadata (including an abstract), the thesis as a PDF (which is the approved submission format), and the full (unstructured) text extracted from the PDF.

The service consisted of several cooperating software components: a Fedora 4 repository, which held the metadata and textual artifacts, an Elasticsearch index, used to query the full-text, as

well as the metadata, an API server which formed the front-end, exposing the ways users could interact with the index and repository, and various queues and caches to connect these components. Each component was deployed in a container to a Kubernetes-orchestrated environment in a cloud service (Google Container Engine).

Several challenges the project encountered, to name a few:

- The quality of the PDFs in the collection varied considerably, with numerous encoding and other errors that affected or impaired use. Some theses were created in digitization workflows from analogue originals, whereas others were ‘born digital’, and both content streams were created over a long span of time using different software, workflow practices, etc. We vacillated between attempts to ‘repair’ the theses, or enhance the metadata with quality indications so that machine use could adjust for it: the final prototype included aspects of both approaches.
- The cloud environment required considerable knowledge of deployment and orchestration tools and platforms that the team lacked. While we were able largely to surmount these deficiencies, we did so at some cost to the overall project deliverables. Our initial resource model for the project included a ‘devops’ role (unfilled) that would have greatly assisted.
- It was difficult to identify and attract a broad variety of potential users to help define the product design. We gained valuable insight from those we engaged with, but suspected there were many more research objectives, techniques, requirements, etc that would have beneficially shaped the design of the API and the whole service. This stemmed in part from the fact that we were asking for input without a working system to react to.

4. Share the docs

Project documents forthcoming, but the code that was used to run the prototype is available on Github.

5. Understanding use

The team solicited potential users of the thesis service, and conducted a small number of interviews to elicit both their intended use, but also what affordances such a service should provide to researchers.

We learned that the metadata we exposed (academic department, completion year, degree type, etc) were considered useful ways to plumb and select within that particular corpus (theses), in addition to keyword search over the full-text.

The service itself was designed to gather data about how it was used, but working against this was the desire to make the data openly available to all, without ‘user tracking’. In the end, the service emerged with a lightly tiered structure: all content was freely available, but certain advanced functions required obtaining an API key (which allowed much better analytics).

6. Who supports use

While the cloud-hosted service compute infrastructure was supported by the libraries technology team, the project required considerable support throughout the libraries and archives. At MIT, the responsibility for collecting and curating theses and dissertations falls to the Institute Archives, who were a key stakeholder in the project. They did extensive research (including soliciting advice from the Institute's legal counsel) on the IP and rights issues surrounding such a new service, since this kind of use was not originally contemplated in the policies governing theses. They also assumed general responsibility for the rare but complex decisions around takedown requests, etc.

Since this service obtains content from existing digitization workflows, the digitization team was also closely involved in providing access to scripts, software tools, etc used to create thesis artifacts.

If the service were launched in production, repository managers would need to both administer the service, but also field questions and provide support for end-users (API key management, etc). In addition, the IT operations group would need to follow the standard set of practices for system backup, performance monitoring, etc. We learned that data-intensive services such as this (where gigabytes of package downloads were routine) had to be managed carefully from a resource perspective.

7. Things people should know

One key insight we gained was the need to perform a thorough appraisal of the collection from a data completeness, uniformity, and consistency perspective: when discovery and access are confined to siloed legacy applications, these quality dimensions may be difficult to observe.

8. What's next

ETDs were a great candidate collection for understanding the requirements of a text and data mining service, but we have numerous text-based collections of high value, including our extensive open access articles collection, conference proceedings, technical reports, working papers, etc. An analysis of these corpora (what are useful metadata discriminators, etc) in light of the insights gained in the theses prototype, could lead to a general, flexible service for offering the wealth of content the Libraries has to new forms of scholarly inquiry.

Facet 2: Carnegie Museum of Art Collection Data

David Newbury, Carnegie Museum of Art; Daniel Fowler, Open Knowledge International

1. Why do it

As stated on the Carnegie Museum of Art (CMOA) website, the Collection Data project is meant to be used for “discovery, inspiration, and innovation, allowing people to creatively re-imagine and re-engineer our collection in the digital space.” CMOA Collection Data is stored in [EMu](#), a collections management system from Axiell. This Collections as Data Facet documents the release of this data: It was exported to both CSV and JSON as a “data dump” and [released on GitHub](#) for public consumption to help enable this creative reuse.

CMOA acknowledges that this project is continuously evolving and that the data will be periodically revised to reflect changes in how its curators understand the objects stored in the database. This acknowledgment is reflected in the choice of a platform (GitHub) which natively

supports storing version-controlled data. CMOA made the choice to publish using CSV, JSON, and GitHub because of their relative ease of use for researchers and developers—these platforms enable easy access to large amounts of data without the need for tools beyond what the researchers already possess, or requiring potential users to learn an API or write SQL against proprietary databases.

In addition to publishing the data itself, it was also important to provide a human- and machine-readable description of the data, its structure, and guidance on how to actually use it. CSV, while easy to work with for many users, is a notoriously underspecified format: developers often have differing opinions on what constitutes a “valid” CSV file. The [Data Package specification](#) developed by Open Knowledge International is a “containerization” format for data which is meant to provide a consistent interface (or “wrapper”) to a diverse range of datasets, especially those containing tabular data (e.g. data stored in CSV files). A single file, `datapackage.json`, stored with the dataset documents where each data file is saved (either on disk or a remote server) as well as its “schema” (number of columns and expected values per column). Releasing this dataset as a Data Package was a good start for providing a minimum machine-readable description of a dataset for processing. A growing set of software libraries and tools can read the Data Package specification so that artists, data analysts, and other users interested in CMOA’s collection can benefit from this consistent interface regardless of the software they use.

A human-readable version of some of this same information is provided through a supplied “README” file.

Collection Data on GitHub: <https://github.com/cmoa/collection>

Data Package specification: <http://specs.frictionlessdata.io/>

2. Making the Case

The case to provide the public increased access to museum data was not a difficult one at the Carnegie Museum of Art—the museum considers engagement and education to be a core part of its mission, and firmly believes in Open Access as essential to museum practice. Also, we were helped immensely by the fact that several large institutions, in particular MoMA, [had already done so](#)—rather than having to explain exactly what we were doing in detail, we could tell our administration and board that “we were doing it the way MoMA did it”. Being able to model our work on the previous work and decisions of others helped reassure non-technical stakeholders that we weren’t doing anything risky or controversial.

The most significant barrier was determining how to coordinate the various expectations across departments—to publish this data required coordination across registrarial, publishing, digital, and curatorial teams. Additionally, it was clear that it would be important to provide all stakeholders with the ability to maintain control over their data. We provided at least six months

of notice to allow the various departments time to correct any information that they felt was essential, and we also allowed anyone to hold back data that they didn't feel was ready. All we asked for was a single sentence written description of why the information should not be published. This allowed stakeholders to maintain agency, while avoiding the temptation to withhold large amounts of the information by default.

Finally, we had many internal discussions about how regular updates would be possible, and we worked with all the departments to craft language to communicate this within the GitHub documentation as being living data. This helped set the expectation both inside and out that this is not a publication that had been vetted by a curator for accuracy and completeness.

3. How you did it

The Carnegie Museum of Art collections data publication was an offshoot of the Art Tracks project at CMOA, a data visualization for provenance. Because of the sensitive nature of provenance, one of the most important goals of the project was to ensure that the professionals with the best understanding of the nuances of the data had control over which works were available for publication. To do so, we worked with Travis Snyder, the Collections Database Administrator, to craft a series of reports, using filter criteria he devised and fields he approved, that created a collection of XML reports, one per-table, from the collections management system. These reports run as needed nightly, and the resulting XML is uploaded to an internal FTP site.

A second set of custom scripts, written by David Newbury, the Lead Developer of the Art Tracks project, download and transform the XML, replacing internal field names with friendlier labels and joining data across tables. Additionally, these scripts add additional information that is not explicitly held in our collections management system such as the URLs for the object website and images of the work. These scripts, written in Ruby, are run whenever the institution wants to update the publication data.

Our intention was to automate this process, but at this point, the benefit of regular, automatic updates is not yet worth the overhead of what is needed to maintain a complex automation system, for example, the time and effort required to provision servers and handle error reporting robustly. Instead, they've been wrapped into a single command line command using Rake, a Ruby library designed to automate repetitive tasks for programmers. The single command will download the XML, reprocess the files, generate both the JSON and CSV representations, and then upload the generated files to GitHub. Currently, if there are problems in the export, a human is running them and will notice (and hopefully correct) the problem before erroneous data is published. One interesting fact is that this script also has to update the documentation on GitHub. For example, we provide in the documentation the number of items in the collection.

We've included several data formats within our the export. First, we include a CSV export. In discussions with members of the Pittsburgh digital humanities community, CSVs were seen as

the most readily-accessible format for researchers interested in quantitative analysis of our collections information. It doesn't require any programming ability to read it, just a copy of Excel, which also means that it's the version we show internal, non-technical people. It is, however, somewhat limited—for instance, artworks can have one or more creators, and tabular formats like CSV are not designed to handle hierarchical relationships. We encode this data using an internal microformat (pipe-separated values), but we've learned from watching users that this is confusing and non-optimal. We're still working to determine if there's a better way to handle this sort of data.

The Data Package descriptor file, `datapackage.json`, which provides metadata for the CSV files in the dataset is written separately as an encapsulation of the expected output of this CSV export pipeline. Information about contributors to the dataset, its licensing, expected values per column per file is stored here.

We also provide a single large JSON export of the data. This is designed primarily for developers, who can load it into memory and process it directly. It's a large file (41 Mb), but not so large that it can't be held in memory using a modern computer. When we've held hackathons or worked with web technologists, this is the form of the data that they've been most comfortable with.

We also provide a directory containing a single JSON file for each object in the collection. This was created to approximate an API—there's a single URL that will return information about each object, as well as an index file containing a list of ids, titles, and a URL to an image. However, our experience has been that this format is too complicated for both developers (who prefer the single JSON file) and non-developers (who prefer the CSV), and is not used.

An additional complication for our data is that we have broken out the 80,000+ photographs of the Teenie Harris collection into their own file. This collection is part of the CMOA collection, but is significantly larger than the rest of the collection combined. We found in exploring other collection data releases, such as the Tate London and their collection of J.M.W. Turner's sketchbooks, that large-scale special collections tended to drown out the rest of the collection in data analysis, and might be best considered separately. We discussed with the museum stakeholders our options, but the decision was made that publishing them as a separate files, using the same format and structures, and both documented the same way in the GitHub, was an acceptable pattern.

4. Share the docs

One of the most important decisions that we made was to treat [the documentation](#) for the release as of equal importance to the data. Tracey Berg-Fulton, the collections database associate and Art Tracks team member, spent a long time crafting the documentation to be thorough and friendly. Friendly was important, because we knew that many of the people who would be looking at this data would be students or members of the public, and we wanted them

to feel welcome to use the data. Big legal disclaimers and restrictions, or dense technological jargon might have prevented them from feeling like they were welcome.

We also included within our documentation a table that indicates not just what the field is, but what it means, what type of data you can expect, and a real-world example of the sort of data that field contains. We wanted to make sure that people were able to find out if our data would meet their needs without having to download it and review it.

Once we had completed our documentation, we sent it through several rounds of internal review—not just editorial review, to confirm that we’d spelled everything correctly, and legal review, to make sure that we’d appropriately used the correct licenses and disclaimers, but also content review, to make sure that our examples were factual, and that our descriptions captured the nuances of the content experts. This helped, but even more it fostered the sense that this was of the museum, not just of the Art Tracks project or the technology department.

Beyond internal review, we’ve tried to consult with developers and researchers to verify that the information that we’ve provided is what is actually needed to understand our release. We also explicitly reached out to others in our field with a history of being critical of museum documentation and data, such as Matthew Lincoln, to critique our documentation and provide feedback on utility, comprehensibility, and completeness. We’ve also monitored other data releases across the museum field, and have worked to integrate good ideas around documentation from our peers. Finally, we model good collaboration by explicitly linking and thanking the institutions that helped us through example and direct advice on this project.

Finally, we’ve been working with Open Knowledge International to explore the use of Data Packages to provide an additional level of documentation for the collections data release. This provides a machine-readable description of the contents of the CSV file, which allows software tools and agents to both understand and validate the structural content of the data. We use it as a validation tool to ensure that all of the data published is structurally correct—for instance, that every URL is a valid URL, or that our ID numbers are in the correct format, or that every work has an accession date. Our hope is that in the future additional software tools will leverage this format, but the most direct benefit to the institution has been as a exhaustive check against our data to verify that the rules that we believe are enforced actually are—and we have been regularly surprised by the exceptions that we’ve found.

Collection Data on GitHub: <https://github.com/cmoea/collection>

5. Understanding use

Compared to an API, providing access to Carnegie Museum of Art Collection Data through a data dump is a lower support cost option in terms of time and money. There is no server we need to run: CMOA are, for the moment, hosting the public data on GitHub’s infrastructure. Providing a data dump also benefits users, both academic researchers and software developers, who might

not be not be interested in writing code to hit an API endpoint 75,000 times to get 75,000 objects. A single file containing all the required data seems to be much easier for certain use cases.

6. Who supports use

Mid-size museums are not well-equipped to offer support for digital resources. Unlike, for instance, a library or archive, the information management and technology staff are internally-focused, not public-facing. Curators, educators, and docents, who are often the public face of the museum, are often unaware that our digital resources exist.

Because of this, we have worked closely with local universities, in particular the University of Pittsburgh's Information Science program, the Carnegie Mellon Digital Humanities program, and the Frank Raytche STUDIO for Creative Inquiry. We've worked with faculty and staff there, providing access to curatorial and digital team members one-on-one to help them enable use of these collections in their programs for teaching, research, and artistic reuse amongst their students.

Finally, our hope is that through the standardization work that we've been undertaking with Open Knowledge International, we can work to make it so that enabling reuse and support can be shared across the industry—we can facilitate working with Museum data, not just Carnegie Museum of Art data.

7. Things people should know

One of the most important decisions we made was to release our data under a Creative Commons Zero (CC0) license. We were strongly influenced in this decision by Cooper Hewitt and the Museum of Modern Art, as well as from conversations with the digital humanities community. Attribution is extremely important to us, and we're extremely proud of our data. But the case was made convincingly that requiring attribution would be a burden to the most innovative and essential use we wanted to enable—projects that synthesize our data with others to generate new knowledge. By putting any restriction on the reuse of the data, many potential users would feel obligated to involve legal counsel to review their use, and that burden would be sufficient to prevent their use of our data. Instead of requiring attribution via a CC-BY license, we made it easy for people to give us credit—we told them how we'd like to be credited, and asked them kindly to do it. In our experience, almost every project that has used our data has credited us in some way or another.

A surprising takeaway for us has been that one of the primary users of our public data has been the museum itself. Easy access to our own data has enabled internal projects to be built on top of the published data, both because it's in an easy-to-use form, but also because of the permissive license. All of the data available is already approved for public use, so the approval process for remixing it and reusing it is significantly easier—"It's already public" is a wonderful

way to eliminate debate as to the appropriateness of using that information in public presentations.

Another important point that we missed on our initial communications is that we didn't adequately explain how we were using GitHub. GitHub is an essential tool in the Open Source community, and that community has a set of norms around how to provide feedback and suggestions on work that is released via the tool. Typically, if you found a mistake or wanted to improve a project that was available on GitHub, you would do so through a provided mechanism called a "pull request", where you would create a copy of the work, make the change, and ask the owner to approve merging your new version with the official version. Because collections data is not a standard use of GitHub, people were unclear whether or not we would accept corrections to our collections information through this mechanism. Matthew Lincoln, who originally brought this to our attention, suggested that it wasn't important what the answer was, as long as it was clear, and so we explicitly indicated that we would not take suggestions this way, and offered an email address that would accept such changes. This has been entirely satisfactory to all of our users, as well as our internal staff who were happy to accept suggestions, but were very pleased to learn that they didn't have to learn how to use GitHub to do so.

Open Knowledge International is keen to work on pilots with others considering releasing high quality tabular datasets in the open: <http://frictionlessdata.io>

8. What's next

Carnegie Museum of Art is hoping to release further iterations of its collections data over time. There are also now more tools that consume and generate Data Packages. It would be an interesting exercise to more deeply integrate features enabled by the Data Package descriptor. For example, CMOA can now add steps in the workflow that validate the dataset using tools like [Good Tables](#) to ensure that the data and the expectations declared in the datapackage.json match before publishing. Additionally, given the additional information stored in a Data Package, semi-automated export to other backend formats or databases can be offered relatively easily depending on interest.

CMOA and Open Knowledge International also hope to do work that supports the automatic generation of dataset documentation to ensure that documentation provided on GitHub through the README file matches that contained within the datapackage.json.

Facet 3: CalCOFI Hydrobiological Survey of Monterey Bay

Amanda Whitmire, Stanford University Libraries

1. Why do it

Researchers are beginning to understand the magnitude and complexity of the effects of climate change on our Earth system, and all research in this area is grounded in what we know about the past. Data collection at sea is labor-intensive and relatively rare, and technology has lowered that barrier only within the last couple of decades. Through this lens, we understand why in the marine sciences, the most valuable data collections are observational time-series studies, and the older the better.

When I realized the scope of the analog oceanographic data collections being housed at the Miller Library (a marine biology branch library in the Stanford Libraries system), there was no question that these materials needed to be digitized and shared openly. There are very few oceanographic time-series studies from the 1950s - 1970s, and these particular data only exist at our location. These data are an important contribution to studies in the marine sciences, climate change and coastal ecology. Our library is located in a tsunami zone, and since we have the only copy of these data, they are at significant risk of being lost.

2. Making the Case

Stanford Libraries has a Digital Production Group (DPG) whose primary focus is digitization of library collections for the purposes of preservation and access. Given the scientific relevance of the oceanographic data and its risk of being lost, it was not difficult to convince my boss (the Associate University Librarian for Science & Engineering) to support digitization of the material.

Our process for internally funding digitization projects is kept intentionally simple. Any librarian in our Science and Engineering Research Group is welcome to write a "Collection Project Proposal" (CPP; limited to a single side of one page) that describes the materials to be digitized, why they are important, what the goals for digitization would be, and an estimate of the costs. Our AUL reviews these on an annual basis and grants as many requests as are justified and he has the budget for. If a project idea comes up mid-year, we can also submit a CPP as needed. I proposed a pilot project to digitize a subset of the collection, and it was funded at \$5,000.

3. How you did it

My goals for this collection include moving a step beyond digitization of materials to create actionable datasets, but I am not prepared to address that because I am still investigating how best to accomplish such a task (automated text recognition processes, crowdsourcing, transcription services, etc.). This section will be a LOT more interesting once I get there, and the project will make more sense as a CAD Facet at that time.

For now, I'll focus on the process of material curation and how the digitization workflow works. Some of the process is being captured in an Open Science Framework project page. In concise terms, this was the curation plan that I made before I started (adapted from a great poster and using common sense), and it has largely been accurate:

1. INVENTORY - What do we have? How much do we have? What kinds do we have?

2. ORGANIZE - By cruise, station, variable, year? Standardize dates, stations, variables, cruise names...
3. APPRAISE - Are there duplicates? Is anything missing? Prioritize: what is most valuable or in the worst shape?
4. METADATA - Create descriptive & administrative metadata to guide digitization process: titles for collections in the digital repository, file names, etc.
5. DIGITIZATION - Stanford Libraries Digital Production Group has a well-equipped lab and staff for systematic digitization & deposit into the Stanford Digital Repository (SDR)
6. METADATA - Data need readme files and item- & data-level metadata to facilitate understanding & reuse; metadata from the DPG needs quality assurance and remediation.
7. MAKE ACTIONABLE - Conversion from PDF to actionable tabular data is critical for enabling reuse of the data. How do we make it happen at scale?

Steps 1-6 have been completed for the first batch of materials (data from every third year over the 23-year time-series). Steps 1-3 are time-intensive and the effort logically scales with the size of the collection. The DPG requires relatively little metadata to get the digitization process going, so Step 4 was brief. I am fortunate that we are so well supported by the experts in the DPG. They require submission of a digitization proposal via a standardized form that they provide, which ended up to be about 4 pages long. Based on the proposal, they provided an estimate of the digitization timeline and costs, and then moved forward.

4. Share the docs

As mentioned in the previous section, some content can be found at, “Whitmire, Amanda L. 2016. “Hopkins Marine Station CalCOFI Hydrobiological Survey of Monterey Bay, CA: 1951 - 1974.” Open Science Framework. November 30. osf.io/c3egt.”

The digitized items are not yet in the library catalog (also the discovery layer for the repository), but you can see a few examples of digitized material via direct links:

- A quarterly report: <https://purl.stanford.edu/qt035cq4651>
- An annual report: <https://purl.stanford.edu/dz088js0926>
- Field data: <https://purl.stanford.edu/xj314cj5427>
- Phytoplankton data: <https://purl.stanford.edu/qw382yy6150>
- Zooplankton data: <https://purl.stanford.edu/hy617cx4382>

5. Understanding use

The primary audience for these data is researchers, but I believe that they will not use the data for research purposes unless it is in a format that that can use. Meaning, text files with tabulated data. That is the main driver behind my desire to move a step beyond digitization (while recognizing that digitization is a critical action for these at-risk materials). I believe this because I used to be an oceanographer and I understand both their need for data like this and also the

constraints on their time and workflows. PDFs of legacy data are nearly worthless to a marine scientist who seeks to answer research questions.

6. Who supports use

After the data have been fully documented and converted to spreadsheets, the goal is that they can be used largely unsupported (setting aside the tremendous amount of work that goes into maintaining the digital repository). As a subject specialist and the curator of the collection, I am available to support data users. Interacting with 4-dimensional oceanographic data is generally handled in Matlab (the software of choice for most oceanographers) or R (an emerging choice in this domain). I expect most users of these data to be outside of Stanford.

7. Things people should know

This project feels important. Analog research data is everywhere - EVERYWHERE - and we need librarians and archivists to engage with faculty who are retiring to guide them in sorting through the maelstrom. I am focused on facilitating reuse in the digital space because my audience for these data are my former colleagues and I know that's where they operate. That said, identifying, curating, and archiving analog datasets to facilitate discovery and enable future reuse is critical. In my opinion, collections as data must necessarily extend to the analog world in order to keep up with the upcoming influx of materials from retiring faculty who worked in the pre-digital era. This project is an example of how we bring those data into the digital realm, but I encourage anyone interested in this type of work to reach out to faculty regarding their data. Do it today.

8. What's next

The most challenging part of this process is next: go from image or PDF to spreadsheets. This is the part of the project that has the potential for real-world impact. Nothing that I've accomplished so far is unique (important though it is). We've seen crowdsourcing, and we've seen transcription. What researchers really need is a way to liberate all of the older, analog data from paper into the digital medium that they use. If I can make progress on addressing how we might be able to do that at scale, I'll consider this effort a success.

Facet 4: American Philosophical Society Open Data Projects

Scott Ziegler, American Philosophical Society Library

1. Why do it

The American Philosophical Society Library (APS) has been digitizing historic primary sources for just about a decade. We've spent a lot of time smoothing out our workflow, and we feel like the

process is pretty well developed. However, we've known for some time that the audience for these scans are limited. The vast majority of our scanned material is hand-written (correspondence, diaries, ledgers, account books, for example). Reading this handwriting can be slow, and at times is a specialization in its own right.

We wanted to make this material available in a more approachable manner. We also wanted to give researchers an opportunity to easily interact with the material in different ways, including mapping and text analysis. Lastly, we see this as an outreach opportunity. We hope to build tutorials for students at the high school and undergraduate level to learn about visualization creation and digital history.

2. Making the Case

The administrative case for creating datasets from our collection was based entirely on our mission to increase access to our collections. This was a relatively easy case to make. However, there were additional hurdles to overcome.

Primarily, there are administrative concerns that the data we put out will have mistakes. This has proven to be the case. We try to include warnings that our datasets are created with attention to detail, but that errors happen. We're also cautious about how we label these datasets. We tend not to say that they are transcriptions (though, due to a dearth of synonyms, we do use the verb 'transcribe'). As an organization, we benefit greatly from large and professional transcription projects, including the Papers of Benjamin Franklin and the Papers of Thomas Jefferson. These projects are definitive representations of primary material. Our datasets are not. Our datasets are our attempt to make our material more usable, and usable for different types of projects.

In making the case for doing these datasets, we agreed to be clear about what we're putting out, to help draw a distinction between our datasets and professional transcriptions, and to supply feedback options for people who find mistakes.

3. How you did it

We identified the requirements for dataset creation to be:

1. ability to view a scan of the page being transcribed
2. ability to simultaneously view the software that the text is being typed into
3. versioning and/or revision history
4. ability to share among multiple people

We experimented with a number of crowdsourcing tools, including Omeka/Scripto, Omeka/Scribe, and Scribe Project. However, we quickly realized that the team we were assembling was small enough to rely on more modest tools.

We ended up using Google Sheets as the primary tool. We used dual monitors to ensure that the person creating the dataset can easily see the scanned page as well as the spreadsheet.

For the [historic prison data](#), our first major step toward thinking of our collections as data, we were lucky to have two talented and devoted volunteers: Kristina Frey and Michelle Ziogas. Kristina assisted in the early stages of the project, and Michelle did the majority of the dataset work.

4. Share the docs

We don't currently have any documentation, though we expect to create some during future projects.

5. Understanding use

We understand the use of our data primarily anecdotally. We think of our datasets as a means of identifying new institutional partners and collaborators. We monitor the use of our data via these partners. For example, we created the historic prison dataset from material in our library related to Eastern State Penitentiary. As we did this, we contacted the staff of the Eastern State Historic Site, and this has flourished into a fruitful partnership. Researchers come to our data through them, through our digital repository, and through the various third-party services we use to host our data. Several of these researchers have contacted us to offer their own data, to discuss additional projects, to show what they're building, and to offer corrections. This has been our principal measure of success.

We do maintain some metrics. The [Magazine for Early American Datasets](#) records the number of times datasets are downloaded. We also have a count of how many people download from our digital repository. These are helpful and appreciated. However, the motivation continues to be the new connections we make with individuals.

6. Who supports use

[blank]

7. Things people should know

When discussing this with people at libraries similar to my own, I tend to focus on the following:

- Datasets are easy to create. All you need to get started is a spreadsheet and something to transcribe.
- Material is easy to identify. We look for material that will work well as spreadsheets. Ledgers, printed forms, tallies, account books, are all examples due to their recognizable and repeatable format.
- Datasets are useful. You can save researchers' time by removing the challenge of reading handwritten notes; you can put material in a format that makes it easy to map; the material can be sorted, searched and filtered; you can promote the mission of your library.

However:

- Datasets need to be managed: Mistakes will slide in, and researchers will point them out; editorial decisions will need to be made, even in the most straight-forward-looking material.

8. What's next

Our flagship project to date – historic prison data – has gotten some positive attention, and we're eager to keep moving. We'll be hosting a digital humanities fellow to focus specifically on using the historic prison data. He'll be exploring various types of visualizations and analysis. We also hope to build a number of tutorials to encourage others to use the data for their own projects.

Additionally, we're working on two other open data projects. One involves a post office book kept by Benjamin Franklin during his tenure as Postmaster of Philadelphia. The other will involve a record of indentured individuals arriving in Philadelphia during the years of 1771-1773. Both of these projects will have academic advisory committees to help us strategize use cases and promote the data.

Facet 5: OPenn

Dot Porter, University of Pennsylvania Libraries

1. Why do it

We believe that users of manuscript data should have access to first-quality images and metadata free of technical or licensing constraints and this is what OPenn provides. First quality means the resolution at which the images were captured, and authoritative metadata in archival formats presented for easy reuse by humans and machines. Everything in OPenn is licensed as a Free Cultural Work.

2. Making the Case

The administrative case for creating datasets from our collection was based entirely on our mission to increase access to our collections. This was a relatively easy case to make. However, there were additional hurdles to overcome.

Penn Libraries has a commitment to Open Data, and the study of manuscripts in a digital age is the central mandate of the Schoenberg Institute for Manuscript Studies (SIMS) which is an integral part of the library and was founded in 2013. Much of the work of SIMS involves the reuse of our own digital manuscript materials, and we knew in 2013 that we could not do our job without a resource like OPenn. So we had to make one. The director of SIMS made the case for OPenn to the Director of Libraries, who made the decision to invest in the creation of OPenn.

3. How you did it

In 2013 Penn Libraries hired Doug Emery, who had created systems similar to OPenn for other projects, and he conceived the framework. The Penn Libraries did not at that time have a repository, so it was not in a position to host OPenn in an existing system. The Director of SIMS asked the Director of Libraries if we could set it up through Penn Central Computing. We started to populate OPenn with existing medieval manuscript image data; this was a challenge because although most of our manuscripts had already been photographed and cataloged, the master TIFF files were located in scattered hard drives and servers stored in various corners of Penn Libraries. This work was very intensive, and was carried out primarily by Jessie Dummer. We chose the manuscripts because they were central to the mission of SIMS and because the data was good. Doug Emery and Dot Porter designed a package and metadata structure for converting descriptive MARC and structural metadata into a TEI format designed for use and consumption integrating metadata with images.

Once OPenn was populated with Penn Libraries manuscript data we moved on to a second project. This project took advantage of the OPenn platform to gather into one location holdings from many different institutions, based around a common theme - 19th century travel diaries. This project has its own website, but the data served from there is all extracted from OPenn (<http://diaries.pacscl.org/>). OPenn now is the host for the Bibliotheca Philadelphiensis project, a project to digitize most of the Western medieval manuscripts in Philadelphia which received a \$500K grant from CLIR. SIMS's Curator for Digital Research Services, Dot Porter, is a co-PI on this project.

OPenn was designed to use the simplest and least expensive technologies available for sharing image and metadata. As such, technologically it is nothing more than a webserver with a very large hard drive that runs Apache and exposes the directory listings of its content. The content itself is static, comprising only images, TEI/XML metadata, text manifests, and HTML files. This data is exposed for ease of access and ease of movement via simple, well-established internet protocols: HTTP, anonymous FTP, and Rsync. One challenge that we had during implementation was convincing our service providers that what we wanted was something as simple as OPenn, without a query interface or an Application Programming Interface. Technologically, OPenn is more like an old-style software sharing website from the 1990s than it is a modern web application.

However this approach does have sustainability issues. Penn Libraries is currently designing and building a Samvera repository, and in the future we would like the data in OPenn to be stored in this repository, but served in ways similar to how it is done now. Storing the data in the repository will help with sustainability, and will also provide additional ways to serve the same data (e.g., using IIIF protocols). However we do plan to keep serving the data as friction-free as possible.

4. Share the docs

We have both a ReadMe and a Technical ReadMe file on the OPenn site:

<http://openn.library.upenn.edu/ReadMe.html>

<http://openn.library.upenn.edu/TechnicalReadMe.html>

5. Understanding use

Through OPenn, we provide well-structured standard packages that allow for machine and human reuse without putting any preconditions on how it may be used. We provide the data; users can do whatever they like. We are undoubtedly OPenn's primary user. We have built online bookreaders (generated with scripts from the TEI/XML files) that stream image files from the OPenn server, and we have also built downloadable epub electronic books (also generated with scripts from the TEI/XML files) that have copies of the manuscript images as part of the book.

6. Who supports use

ISC (Penn Central Computing) maintains the computer and storage, Jessie Dummer and Diane Biunno carry out the day to day work of managing and adding materials to the OPenn website. Dot Porter provides curatorial advice and oversight (and is also a superuser), and Doug Emery wrote and maintains the software and manages the project.

7. Things people should know

We serve digital assets on OPenn that represent physical materials that Penn Libraries doesn't own. OPenn is seen by us as an outlet for materials

OPenn treats digital assets as originals and seeks to build up a distinctive library of assets whether those originals are housed by Penn Libraries or not. The Open licensing in OPenn allows for easy collaboration with institutions local and international, many of which could not deliver this data in this quantity by themselves. It is a mistake to think that either the licensing or the ease of access to the materials is less important than the other - they are equally important.

8. What's next

We are going to move OPenn to the Library's Samvera repository to ensure preservation standards and long term sustainability and scalability. We will maintain an OPenn interface to this data, but the same data will also be able to be served through other methods including IIIF. We will also be expanding the content of OPenn from mainly medieval manuscripts to printed books and archival material.

Facet 6: Chronicling America

Robin Butterhof, Library of Congress; Deborah Thomas, Library of Congress; Nathan Yarasavage, Library of Congress

1. Why do it

American newspapers are a valuable primary source for research and study across a wide variety of disciplines – from political history to economics to epidemiology and more. The primary goal of the National Digital Newspaper Program is to enhance and expand access to American newspapers by providing free and open access to the data selected and gathered from institutional collections around the country to create one unified national collection of historically significant newspapers. By utilizing open data formats and schemas, communication protocols, and providing bulk data downloads, we can expose the collection to a very different type of use than through an individual user-based Web interface and extend the research value of the collection.

2. Making the Case

The administrative case for creating datasets from our collection was based entirely on our mission to increase access to our collections. This was a relatively easy case to make. However, there were additional hurdles to overcome.

The case for providing extended access to data had two aspects. Extending uses of the collection beyond the individual user was an opportunity to allow for new and enhanced uses of the content. In addition, the software developed for managing and displaying the data created under the program uses internal APIs and standard Web protocols for accessing data and communication within the software. To expose these internal mechanisms to external users was a low barrier to extending the use of this important federally-funded resource.

3. How you did it

An important component of envisioning the collection as a dataset was accomplished through emphasizing consistent and verified technical standardization of the file formats and metadata created under the program. To ensure this outcome, primarily for the purposes of creating a sustainable collection, the program developed highly-detailed technical requirements for data producers and provided a JHOVE1-based JAVA validation tool for ensuring conformance to key requirements. While minor changes have been made over the course of 12 production years so far, the dataset is largely internally consistent. (Most changes have been loosening of precise requirements rather than outright changes to technical specifications.) With a long-term vision for the program and specifically scoped goals (eventually involve all 50 states and territories, produce x amount of data per producer per 2-year grant, etc.), we strove to ensure that the data we received at the end of the program (some 20 years later) would be compatible with the data received early in the program. To that end maintaining strict data standards using open

well-document technical formats and a robust inventory management system has allowed us to achieve that goal to date.

With a reliable and consistent dataset, an access system could be built that both supported broad access to the collection and provided robust and flexible technical environment. The current system is based in the Django web framework written in Python which includes implementation of various open data access points and supports others. More information on [these access points](#) is available and the [code-base itself](#) is available.

Collaboration is a notable characteristic of the program not only with regard to the institutions providing data, but also with regard to the staff within the Library of Congress. Developers, digital library staff, program managers, and collection specialists alike had a stake in the development of the web site. Various views were created not only to assist programmatic access to the open data for digital humanists and researchers but also for digital library staff, program partners, and collections managers at LC.

4. Share the docs

Technical requirements for creation of the dataset are part of the [Technical Guidelines for the National Digital Newspaper Program](#). The National Endowment for the Humanities funds state representatives to select and digitize historic newspapers from their collections to conform to technical specifications established by the Library of Congress. All data created under the program is delivered to the Library for aggregation and public presentation, creating a large consistent dataset for historic newspapers (currently 12 million pages/45 million files).

Harvest and use of the data is documented on the [main web interface](#). A built-in reporting feature of the Django framework provides information and RSS feeds supporting use of the data at <http://chroniclingamerica.loc.gov/reports/> . The Django framework and Python code itself is [available on GitHub](#). In addition, a [listserv](#), hosted by the Library of Congress, supports data users through community input.

5. Understanding use

Learning about uses of the data is often indirect. As no API key is required to use the data, there is no register of people interested in using the data. On one hand, this is a primary driver for the adoption of the content in, for example, classroom settings. No API key means that it is very quick to get going with the content. On the other hand, it means we must infer use through various alerts and searches, for example, when we see a published article. In addition, as the content is public domain, there are no restrictions on the use of the content. This has led to a wide variety of uses, from commercial harvesting of the site to serving as a test dataset in a digital humanities class.

Some methods of finding out about the data use include Google alerts for the project name or social media posts, using common #hashtags like #ChronAm or retweets. (A former web

developer created a Twitter bot [@paperbot](#) that retweets when someone posts a tweet with a link to one of the NDNP pages.) Other methods include tracking metrics for the site; a huge traffic spike on a particular day to a particular page turned out to be a popular Reddit post. Similarly, if the content harvester or researcher is running into problems getting content from the site, they will reach out to us to figure out a better method. Researchers will also reach out for information about how to credit the site or ask questions about the parameters of the data, both through direct contact or through the [chronam-users listserv](#).

NEH also ran a [data challenge in 2016](#) to encourage direct use of the content. This led to some outstanding projects. One tracked how biblical quotations were used within the newspaper context; another combined the data with another dataset (Project Hal, a national lynching database) to provide more information about specific lynchings. Other researchers tracked the etymology of the word “Hoosier,” extracted the agricultural news, and created an interactive visualization for following a phrase over time/location. In the K-12 arena, an AP History Class used digital humanities tools to look at different historical topics in the newspapers.

6. Who supports use

There are a number of different layers that support the use of the data. Inside of the Library of Congress, the NDNP program specialists are often the first line of contact. The Library of Congress site provides an email contact option (Ask-a-Librarian), and reference specialists typically refer these questions to the NDNP program specialist. (Most users review all available documentation first and tend to use email contact as the last possible option.) The NDNP program specialists tend to answer some technical questions (pointing users to csv files), data questions (questions about OCR, limitations of the dataset), or query tweaking (instead of looking for fish pricing, search for specific fish prices in specific markets, such as market price of salmon in Portland versus local nearby markets).

For complicated questions, there are a number of other options. Sometimes the method the researcher/user is using can impact the performance of the website. In that case, the LC technology staff figure out how the researcher/user can get at the data without impacting performance (like downloading the bulk OCR bags instead of scraping the site). In other cases, the question is best answered by other users of the data. In this case, we recommend that users contact the ChronAm-users listserv (chronam-users@listserv.loc.gov). For example, another user might have already figured out a way to visualize given issues in a specific state by year. As more and more users work with the data, we encourage researchers to look at prior research, and point researchers to known current research efforts underway.

Publicizing and encouraging the use of the data is also mixed in with encouraging the use of the collection in general. The NEH supports the use of the data, such as the data challenge described above. Similarly, our education outreach team as well as National History Day serve as boosters for the use of the collection in general and the use of the data. As the project is a distributed model, our state project partners (universities, state libraries, and state historical societies)

encourage the use of content in the classroom, provide greater awareness of the content and what can be done with it via talks at conferences, etc.

7. Things people should know

Beyond the features that support individual Web browsing, Chronicling America also supports access to all data through common Web protocols and formats, providing machine-level views of all data for harvesting and large-scale bulk download. As examples, researchers can harvest batched digitized page images as JPEG2000, PDF and/or METS-ALTO OCR, or bulk OCR-only batches. Each newspaper page includes embedded Linked Data using a number of ontologies and supports JSON and RDF views. US Newspaper Directory bibliographic records are also available as MARCXML. The open API includes industry-standard endpoints like OpenSearch and supports stable intelligible URLs.

To accommodate data harvesting activities, the Chronicling America Web site infrastructure and workflow includes several features specifically designed to support such work:

1. During data ingest, additional text-only data sets are created and stored separately ready for bulk download.
2. To create transparency and ease of access to the bulk downloadable data, feeds for the downloadable files, in both ATOM and JSON format were added. Researchers can subscribe to the feed to ensure they get any new data that is added.
3. For the interactive API (JSON & RDF) caching was added to provide fast responses for pages that need to be created “on the fly” by the server (as opposed to the bulk processed data that exists in flat files).

For the user, we intentionally provide access and support to users with a wide variety of needs and skills. For example, a student can download a csv file of all of the digitized newspapers available on the site; the csv file includes information about the title, first issue digitized, final issue digitized, state, etc. A researcher might be interested in large-scale text analysis; for that user, all of the OCR files have been bagged and are available for bulk download.

8. What’s next

Planned infrastructure and interface design upgrades as well as endeavors to integrate and streamline digital content presentations at the Library present challenges and opportunities related to API access to data collections. Planning is underway to integrate the Chronicling America dataset into the general digital collections of the Library. Providing API and bulk data download access to Chronicling America data has proven to be a valuable service, and as such, maintaining equivalent or improved access after integration is a priority for the Library. Much of the available digital collections at the Library of Congress lack API documentation or bulk data access. Leveraging the work done with Chronicling America in these areas, more data collections at the Library are expected to take advantage of the same approaches used by Chronicling America in the near future.

Facet 7: La Gaceta de La Habana

Paige Morgan, University of Miami Libraries; Elliot Williams, University of Miami Libraries; Laura Capell, University of Miami Libraries

1. Why do it

The University of Miami Libraries Cuban Heritage Collection (CHC) received funding from LAMP (Latin American Materials Project) and LARRP (Latin American Research Resources Project) to digitize its holdings of La Gaceta de la Habana in 2015. La Gaceta is a significant historical resource, in that it was the paper of record during the Spanish colonial occupation of Cuba; and the CHC holds one of the largest collections of the newspaper outside of Cuba, with nearly 50 years of issues (from 1849-1899).

As part of our regular digitization workflow, we also create a plain-text file generated through Optical Character Recognition (OCR), in order to make digitized material discoverable through our digital collections user interface. Our standard practice within this workflow has been to use uncorrected OCR. However, our digital collections interface (currently CONTENTdm) only allows discovery, rather than any sort of analysis. Associate Dean for Digital Strategies Sarah Shreeves was aware of the increasing interest in text analysis as a result of digital humanities activity, and she suggested that creating a dataset that was easily accessible for use in text analysis tools would be a useful experimental project for a few members of the Library's Digital Strategies team. Everyone involved was aware of the imperfections of the OCR'd files; but we were also aware of the relative scarcity of Spanish-language datasets, and aware that if we made high-accuracy OCR a condition for release, that we might never reach the point where the files were ready. At this point in time, we are more interested in learning what is possible with imperfect OCR, and learning how we can make significant small improvements, than we are in striving for perfection on first release.

We think that it is worth emphasizing the creation of this dataset as a learning project on multiple levels. One of those levels was institutional: our goal was to understand how much work was involved in preparing a large dataset (approximately 50,000 files), and what specific steps would be part of the workflow, both for La Gaceta and potentially for other datasets we might want to release in the future. On another level, it was a learning project for the three of us who were chiefly responsible because of our different backgrounds. As a Digital Humanities Librarian without an MLS/IS, Paige Morgan brought hands-on experience with text mining, and with creating and preparing corpora, but lacked experience with corpus creation in the context of library systems for large-scale file management. Conversely, Elliot Williams (Metadata Librarian) and Laura Capell (Head of Digitization) had experience with library file management, but were unfamiliar with the specific needs of researchers who might want to work with the La Gaceta materials. This project was an opportunity for all three of us to begin fitting our expertise together and teaching each other enough to be able to produce materials efficiently. We see this as valuable preparation for future similar projects where we bring in people who may have vital

expertise with a particular set of materials, but who may be less familiar to the processes involved in creating machine-readable data.

2. Making the Case

There was considerable enthusiasm for this project, both from library administrators, CHC curators, and library faculty who were excited about providing deeper access to materials than the Digital Collections interface allowed. La Gaceta is a significant set of texts for Cuban and colonial studies, and we are excited about being able to introduce interested CHC researchers and UM students to text-mining techniques with materials that are directly relevant to their studies.

Acting on that enthusiasm was not difficult precisely because we deliberately kept this project as low-key and low-resource-intensive as possible: three people were primarily involved, with brief consultations or assistance from three others. Generating the OCR'd plain-text files is part of our existing digitization workflow, so the new activity within this project was focused on finding the best way to share the files and document how to use them. Our estimate is that the total time spent on this new activity was around 4-6 hours. Keeping the project fairly low-stakes and experimental made it a more comfortable site for learning and collaboration for everyone involved. It was also helpful that our goal for this project was not just the end product of the La Gaceta dataset, but also a clearer understanding of the work involved, and the resources we might need in the future (i.e., an internal data repository, rather than an external GitHub site).

La Gaceta is an interesting test case for text mining release because it's an imperfect dataset. The paper is thin enough that opposite page images tend to bleed through, and creases and sometimes blurred text complicate the OCR process. The dataset is too large for every page to have its OCR checked individually – however, that makes it a more interesting test case. And even with imperfect OCR, distant reading still yields interesting results. We're looking for repetitive errors that might be fixable using a bulk find-and-replace – and hoping that doing so will be another aspect of useful learning for our team.

3. How you did it

For the initial digitization process, roughly half of the La Gaceta volumes were digitized in-house by UM Libraries personnel; and the other half were outsourced with funding from LAMP and LARRP. The combined output of this digitization process was approximately 4.2 terabytes of TIFF files (one file for each page of the newspaper), which were OCR'd in-house. Both the TIFF and plain-text files are stored in our dedicated digital collections server for preservation purposes, but for this initial release, we decided to focus on providing just the plain-text files as a bulk download, available through a GitHub repository.

The majority of our work was about deciding how to structure the files, and how they should be named – and for all of us, that meant learning about the differences between file management practices within a library context and the context of a DH researcher working with the files in a

text analysis tool such as Laurence Anthony's AntConc or Geoffrey Rockwell and Stefan Sinclair's Voyant.

To explain: when our La Gaceta holdings were prepared for digitization, they were separated in one-month chunks. Within each month, there would be separate text files for each page of the newspaper, so each month would contain about 100 files, since each issue is 3-5 pages long. We broke up the newspaper this way because although La Gaceta was a daily paper, breaking it down by day would have required substantially more time – enough to be unsustainable within our standard digitization workflow. We experimented with regular expressions to see whether it would be possible to break the months into days using the first few lines – but the results weren't quite reliable enough to be worthwhile. One month chunks of the newspaper worked fine for displaying La Gaceta within our Digital Collections interface. But what would it be like for researchers to navigate those materials in bulk within a text analysis tool?

The question that emerged from this thinking was about the ID for each individual .txt file, i.e. each page of the newspaper. Our standard digitization workflow also generated a 20-character filename for each .txt and .tiff file (e.g. chc99980000010001001.txt). This filename is the product of our house schema for internal file management, which has worked very well in that context: library faculty and staff who use it are familiar with how the filename breaks down into segments that identify the repository, collection, object, sequence, and format. However, this filename structure is not easy to parse for external researchers, especially not in tools like AntConc and Voyant. Would we need to change the filename to something more human-readable in order to make the dataset useful? What would the stakes of that change be? As a researcher, Paige wanted more legible filenames, while Laura and Elliot were resistant to the idea of multiple filenames for the same object, and what it would mean for the Library to potentially have to develop an alternative filename schema designed for functionality within text analysis tools.

Making a decision about the filenames was probably the most controversial/high-stakes aspect of this project, since it felt like it had major implications both for users and for the library personnel involved. In the end, for our initial release of La Gaceta files, rather than create simplified and human-readable filenames for each document, we created a roster that will allow users to match any filename to its month and year. Keeping the 20-character filename is advantageous since researchers can use the same ID number to access the page image through our catalog if they want to check the original image. As we make more releases, the question of a more human-readable filename will almost certainly come up again, and perhaps we'll work towards that alternate schema that's designed more for external researchers, rather than for internal library file management.

4. Share the docs

This project is still new enough that we're still in the process of adding more formal documentation – as we have it, we'll make it available through the [UM Libraries Collections As](#)

[Data website](#). Our current introduction to the dataset (including an explanation of the filenames) is here, in our main repository.

For now, however, we recommend exploring this dataset with Laurence Anthony's [AntConc](#). We recommend AntConc for three main reasons:

1. It's lightweight and easy to download and run on Windows, Mac, and Linux machines.
2. The main interface is adjustable in a way that will work well with the La Gaceta filenames.
3. AntConc is widely used enough that there are plenty of excellent tutorials, and even a [corpus linguistics MOOC](#) based at Lancaster University that features it – in short, lots of support for users who might want to use this dataset as they learn more about text mining.

While this dataset could also work with [Voyant](#) (particularly Voyant Server, which doesn't require an internet connection), the experience might be a bit rougher, just because of the sheer number of files involved, since even a single month includes around 100 pages.

5. Understanding use

Because of the early stage of this project, this is an area that we're still figuring out: we want to learn from what our users do and what they need, and continue refining this dataset or use the info to produce better datasets with future materials. One important aspect of this project is that the local campus community is relatively new to DH, and so getting to the point where we can better understand the use will involve at least some work on our part to model what use looks like. Since we released this at the end of the school year, we anticipate more opportunities to figure that out till this fall. We understand that our success in this area will depend on how much work we put into making sure that various communities are aware of this dataset and how to use it, and plan to produce more materials that help them learn what they can do.

We're very interested in responding to the needs that our users raise, and we welcome feedback and requests.

6. Who supports use

The fully digitized version of La Gaceta is supported by University of Miami Libraries faculty in the Cuban Heritage Collection and faculty who work with our distinctive collections. Use of the current release of the plain-text dataset is supported chiefly by Paige Morgan (Digital Humanities Librarian), in collaboration with Laura Capell and Elliot Williams, as we continue to refine the dataset according to user feedback. In addition to making the dataset available for individual researchers, we are also developing lightweight plans that instructors could adapt if they wanted to use the dataset as a smaller or larger unit within a particular course.

7. Things people should know

Our approach might be described as “ambitiously unambitious” in its scope – and that gave us room to think calmly and clearly about the new dataset that we were producing, and how it fit

(or didn't fit) with our existing digital collections and schema, and our local institutional practices, etc. Creating this dataset has helped to make some inchoate questions more explicit, and we think that seeing those questions more clearly is just as valuable as answering them – which we hope to do in future projects. We recommend this approach, especially for any institutions that are hoping to use the Collections as Data initiative as a means for helping their faculty/staff develop new skills and expertise.

8. What's next

In the immediate future, we want to make sure that we put sufficient energy into outreach, promotion, and support for the La Gaceta dataset, which should be valuable both as a training object for our local community, and for gathering feedback for future data releases.

We will also be looking for other materials in our collections that could be good candidates to be processed and released in formats that will be useful for digital humanities researchers. One obvious future project will be various parts of the [Pan American World Airlines Collection](#), which is in the process of being digitized – but we're certain that the Pan-Am Collection is just one of many potential projects.

Facet 8: Text as Data Initiative

Zach Coble, New York University Libraries; Scott Collard, New York University Libraries; Nicholas Wolf, New York University Libraries

1. Why do it

As part of a broader text-as-data initiative, New York University (NYU) Libraries is in the process of expanding access to the ProQuest Historical Newspapers collection. This project involves negotiating with the vendor for access to the corpus as a set of text files, acquiring and storing the data, and creating infrastructure to promote discovery, access, and creative uses of the new collection. At a high level, this is the type of work that librarians do every day, but the technical components of the project have presented a fresh set of challenges.

We are seeing an increasing number of requests for machine-actionable data at NYU Libraries, whether in the form of full-text collections, bibliographic metadata, or both, from data researchers seeking corpora to perform topic modeling, network modeling, machine learning, and other natural language processing tests. The most predominant disciplines at our university that are interested in these methods have thus far come from political science and the [Center for Data Science](#). We are simultaneously tracking the changes among publishers with regard to API access to collections, provisions for researcher worksets of publisher data, and other affordances for machine-actionable research using previously licensed content. In anticipation of an emerging trend, several departments at the library, including [Digital Scholarship Services](#), [Data Services](#), and [Digital Library Technology Services](#), are eager to get ahead of this changing landscape, to shape how our relationships with content providers can enable this type of research, and to reconsider what library-provided content will look like in this environment.

2. Making the Case

As with all of our new initiatives, it begins as a pilot. We are interested in exploring several significant questions: What is the best way to provide access to the data? How will researchers use it? A pilot provides a low-stakes mechanism to work through a set of faculty requests in order to answer these questions and then evaluate if and how we want to continue. In our experience, when we are upfront with patrons about the pilot status of a project, and make clear that we are not promising new services and that the whole thing might disappear in, say, six months, they respond favorably and appreciate the candidness.

We have also found that pilots are most successful when they have wide scale buy-in. A project like this has a variety of stakeholders - both internally from liaison, reference, collections management, data services, and metadata librarians, as well as externally from faculty and central IT. Clear and consistent communication with everyone during pilot process not only helps prevent surprises but also establishes buy-in through a collaborative work process.

3. How you did it

The project began with a faculty member asking a liaison library for access to government documents corpora. This prompted us to revisit our licensing terms for similar types of content, such as historical newspapers, and to look for cases where our licensing terms allows us to provide full-text content to our research community. Once we realized there was potential to meet an emerging need among scholars and to leverage existing resource agreements, we convened a working group to investigate the issues.

The project has been a joint endeavor bringing together several departments, including Digital Scholarship Services, Data Services, Digital Library Technology Services, Subject Liaisons, and Collection Development. Each brings strengths to this team project. Digital Scholarship members speak to researcher needs working with content not traditionally seen as “data,” in this case full-text historical content. Digital Scholarship can also draw on past experiences in digital humanities projects that have developed key techniques in text mining that we can bring to bear on how we shape the form of the data we distribute. Data Services team members bring an awareness of how researchers are wrangling, transforming, and analyzing data-driven projects, assisting patrons and librarians alike in how they conceive of the data embedded in the full-text content. Subject Liaisons will have interacted with faculty members and understand the scope of their needs. Collections Development can speak to the terms of licenses, will often know the institutional history of data collections acquired by vendors (often previous shipments of CD-ROMS, hard drives, and other storage media), and can help negotiate new terms as vendors begin to take notice of data-drive access requests.

The pilot is also a helpful use case for new mass storage services coming out of [Research Cloud Services](#), a joint initiative from NYU Libraries and central IT. Specifically, we are considering providing access to the collection through NYU’s mountable storage (another pilot!), which provides remotely accessible fast-as-desktop storage that is protected and backed up. Here, we will use this new storage service as a distribution point to researcher to enable restricted access that is both convenient and controlled.

4. Share the docs

We do not have any documentation that we have permission to share at this point, although we will share it via our various channels as it becomes available.

5. Understanding use

We have researchers interested in using the historical newspaper corpus for machine learning, topic modeling, network modeling, and other natural-language processing. To better facilitate a variety of research uses, we are currently investigating ways to reduce the data cleaning and preparation steps that individual researchers are required to perform. One example of this is OCR correction, as preliminary samples indicate there is a fair amount of incorrectly transcribed text. Additionally, the library would like to create mechanisms to query the corpus and create

subcollections (e.g. by a specific newspaper, timespan, or keyword) to facilitate use by researchers interested in working with the content but are not interested in massaging the data. At a broader level, the library sees this pilot as a new and creative approach to library forms of ingest, collection development, and information distribution. We want this use case to help inform our vision for next-generation library services and library collections.

6. Who supports use

Use of the historical newspapers corpus is supported primarily by Data Services and Digital Scholarship Services. Liaison librarians also have a significant role in outreach and patron support.

7. Things people should know

We are still early in the process and are eager to learn from our experiences. Thus far we have found that positioning the initiative as a pilot was helpful in making the administrative pitch because it allows us to try new things and, equally important, gives us room to make mistakes. Additionally, bringing in several departments has been helpful in scoping the project as well as getting buy-in from our diverse group of stakeholders.

8. What's next

Our next steps include plans to improve access, discovery, and outreach for the collection. After our data cleaning and processing work is complete, we want to ensure the collections is discoverable in the library catalog and other primary discovery avenues. Finally, we plan to begin outreach for the collection, which could include workshops as well as class-based instructional sessions, as we've found that sessions working with pre-packaged data sets are better.

Facet 9: #HackFSM

Mary Elings, University of California Berkeley, Bancroft Library; Quinn Dombrowski, University of California Berkeley, Research IT

1. Why do it

In April 2014 to celebrate the 50th anniversary of the Free Speech Movement at UC Berkeley, The Bancroft Library, the Research IT group in the Office of the CIO, and the School of Information at UC Berkeley held [#HackFSM](#), a hackathon around the [Free Speech Movement Digital Archive](#), as part of the Digital Humanities @ Berkeley initiative. The event brought together thirteen teams of UC Berkeley students to design a new interface for a subset of Bancroft's digital holdings on the Free Speech Movement.

The Free Speech Movement was an appealing, immediately recognizable subject of the hackathon. The Free Speech Movement is felt to be quintessentially "Berkeley", and while most students are aware of the movement, it is not necessarily well understood by those students. The hackathon offered an opportunity to raise awareness of the subject and there was an available dataset to work with in the Bancroft Library's Free Speech Movement (FSM) digital archive.

2. Making the Case

The hackathon served as a valuable opportunity for groups in very different areas of the university, with different priorities and organizational cultures, to work together towards a shared vision. There were areas of administrative overlap, particularly between the Library and Research IT groups, and clearly defining roles and responsibilities was essential. #HackFSM was a highly collaborative and interdisciplinary effort, made possible by the participation of the Library Systems Office, Library Administration, BIDS, the School of Information, Arts & Humanities Division, Social Sciences, and the students from various disciplines, in addition to the Bancroft Library and Research IT. The relationships formed through participating in this hackathon have continued to benefit campus through the development of new collaborative initiatives.

3. How you did it

See the white paper (below).

4. Share the docs

[#HackFSM: Bootstrapping a Library Hackathon in Eight Short Weeks](#)

Abstract: This white paper describes the process of organizing #HackFSM, a digital humanities hackathon around the Free Speech Movement digital archive, jointly organized by Research IT at UC Berkeley and The Bancroft Library. The paper includes numerous appendices and templates of use for organizations that wish to hold a similar event.

5. Understanding use

There was never an explicit discussion of “use”; it was left up to the individual student teams to define the audience for their project, and what “use” looked like. Responses varied, and included a tool for conducting research, multiple browsing / exploration interfaces, and a few that were more like an exhibit.

6. Who supports use

The HackFSM team included The Bancroft Library, the Research IT group in the Office of the CIO, and the School of Information at UC Berkeley. The data preparation for the API involved the Library Systems Office and the Bancroft Library. In order to govern access to the Library’s FSM API, ResearchIT staff used a common-good campus service (no cost to users) called API Central, provided by UC Berkeley’s Information Services and Technology department. The API Central service provides a proxy to the Solr API, and can be configured to require credentials in order to process an HTTP Request (credentials are values of app_id and app_key headers that are set in the HTTP Request Header). University IT staff, I-School faculty, Berkeley alumni, and individuals from local tech companies served as code mentors during the hackathon. Eventbrite was used for registration of participants. Social media accounts (twitter and Facebook) were used to promote the event. During the hacking period, students, mentors, and event organizers communicated via Piazza, a free platform that offers a course-based message board, commonly used in STEM courses at UC Berkeley.

The Library administration offered space, as the new Berkeley Institute for Data Science space and the UC Berkeley School of Information for the opening and closing events. During the hackathon students were encouraged to make use of physical collaboration space provided by our new social sciences D-Lab and library.

7. Things people should know

Projects like this are highly collaborative and require technologists as well as content providers. The most successful outcome of the project was student engagement. Students from across disciplines came together to build something.

Maintaining the winning sites was not successful and we need better method and practices to achieve a record of this work.

While the main work product was a website, the greater product was that developers and humanists learned to communicate and work together. It was humanists and technologists working and talking together, learning from and collaborating with each other in the process of building new scholarly output. Hopefully events like HackFSM can prepare them for future collaborations in a research environment where such interdisciplinary projects will be more common.

8. What's next

Our hope is to prepare more digitized collections as data so they are ready to be used computationally. Current OCR could be improved and brought to a point of being “research ready” for computational use. We plan to write a grant to prepare a large recently digitized archival collection, working with local data scientists on the requisite steps we would need to take to get the data to a point of usefulness.

Facet 10: HathiTrust Research Center Extracted Features Dataset

Eleanor Dickson, University of Illinois at Urbana Champaign

1. Why do it

HathiTrust Digital Library is a massive digital collection, comprising more than 15.8 million volumes, and growing. HathiTrust aims to leverage the scope and scale of the digital library to the benefit of research and scholarship. The collection includes considerable material under copyright or subject to licensing agreements, which prohibits HathiTrust from releasing much of it—either in the form of plain text files or scanned pages—as freely-available data. The HathiTrust Research Center therefore develops tools and services that open the collection to data-driven research while remaining within the bounds of copyright and licensing restrictions, allowing only non-consumptive research.

One way the Research Center approaches this goal is through tools and technical infrastructure that mediate access to the data, including web algorithms researchers can run on HathiTrust data, the HathiTrust+Bookworm visualization tool, and the HTRC Data Capsule secure computing environment. Results from a user-needs assessment for text analysis conducted by the Research Center, as well as anecdotal evidence from researchers affiliated with HTRC, evinced the value of flexible, open data for text analysis research. To this end, the Research Center released the HTRC Extracted Features Dataset in 2015, which includes metadata and data derived from the HathiTrust corpus. The derived “features” in the dataset include page count, line count, empty line count, counts of characters that begin and end lines, and part-of-speech tagged word counts. The first release (v.0.2) included 4.8 million public domain volumes from the collection, and second release (v.1.0) opened 13.7 million volumes from the collection, representing a snapshot of the entire HathiTrust Digital Library circa 2016.

2. Making the Case

The HTRC Extracted Features dataset was in part born from other projects at the Research Center, including the Andrew W. Mellon-funded HathiTrust+Bookworm project, that required the HTRC to process full volume text into alternate formats. The team working on these projects realized that the data they were deriving would likely be useful to researchers and satisfy the HTRC’s policy for non-consumptive research.

Much text analysis research begins with the process of generating so-called features from the original text, which are then counted and calculated to draw conclusions about the data. HTRC Extracted Features aids the researcher by providing the data already in feature format. Furthermore, this shift in format from full text to features distills the contents of the volumes into facts and metadata, discarding the original expression of the full text. The Extracted Features dataset therefore strikes a balance of meeting the needs of researchers in a non-consumptive manner.

The research opportunities created by the release of HTRC Extracted Features was understood throughout HathiTrust and HTRC, and after review, the dataset was released.

3. How you did it

Deriving the HTRC Extracted Features was largely the work of Peter Organisciak (University of Denver), Boris Capitanu (University of Illinois), and Ted Underwood (University of Illinois). Together they collaborated to create a data model and write code to derive the extracted features.

The resulting dataset includes: *For every volume: metadata, including bibliographic metadata, word counts, and page counts. *For every page in a volume: part-of-speech tagged tokens (words) and their counts. Metadata, including information about the page (number of lines, number of empty lines, counts of characters beginning and ending lines), and the language, which has been computationally determined.

HTRC Extracted Features are available in JSON format, where each file represents a volume. Within the JSON files, data is organized by page in the volume. JSON is a hierarchical file format popular for exchanging data, and it lends itself well to representing book data.

HTRC Extracted Features are available using [rsync](#), which HathiTrust tends to use to share data and is considered an efficient file transfer protocol. Volumes download in [pairtree](#) format, a highly-nested directory structure.

The data can be retrieved with a structured URL that includes the standard HathiTrust volume identification number. The rsync URL format is: `data.analytics.hathitrust.org::features/`. More information about generating the rsync URL can be found here: <https://wiki.htrc.illinois.edu/x/oYDJAQ>.

4. Share the docs

The following sources contain more information about HTRC Extracted Features.

Code to extract features:

- <https://github.com/htrc/HTRC-FeatureExtractor>

Data paper:

- Organisciak, P., Capitanu, B., Underwood, T. & Downie, S.J. (2017). "Access to billions of pages for large-scale text analysis." iConference 2017. Wuhan, China. <http://hdl.handle.net/2142/96256>

HTRC Extracted Features documentation:

- <https://wiki.htrc.illinois.edu/x/WQCGAQ>

HTRC Feature Reader toolkit:

- Python toolkit for interacting with HTRC Extracted Features: <https://github.com/htrc/htrc-feature-reader/>

5. Understanding use

The HTRC Extracted Features dataset is useful for both research and teaching. As discussed in section 2 above, the feature format provides the data in a derived manner that aids the research process without over-mediating access to the data. As structured and pre-processed data, it does not meet the needs of all users, for example those whose work requires access to bigrams or greater, though it is useful for research that follows the bag-of-words model or that starts from token counts. Demonstrated uses have shown the data's value in large-scale computational text analysis, such as text classification using machine learning techniques, and in-classroom for teaching data science and digital humanities. Exemplary uses are outlined below.

Text classification with HTRC Extracted Features

Ted Underwood at the University of Illinois has drawn on HTRC Extracted Features in his research on literary genres. His work in machine learning uses the features data, including words and word counts, characters, and computationally-inferred, page-level metadata, to make inferences about genre in HathiTrust. Dr. Underwood classified volumes in the broad categories of fiction, poetry, drama, nonfiction prose, and paratext. His work classified over 800,000 volumes at the page-level, and resulted in a derived dataset containing word counts by genre and by year for volumes from 1700-1922.

More information about this research is available on FigShare: <http://dx.doi.org/10.6084/m9.figshare.1281251> .

Pedagogical application of HTRC Extracted Features

Chris Hench and Cody Hennesy at the University of California, Berkeley have developed a module for the Berkeley Data Science Education Program that makes use of HTRC Extracted Features. In the first iteration of the module, students documented the use of Extracted Features in data visualization, mapping, and classification in Jupyter Notebooks. Their Notebooks will be re-used in the classroom over the next year. Chris will introduce the curriculum to students in his course, "Rediscovering Texts as Data." In that multidisciplinary, digital humanities class, students will build on the existing Jupyter Notebooks as they develop coding skills. Chris also imagines using the Notebooks in workshops with non-programmers, where they will provide a legible introduction to text analysis by revealing how Python code is used to interact with the data without requiring attendees to program.

The Jupyter Notebooks are shared on GitHub: <https://github.com/ds-modules/Library-HTRC> .

6. Who supports use

Use of HTRC Extracted Features is supported by two main groups within the HTRC: the HTRC Tech Team and the HTRC Scholarly Commons. The HTRC Tech Team is comprised of research programmers, software engineers, and researchers (faculty, postdocs, and graduate students) affiliated with the [University of Illinois School of Information](#) and [Indiana University Data To Insight Center](#). The HTRC Scholarly Commons group is made up of librarians from the University of Illinois and Indiana University who are affiliated with digital scholarly initiatives at their local campuses.

The Tech Team provides technical support for the data, including writing the code to generate the features, processing data on supercomputers at the University of Illinois and Indiana University to derive the dataset, and providing reliable access to the data. The HTRC Scholars' Commons supports research and teaching with the suite of HTRC Tools and Services. The Scholars' Commons leads workshops, conducts outreach, and offers support to researchers who have questions about using the dataset. The HTRC Tech Team and Scholars' Commons have collaborated on questions of data curation and preservation of the dataset, discussed in more detail in section 7 below.

7. Things people should know

At the scale of HathiTrust, challenges to access and storage become particularly acute. Crunching feature data for millions of files is computationally expensive, and requires access to high performance computers. HathiTrust is also a non-static collection: Volumes are added daily, and (with less frequency) volumes are removed. For these reasons, HTRC has versioned the dataset following a "snapshot" model. Due to the time it takes to generate the features, the dataset will never be exactly current with the HathiTrust Digital Library, but instead captures the collection at a moment in time. The Research Center continues to provide access to both extant versions of the dataset, [v.0.2](#) and [v.1.0](#), but in the future, may have to look to alternate models for access to versions. Each version of the dataset is terabytes in size and storage may prove an issue if every new version includes features for the entire corpus.

Others interested in creating derived datasets as a model for opening access to restricted collections should consider what features would be useful to their researcher community. In addition to the token (word) counts, HTRC Extracted Features includes additional metadata, some of it processed from MARC records and others calculated during feature-extraction, that we hope provides valuable context for researchers who want to make use of the dataset. Other collections with other perceived user communities may want to include additional features.

8. What's next

As HathiTrust continues to grow, the HTRC Extracted Features dataset will be periodically updated with new versions. Between the first and second releases of the dataset, significant changes were made to simplify the data model that required all of the data to be re-crunched. In future releases, only new or differing files may need to undergo feature-extraction. Still, there

are some issues in the existing data, primarily related to the tokenization of Chinese-, Japanese-, and Korean-language text, that HTRC plans to improve on in future releases.

Facet 11: Beyond Penn's Treaty

Michael Zarafonetis, Haverford College; Sarah M. Horowitz, Haverford College

1. Why do it

At Haverford, we believe that libraries should move beyond the creation of digital images of original sources. Digital materials should allow scholars to do interesting and amazing things with our unique collections beyond what is possible with their physical incarnation rather than trying to replicate the experience of the original. We believe that “digitization” encompasses all of this work, rather than just the creation of images. As part of our efforts to make our collections available to a wider set of users and to be used in new and interesting ways, we have developed a number of projects that use this expansive definition of digitization with public facing websites that facilitate exploration of the collections.

Beyond Penn's Treaty fits into this effort for a number of reasons. While it includes digital images of materials—primarily journals and letters written by Quaker travelers in the late eighteenth and early nineteenth centuries—it also has added value in the form of TEI encoded and linked text, as well as further information on the people, places, and organizations encoded. The materials from Quaker & Special Collections included in the project are frequently requested, making them good candidates for digitization and wider distribution.

2. Making the Case

The types of materials included in this project are some of the most requested by researchers and scholars using Quaker & Special Collections. Many of the included documents had only recently been cataloged as part of a grant-funded project. Because much of the work for the project was in-scope for the Digital Scholarship team (creating databases, writing code, etc.), we needed only informal approval from the library director. She approved it based on the project's ability to showcase these newly-cataloged materials and add to our growing collection of digital collaborations between Quaker & Special Collections and Digital Scholarship.

3. How you did it

We collaborated with colleagues at the Friends Historical Library (FHL) at Swarthmore College to add their materials to the digital collection of travel journals and letters. Items from Haverford and FHL were scanned in their respective departments. The Digital Scholarship team at Haverford, at the time composed of two DS librarians and several student assistants, then migrated the digital objects from a CONTENTdm instance to a locally hosted Omeka instance with the Scripto/Scribe plugin and theme to facilitate transcription. Student workers in the library (in both DS and Quaker and Special Collections) transcribed materials during their shifts. Summer interns at Swarthmore (2016) and Haverford (2017) encoded the materials in TEI XML and shared those transcriptions in a Google Drive folder while also producing a master database (Google Sheet) of biographical, location, and organization records. An additional intern also

worked on cleaning geographical data and building maps tracing travel routes recorded in the documents. Student interns were overseen by staff from Quaker & Special Collections and Digital Scholarship with expertise in the subject, technologies used, and metadata. Pat O'Donnell at FHL provided subject expertise in Quaker biography and history, as well as experience with authority control for Quaker records, to help build out the database and provide quality control for the records created. The transcribed and encoded documents are made accessible to the public in a custom-built Django site—Beyond Penn's Treaty—that provides multiple entry points to the collection. Users can explore several maps that trace the routes of Quaker travelers and search across the entire collection for person, place, and group names. The encoding of the documents creates future opportunities for visualizing the collection based on researcher interests.

4. Share the docs

The TEI XML documents are publicly available in a [Github repository](#), as is the code for the [Django site](#). We have a [Google Doc](#) with instructions for scanning, transcribing, and encoding materials.

5. Understanding use

Like most of our digital scholarship projects, Beyond Penn's Treaty is outfitted with Google analytics to allow us to track basic metrics of use on the page. However, beyond that, our data about use is mostly anecdotal. Since we provide all the materials for people to download and use, we only hear about these uses if they get in touch. As a relatively new project, we are not aware of any major uses of this data.

6. Who supports use

Use of the data is supported by Digital Scholarship and Quaker & Special Collections. The Coordinator for Digital Scholarship and Services and the Digital Scholarship Librarian have led the development of the Django site, with regular input from the Head of Quaker & Special Collections. In the past year, encoding and transcription work and some of the Django development has also been managed our Metadata Librarian, who has dedicated time for DS projects built into their job responsibilities and is a member of the DS team. Special Collections and DS staff continue to work together to identify funding opportunities and to create student internships to continue the digitization, transcription, and encoding of new materials.

7. Things people should know

Much of the work involved with this project was done by student interns. This is a familiar model for us, and one that works well in an undergraduate liberal arts setting. Using students is not necessarily less work than doing such a project in other ways, however, as they need lots of oversight and supervision. Such deep opportunities can be transformative experiences for students and rewarding for all those involved in such projects.

While this was a new project for us, it is built on other work we had done. We have used Django as the framework for a number of other projects, such as [Quakers & Mental Health](#), and the

transcription and transformation process we employed was similar to that of the [Ticha project](#). The project also built on the strong collaboration between Digital Scholarship and Quaker & Special Collections.

8. What's next

Since all of the documents in the project are encoded in XML, we can create visualizations of many different kinds to explore the collection as a whole and the connections between people, places, and groups within it. We also hope to integrate the people, places, and organizations that have been encoded into a Quaker linked data project that we are building. This application will allow researchers to explore connections across our entire suite of Quaker projects.

Facet 12: Ticha: A Digital Text Explorer for Colonial Zapotec

Brook Lillehaugen, Haverford College; Michael Zarafonetis, Haverford College

1. Why do it

The digitization, transcription, and encoding of these documents is part of Dr. Brook Lillehaugen's linguistics research on the Zapotec family of languages in the Oaxaca region of southern Mexico. The documents include printed texts and manuscripts written by Spanish monks, bills of sale, religious testaments, land deeds, and other manuscripts that include the Spanish, Latin, and Zapotec languages. The work has been done over the past several years and continues as the project team explores more archival material in Mexico. The transcription and encoding is crucial to creating a digital annotated version of colonial period texts that include the Zapotec language, which include morphological analysis within the texts. Additionally, the public interface features a transcription tool that allows the public to transcribe documents, providing avenues for students, other scholars, and indigenous community members to engage with the materials.

2. Making the Case

No administrative case needed to be made, as digital scholarship staff in the Haverford library supports faculty and student research. This project is essential to Dr. Lillehaugen's research. The main institutional or administrative barrier is obtaining permission from various Mexican archives to make the images publicly available.

3. How you did it

The project is composed of several workflows. The first is digitization of archival manuscripts (bills of sale, religious testaments, etc.), which is done primarily by project team members—faculty, student research assistants, and librarians. The Ticha project employs a postcustodial approach to the creation of the digital archive. The digital images are organized and stored in a Dropbox folder, and uploaded to an Omeka instance with the Scribe/Scripto theme and plugin combination. There they are described by student assistants, and made available for transcription. Once the transcriptions are complete, they are visible alongside the image of the manuscript.

For printed texts and bound volumes, transcription and encoding is done by students in Dr. Lillehaugen's Colonial Valley Zapotec class. Using Git and Github for version control, students transcribe texts digitized at the Internet Archive and push their work to a remote repository. Making several passes at their assigned sections, they encode for language, outline structure, and formatting in TEI XML markup. We chose TEI to adhere to an encoding standard for texts, and to draw comparisons across texts in the growing collection. This XML markup is merged with an export of morphological analysis from the Fieldworks Language Explorer (FLEx), a popular software package in the field of linguistics, which is then rendered into HTML for the public site.

The public website is built in Django, a Python framework for the web, because many of our student assistants are Computer Science majors who learn Python in their introductory courses. Using the Omeka API, we can update the data and metadata in the archival materials section of the site by running a Python script. We also provide a download link to the plain text transcriptions of each page on the website. A bulk download option of all texts is coming soon.

4. Share the docs

Most of our documentation is in the [Github repository](#) for the encoded texts.

5. Understanding use

The materials on the site can be used freely under a Creative Commons Attribution and Share-Alike license. The encoded transcriptions are of research value to Dr. Lillehaugen and linguists who study the Zapotec family of languages. Access to the documents (both the digitized originals and the transcriptions) is important for community members to explore their language and history. By soliciting direct input from these community members and from workshops in Oaxaca that the public interface facilitates this exploration. We continue to consult our Zapotec speaking collaborators on design and interface questions.

By providing access to the encoded texts in TEI XML, we hope that scholars can find interesting ways of visualizing the collection.

We use Google Analytics to track usage of the project, and to help us make design decisions.

6. Who supports use

The Digital Scholarship team in the Haverford library provides technical support for the project, with server space for the public interface provided by Instructional and Information Technology Services. Mike Zarafonitis (Coordinator for Digital Scholarship and Services and a project team member), and Andy Janco (Digital Scholarship Librarian) provide project management and technical support for the project. Technical work (TEI quality control, Django project feature development, etc.) is done by student research assistants and DS student assistants. DS also provides instructional support for Dr. Lillehaugen's class, in which students collaboratively transcribe and encode the larger printed texts.

7. Things people should know

This project is very inclusive of undergraduate students in the work of transcribing, encoding, and developing the web platform for the public site. This is a model that is familiar to us in the Haverford libraries, and one that is aligned with our goals as a liberal arts institution. These students require a good deal of instruction and supervision, but such deep opportunities can be transformative experiences for them and rewarding for all those involved in such projects.

Additionally, members of the project team are very intentional about incorporating feedback from Zapotec-speaking community members. The transcription feature, for example, grew out of a request from speakers of the language who wished to contribute to the project. Thinking expansively about our user base, particularly beyond a strictly scholarly audience, is important.

8. What's next

We continue to add more archival manuscripts and bound texts to the public interface. Students are currently encoding and transcribing Fray Leonardo Levanto's *Arte de la Lengua Zapoteca*, and we hope to have the encoded version completed by the end of 2017. The next printed text for transcription, encoding, and analysis will be Juan de Cordova's *Vocabulario en Lengua Zapoteca*.

We also plan to add interlinear analysis of the Zapotec language to the archival manuscripts in the near future, which break down glosses by component parts. Interlinear analysis is already in place for some of the printed texts (see this [example page from Juan de Cordova's *Arte*](#)).

Facet 13: Vanderbilt Library Legacy Data Projects

Veronica Ikeshoji-Orlati, Vanderbilt University

1. Why do it

The Jean and Alexander Heard Library has become the repository for dozens of digital projects executed across the university. As stewards of these digital collections - encompassing databases, archives, e-editions, and exhibitions - it is incumbent upon us to ensure not only the availability, but also the accessibility of these resources to current and future generations. Every digital project is the product of hundreds, if not thousands, of hours of intellectual labor. To facilitate (re)use of digital scholarship pioneer and practitioner contributions requires that their work be thoughtfully curated, documented, and made publically available.

2. Making the Case

The administrative case for instituting a “data-first” policy of distilling the content and structures of digital projects into machine-actionable datasets is driven not only by ideological considerations but also practical ones. Fundamentally, the infrastructure to support continued development of sunsetted digital projects without personally invested stakeholders is lacking. The time and expertise required to satisfactorily migrate and maintain all sites built in Drupal 6, for example, is not fiscally viable if the library is to care for an ever-burgeoning collection of digital projects. In addition, the CLIR Postdoctoral Fellowship Program in Data Curation has allowed the library to experiment with integrating digital data curation practices into Digital Scholarship workflows.

3. How you did it

The first dataset curated by current CLIR postdoctoral fellow Veronica Ikeshoji-Orlati is the e-edition of Raymond Poggenburg’s Charles Baudelaire: Une Micro-histoire. Poggenburg initially published the Micro-histoire in 1987 as an entry-based chronology of the life of Charles Baudelaire (1821-1867). In the early 2000s, an expanded e-edition of the Micro-histoire was published by the Vanderbilt University Press and Jean and Alexander Heard Library. In 2016, due to the deterioration of the perl framework on which the e-edition was built and the library’s desire to increase the accessibility of the Micro-histoire’s contents, the data and metadata from the relational database underlying the e-edition were extracted into CSV format. Data cleaning was accomplished with OpenRefine, and the Library of Congress Metadata Object Description Schema (MODS) version 3.6 was selected for structuring the data and metadata in XML format. The dataset is currently in a github repository awaiting legal counsel’s approval for public release. The process of curating the Micro-histoire dataset was presented at the IDCC 2017 conference.

4. Share the docs

Legacy data curation protocols and institution-wide data management policies are currently being drafted. Each project, in its public release through the [library GitHub](#) account, is accompanied by documentation specific to that project.

5. Understanding use

Our goal in making Vanderbilt's digital project datasets publically available under CC0, CC-BY, or CC-BY-NC licenses (as appropriate) is to facilitate (re)use of the data in research and teaching contexts. It is anticipated that the communities currently utilizing the digital projects will engage with the curated datasets for their research purposes. In addition, new users interested in scholarly meta-analyses or large-scale quantitative research may incorporate the library's datasets into their work. In the case of the Poggenburg Micro-histoire dataset, for instance, Baudelaire scholars are the most likely audience, but those interested in broader questions in French history and literature may find the data of use, too. While the users for each dataset may differ, it is hoped that the curated datasets will also be of service to teachers working with students to learn how to interrogate humanities and social science data in meaningful and methodologically sound ways.

6. Who supports use

Members of the [Digital Scholarship and Scholarly Communications team](#) in the Jean and Alexander Heard Library are the primary facilitators for data acquisition, curation, publication, and use projects on campus. A new position, the Curator of Born-Digital Collections, has been created in order to continue curation efforts on library-housed digital datasets. In order to encourage campus use of the datasets, the Digital Scholarship team conducts regular workshops and hosts working groups in Linked Data and the Semantic Web, Tiny Data (data curation for the humanities), GIS, and XQuery to develop a cohort of data-literate faculty, staff, and students around campus.

7. Things people should know

As many data curators may already know, an overwhelming majority of one's time is given over to [data cleaning and standardization](#). To successfully run a data curation program within a library, it is critical to translate the lessons learned in curating legacy data sets to training programs in data management for researchers across campus. The data-driven research projects of today are the data curation challenges of the future, so establishing sound data management practices in current digital projects streamlines the process of ingesting them into the library's collection when they are completed. In addition, a data curation program must be grown in tandem with digital scholarship education infrastructure in order to arm teachers and researchers with the programming skills required to grapple with the curated datasets.

8. What's next

Currently, Veronica Ikeshoji-Orlati is curating the TV News dataset, a collection of nearly 1.1 million abstracts of news broadcasts from ABC, CBS, NBC, CNN, and Fox News dating back to August 5, 1968. The [Vanderbilt Television News Archive](#) is one of the richest resources for US

news reporting in the 20th and 21st century, but access to the metadata is limited due to the current web interface. In order to facilitate not only improved discoverability of news segments, but also quantitative analysis of the dataset as a whole, Ikeshoji-Orlati is collaborating with Suellen Stringer-Hye (Linked Data and Semantic Web Coordinator), Steve Baskauf (Senior Lecturer of Biological Sciences), Zora Breeding (Cataloguing and Metadata Team Leader), and Jacob Schaub (Music Cataloguer) to map the dataset to the [IPTC NewsCodes Vocabulary](#). In addition, she is working with Lindsey Fox (GIS Librarian) to enrich the dataset with geospatial data.

Facet 14: The Museum of Modern Art Exhibition Index

Jonathan Lill, MoMA Archives

1. Why do it

Since 1929, The Museum of Modern Art (MoMA) has been and remains the preeminent art institution in the history of 20th and 21st century visual culture. Through groundbreaking exhibitions about Cubism, abstract art, Surrealism, and other art movements, MoMA led the way in promoting artists who are now household names. MoMA established a holistic approach to the understanding of Modernism by exhibiting and establishing curatorial departments devoted to film, architecture and design, and photography. MoMA demonstrated that those fields of activity were worthy of critical analysis and appreciation.

The Museum Archives works continually to tell that history of the Museum, and to organize and provide access to the documents and records that evince those decades of activity. We strongly believe that exhibition history is an important scaffold that can be used to build an understanding of MoMA's accomplishments. [Indexing exhibition artists and curators provides researchers new pathways of exploration while linking archival resources and artworks in the collection.](#) This work helps increase exposure and use of MoMA Archives' historical collections and the dissemination of MoMA's history.

2. Making the Case

In 2014 the MoMA Archives received funding to organize and describe MoMA's exhibition files, which comprised paper records from all curatorial departments and the museum registrar for exhibitions staged since 1929. We decided that an exhibition index could be built as part of that project workflow. Due to our experience fielding public and staff inquiries and guiding user research, the Archives had developed an appreciation of the utility an exhibition index. How this data might be made available to researchers was unknown at the inception of the project.

Simultaneous to the Archives' work on this project, the MoMA hired a new director of web and video who was given the mandate of radically expanding the Museum's web content. She understood that our data could power the deployment of thousands of new web pages devoted to historical exhibitions, which could then be linked to numerous digital resources such as scanned press releases, exhibition catalogues, and installation photographs. Only with the web team pushing this project forward was the Archives able to move to completion. The new exhibition pages launched in September 2016. The data set was [published to Github](#) at the same time.

3. How you did it

The MoMA Archives had long maintained a simple list of historical exhibitions. I built an Access database, parsed that list, and imported a table of over 50,000 artist names from the Museum's

collection management system (The Museum System, TMS, vended by Gallery Systems). I created a simple interface that allowed interns to connect names to each exhibition using drop-down menus and when necessary to create new name records. Additional data was gathered from exhibition checklists scanned as part of the larger exhibition files project. The database structure allowed for easy review of the data, error checking, editing, and other maintenance. Once the indexing was largely completed, names in the index were reconciled to VIAF identifiers using the OpenRefine. The VIAF ids were then used to add Wikidata QIDs and Getty ULAN record numbers. Once this data was used to generate web pages, URLs for exhibitions and artists were added back into the dataset. Gallery Systems assisted with importing the data back into TMS from the Access-generated csv files. The web team extracted data from TMS to ingest into the web system as they do with collection objects and other data. A simple flat version of the data was posted to Github.

This project required close collaboration among several departments: the MoMA Archives, the data asset management system administrators who managed all the digital objects to be connected to our new exhibition web pages, the TMS administrators, and the digital media team. Importantly, this was the first time the Archives took responsibility for historical exhibition data in our collection management system and on the web site, involving us more closely in some key museum systems.

4. Share the docs

All documentation for the exhibition index and MoMA's collection are located on Github, along with the actual datasets: <https://github.com/MuseumofModernArt/exhibitions>

5. Understanding use

The immediate and most practical use of this data is for answering research inquiries: who was in an exhibition, how many exhibitions has an artist been in, how often two artists have been exhibited together, etc. This amounts to significant daily usage by library and archival researchers as well as the general public. With basic database or spreadsheet skills, more advanced inquiries can be answered by this data such as who was the youngest artist to be given a solo exhibition at MoMA? Or which artists have been exhibited most frequently without having works in the collection?

Separate from immediate needs of art historians and scholars, we expect this resource should be of tremendous use in classroom teaching about specific artists, modern art, and museology in America. Further, we believe this data can be used to connect digital and archival resources across the web. The exhibition index is less important for the information it contains than for the people, things, and data it allows a user to connect together. Its real potential is only realized when connected to Wikipedia entries, library union catalogs, and other datasets such as [Social Networks and Archival Context](#) (SNAC) or the American Art Collaborative. Ideally, this index can serve as a model for a multi-institution pooling of exhibition and artist data and online archival resources.

6. Who supports use

[blank]

7. Things people should know

To build an exhibition index with any speed, the materials that provide the data must be located and near at hand, preferably digitized, which is why conducting this work alongside a digitization or processing project is ideal. OCR of archival documents does not yield readily usable data. Facility with database applications and data manipulation software or programming languages is key. But most important is having labor to perform the data entry. Our workflow proved that with a narrowly constructed data-entry interface, precise detailed instructions, and proper supervision and review, that this work can be swiftly and effectively performed by non-professional staff and interns. Beginning with imported name records and other data increased efficiency and reduced mistakes. Error checking of the data showed that the error rate was within acceptable bounds and that most errors were omissions in data.

8. What's next

Our initial funding allowed us to build an exhibition index from 1929 through 1989 (while primarily processing and opening to the public tens of thousands of folders of paper records). A new round of funding is now allowing us to extend that work through 2000, merge it with more recent data created in TMS, and to further enrich the data by adding exhibition information such as department of origin, physical location, and subject tags. We are also working to combine this data with the exhibition index of MoMA PS1 (constructed as a smaller local project five years ago) and can begin to explore merging this data with that of other institutions such as the New Museum, White Columns, and other arts institutions.

Facet 15: Social Feed Manager

Laura Wrubel, Software Development Librarian, George Washington University; Justin Littman, Software Development Librarian, George Washington University; Dan Kerchner, Senior Software Developer, George Washington University

1. Why do it

Social media platforms produce and disseminate a record of our cultural heritage and are a source of data for answering research questions from numerous disciplines. After learning about a George Washington University faculty member's research which involved collecting tweets using a manual process, we developed prototype software in 2012 to connect to Twitter's APIs and help her collect data. Conversations with our university archivist highlighted use cases for collecting social media in the archives for future researchers. We saw a role for the library to build better tools for our community to conduct social media research. This led us to develop Social Feed Manager, which empowers researchers to build collections and enables libraries to proactively create datasets for use within their community. Along with providing data, we offer a consultation service for students, faculty, researchers—and also archivists and librarians—to access and use social media data.

2. Making the Case

Development of Social Feed Manager started through an IMLS Sparks grant and proceeded with support from National Historical Publications and Records Commission and the Council on East Asian Libraries. Library leadership participated and supported these grants which defined work proceeding from our existing relationships with faculty and archivists. Grant funding and project deliverables, as well as researcher and archivist needs, drove the allocation of staff time from developers, archivists, and librarians to support the work. Developing software and building a service supporting social media research might appear to be peripheral to typical library operations. Yet, the growing integration of the library's staff into research projects, including funded research, SFM's popularity with students at all levels, and the prominence of projects supported by data collected using SFM have become compelling evidence of its value and how this work supports library strategic goals concerning research and cross-disciplinary collaboration.

3. How you did it

Our initial project team in 2013-14, funded by a Sparks! grant from IMLS, was small and focused: the library's director of scholarly technology (who served as project manager and principal investigator), a software developer, our e-resources content manager, and a graduate student developer. In this first phase, we developed a suite of utilities and an administrative interface to manage collecting activities against the Twitter public APIs. A basic user interface provided access to data from Twitter user timelines, one at a time. We collected data of interest to the GW research community and in support of specific faculty and student research projects. This

included tweets by members of Congress, news outlets, and public sports and entertainment figures. The project team mediated much of the running of the data collecting and exporting data beyond simple downloads of an individual timeline's tweets.

In our second round of grant funding from the National Historical Publications and Records Commission and the Council of East Asian Libraries, we further developed the software and widened staff involvement in the project. Our grant funded the exploration of social media archiving and thus several of our archivists and our digital services manager participated as team members. The project included a significant software development component, as we added social media platforms, built a user interface to empower researchers to manage their own collections, and added more functionality overall to manage collecting from the Twitter, Tumblr, Flickr, and Sina Weibo APIs. To improve SFM's usability, our grant from NHPRC supported bringing on a UX consultant to conduct an expert review of its interface. We also brought on an experienced digital archivist to review the technical architecture and archival use cases. We wrote documentation and a quick start guide for both end users and other institutions using Social Feed Manager.

As a library, we actively collected tweets related to topics of interest on the GW campus. The largest and most heavily used collection has been our [2016 elections collection](#), containing over 280 million tweets. To facilitate making this data accessible to the GW community and beyond, a team member created [TweetSets](#), which provides a self-service interface for the GW community to download data and for the broader community to download tweet identifiers.

The changing terms of use for social media platforms and accompanying changes to APIs are a challenge both for maintaining working software and supporting research.

A current challenge is tracking and keeping up with the many research projects that use SFM. We want to be able to tell the story about the students and faculty in a wide range of disciplines and schools who are using SFM, and the contributions our librarians make to this work.

4. Share the docs

Documentation for the Social Feed Manager software.

The following documents are available through Social Feed Manager [project site](#):

- [Social media research ethical and privacy guidelines](#): general guidelines for GW researchers focusing on the collecting, sharing, and publishing of social media data
- [Social Feed Manager: Guide for Building Social Media Archives](#), Christopher J. Prom (2017)
- [Building Social Media Archives: Collection Development Guidelines](#)

The details of our software development work are available on [GitHub](#). This includes issue-tracking and prioritization, past and ongoing milestone activity, and release notes. We also

publish [blog posts](#) with each release, highlighting new features useful to the community and sharing tips for collecting and working with the data.

5. Understanding use

Our consultation model means that we typically have contact with users of Social Feed Manager and/or social media data and have an ongoing conversation about the analysis methods, findings, and outcomes of their research. This model also supports including discussion about ethical use of social media data.

In addition to being publicly available from TweetSets, several proactively collected datasets are available publicly on Dataverse, as sets of tweet identifiers. Twitter's terms of use do not allow full tweet data to be shared, but tweet identifiers may be shared for research purposes. A researcher can pull the full tweet, or "hydrate" it, from Twitter's API. Download metrics are available through Dataverse and its collections are highly discoverable via Google. We receive occasional follow-up requests or questions and track citations of datasets we've published.

Within the university, we are tracking schools and departments we've interacted with and monitor for published research that uses SFM, presentations, posters.

6. Who supports use

We have a team of software developer librarians who develop Social Feed Manager, provide consultations with faculty and students, teach workshops, and manage related services. Our subject specialist librarians are a frequent source of referrals. Our data services librarian sometimes participates in consultations, especially where they involve the larger research data lifecycle.

7. Things people should know

Ethical and privacy considerations need to stay at the forefront of this work and are a thread throughout the software development, research consultation, and instructional aspects of this work.

It is not enough to provide a tool for building social media collections: users will need support in understanding and optimizing their collecting parameters, understanding the data, and finding ways to manipulate or reformat it for analysis. We work with freshmen in writing seminars, undergraduates and graduate students from a wide range of disciplines, and faculty, with varying familiarity with CSV and JSON data, social media platforms, and research methods suited to social media data.

Social media platforms are constantly changing. Terms of use and API affordances are designed for commercial users rather than academic or research use. It's necessary to spend time understanding social media platforms, researcher needs, and staying up to date since what is

available is always changing. Advocacy for researcher needs can sometimes lead to change with platform terms, even if only over the long-term.

8. What's next

We are continuing to maintain Social Feed Manager and trying to keep up with changing API affordances. We're further developing our workshops and outreach on campus. The interest in our 2016 elections collection has led to our working with external audiences for this data such as journalists and non-profits, and we participate in conferences related to that work. We're being proactive about the 2018 midterm elections and collecting with future research uses in mind.

Appendix 3: Collections as Data Personas

October 2017 - April 2018

Collections as Data (CAD) Personas represent an initial set of high level role types associated with collections as data activity. While distinctions are fuzzy in the context of disciplinary and professional praxis, roles represented by personas can generally be understood in alignment with data stewardship or use. On the whole, personas aim to surface needs, motivations, and goals in context. These representations are derived from Collections as Data project engagements and project team experience.

In Agile software development, a persona is used to help develop a broadly shared orientation to user experience. Gary Geisler has written, “Personas offer a way to summarize findings from user research and help determine user requirements and priorities. These documents help project teams develop a common understanding of a project’s intended audience and priorities. They also serve as a useful reference for design decisions throughout the development process.”

Collections as Data Personas

persona /ˌpɜːrˈsɒnə/: an archetypal user


Faculty	Motivations	Specific Goals
 <p>Dr. Kylie Yu is an associate professor of US history at a small liberal arts college in the midwest. Their research is focused on representations of immigrants in the media in the 19th century and they teach the survey of US history, 1865 - Present, as well as upper level courses for history majors on immigration, identity, and urban life. They are not technophobic, but they have very little formal technical training.</p>	<ul style="list-style-type: none"> • Desire to uncover previously unknown information and create new insight • Need to earn the respect and trust of colleagues by producing high quality, interpretive scholarship based on solid research • Ambition to push the boundaries of traditional methods and take advantage of new research opportunities • Eagerness to design new assignments that will harness their students' interest in technology and teach them how to be effective and critical digital citizens • Interest in helping to develop better collections 	<ul style="list-style-type: none"> • Locate and download data that is trustworthy and appropriately formatted to work with their preferred software • Develop assignments for a new course with the Library's recently appointed Digital Humanities Specialist • Submit request to the IT department for 1 Terabyte of online storage for data projects so it is easy to review student work and so that it is easy to share with peers during peer review or after publication
<p>“I am very excited about the potential for digital methods to allow me to ask different questions than I am used to, but I don't think I have the tech skills to create useful datasets so I can begin experimenting.”</p> <p>“I was initially very excited about the digitized newspaper collections here, but the software I wanted to use only works on plain text, not PDFs, and I simply don't have time to download each page individually. I was also disappointed in the poor quality of the OCR text. Then I discovered the enhanced availability of computationally-accessible newspaper content provided by the Library of Congress and it blew my mind to consider all the research possibilities. This is going to be fun!”</p>		

Photo credit: [flickr](https://www.flickr.com/photos/kyliyu/). CC BY-NC 2.0

Collections as Data Personas

persona /ˌpɜːrˈsɒnə/: an archetypal user


Undergraduate Student	Motivations	Specific Goals
 <p>Julian Santino is a junior majoring in political science with a minor in music. He plays intramural tennis and is involved in several student organizations. He has already secured a summer internship with an NGO working on poverty in Central America and hopes to work for them full time after graduation.</p>	<ul style="list-style-type: none"> • Gaining valuable skills & experience for future employment • Graduate with a basic understanding of practical applications for the range of data processing techniques • The need to balance numerous curricular and extracurricular demands • Very limited time for research due to the length of the semester 	<ul style="list-style-type: none"> • Find data for a class assignment • Create and document a derivative dataset • Learn methods that support assessment of data representativeness and relevance • Complete a beginner's workshop on R and Jupyter Notebooks • Analyze the data with topic modeling • Credit data source in the final project
	<p>"I know that being able to work with data will be an important part of whatever job I have in the future."</p> <p>"People keep saying that I'm a 'digital native' and they assume I know more than I do. Just because I'm comfortable with social media doesn't mean I'm ready to be a programmer."</p> <p>"Until I took this class, I didn't know that librarians could help me find data and teach me software tricks. I thought the library was just books, online articles and study space."</p>	

Photo credit: University of Missouri System on [flickr](https://www.flickr.com/photos/missouri/1488888888/), CC BY-NC-ND 2.0

Collections as Data Personas

persona /ˌpɛrˈsɒnə/: an archetypal user


Data Curator	Motivations	Specific Goals
 <p>Ava Wright is a Data Curator at a public research library in New York. While the majority of her work is devoted to working with colleagues to acquire new collections and making sure they are processed correctly, she also regularly engages with visiting researchers. Her intimate knowledge of the collection helps her answer user questions and her interactions with users inform her thinking about description and access. She is seen as the local expert on text and data mining (TDM), and is recognized for collaborating with others on several successful pilot projects.</p>	<ul style="list-style-type: none"> • A sense of responsibility for ensuring the quality of data collections and their descriptive metadata • A commitment to data preservation throughout its lifecycle • A desire to make access as frictionless as possible for users, requiring less mediation • Being able to provide helpful answers to requests 	<ul style="list-style-type: none"> • Collect and make available data that are high quality (reasonably free of errors and omissions), well described and easy to access • Maintain and preserve digital collections as well as archiving versions and derived datasets • Manage data acquisitions more efficiently • Create documentation that informs patrons how repository data can be acquired and used
<p>“Our workflow is constantly evolving as the nature of collections and user requests change and we struggle to balance our desire to release high quality data with the fact that it will never be perfect. Data cleaning often takes more time than we have. We try to make up for this with great documentation.”</p> <p>“We use GitHub to keep track of the dozens of scripts we’ve developed to process new collections and slice and dice those collections based on user requests.”</p> <p>“We generally trust our repository system, but we know that our data are getting out into the world. We are committed to open access and love that our data are being used and reused and remixed, but we worry that they will lose their context once separated from our system.”</p> <p>“I work with subject specialists to negotiate data acquisitions, to get access to data via an array of delivery mechanisms (email, mail, FTP). We have many different types of license arrangements that we need to track, and sometimes renegotiate.”</p>		

Photo credit: Justin Lui on [flickr](https://www.flickr.com/photos/justinlui/). CC BY 2.0



Collections as Data Personas

persona / ɪˈpɜːsənə/: an archetypal user


Metadata Analyst	Motivations	Specific Goals
 <p>Becky Roberts, a metadata librarian at a large research institution, enjoys tackling difficult problems and finding solutions to them. She is familiar with XSLT, Python, and SPARQL, and enjoys browsing GitHub and professional lists to learn new tools and approaches, though she finds her time is limited for exploration. Becky spends most of her time analyzing and transforming metadata, ensuring it maintains integrity as it is processed for different endpoints. She is keenly aware of how her decisions affect end user access and discovery.</p>	<ul style="list-style-type: none"> • Increase broad content discovery and use by fully exposing metadata for our digital objects whenever possible • Foster development of interoperable metadata by establishing and proliferating efficient and consistent descriptive practices • Produce accurate metadata that supports (re)use 	<ul style="list-style-type: none"> • Develop reliable connections to authority services and reconciliation tools • Create interoperable metadata that standardizes and harmonizes local & global description • Assign rights statements and licenses to repository metadata to encourage reuse • Initiate automated methods for doing qualitative/computational analysis on legacy data to fix errors, especially as standards change
<p>“As I work with metadata and come across issues, I make an effort to give feedback to the standards communities so we can make things work better.”</p> <p>“I work with metadata in all forms, from all sources. I need to know the standard followed in creating it (but I find that rarely is the standard followed completely). So you deal with what you are given.”</p> <p>“I work with others at my institution to make as much metadata freely available for bulk download with a clear market indicating that the metadata is free of rights restrictions.”</p> <p>“I’m excited about exposing our metadata at scale, but we are still figuring out how best to estimate the resources required to get it done.”</p>		

Photo credit: wocintechchat.com, CC BY 2.0

Collections as Data Personas

persona / 1 perˈsɒnə/: an archetypal user


University Archivist	Motivations	Specific Goals
 <p>Torsten Melhus, a college archivist. With 15 years in this job, he possesses extensive knowledge of the college and is deeply committed to preserving its history. He manages the work of 2 full-time staff and 5 part-time student workers. Torsten is involved in all aspects of the department, from acquisition of new collections, to processing, to working with researchers. He is interested in exploring new ways to promote use of the college's collections, but has limited time and resources.</p>	<ul style="list-style-type: none"> Promote the college's collections, locally and internationally Ensure digitized and born digital material from the archives are presented and understood in their appropriate context Develop new, diverse ways in which archival collections can be used Encourage faculty, student, and broader community use of the collections 	<ul style="list-style-type: none"> Create metadata for digitized and born-digital objects, including provenance, reuse rights, technological context, and relationship to other materials Produce data visualizations for archival collections that support archival processing and provide context to researchers Enable use of computational methodologies with archival collections while still protecting property and access rights of the content
<p>"I am thrilled that we have implemented standardized rights statements from rightsstatements.org. It was a fair amount of work to get it in place, but scholars really appreciate knowing the rights status straightaway."</p> <p>"Last semester the archives collaborated with Dr. Marshall on a digital humanities mapping project. It was a smashing success and we all learned a lot. I want more faculty to incorporate the college's collections in their classes."</p> <p>"I feel archivists and scholars are really hampered by the lack of an easy way to produce useable data from scanned manuscripts and other texts. There must be a techy solution!"</p>		

Photo credit: Martin Dee, University of British Columbia photographer CC BY 2.0

Collections as Data Persona

persona /ɪˈpɜːrˌsɔːnə/: an archetypal user


Data Reporter	Motivations	Specific Goals
 <p>Eva Delgado works at a North American newspaper with a wide circulation in a major metropolitan area. Part analyst, part developer, part data scientist, she also keenly understands state and regional politics and education issues, frequently networking with researchers and policy makers in these areas. She is starting her 3rd year within a growing data journalism department that is increasingly central to the production of the big stories getting lots of attention from readers.</p>	<ul style="list-style-type: none"> • Produce attractive, informative visualizations with cutting-edge techniques to accompany data-driven stories • Improve quality of data wherever possible, developing solid workflows, documentation, and practices • Share work products via a public repository to promote, extend, and persist the impact of journalistic work 	<ul style="list-style-type: none"> • Find well-documented data from trustworthy sources • Move data to and from our cloud service, where it is processed, with minimal overhead and technical headaches • After a piece is published, transfer the processed data, documentation and scripts used to develop visualizations to GitHub or another repository • Keep track of data downloads and use
<p>“I am excited when I publish a citable data visualization, with links to my source data, alongside an article, because I know the viz will catch a lot of eyes and interest in my piece.”</p> <p>“Up to now, I’ve found a lot of my data on government agency web sites, though recent political developments make future access less certain. The rest I find at research centers and policy institutes. It’s getting easier to find things online on my own, which saves me a lot of time.”</p> <p>“I try to share most of my data on GitHub. Sometimes the data is too large. It might be nice to find research institutions that can host the data for reuse by others. GitHub is great, but libraries are likely better for persistent, long-term access to data.”</p>		

Photo credit: wocintechchat.com, CC BY 2.0

Collections as Data Persona

persona /ˌpɜːrˈsɒnə/: an archetypal user


<p>Library Administrator</p> 	<p>Motivations</p> <ul style="list-style-type: none"> • Recognition by Vice Provosts of the talents of library staff for their efforts to leverage collections for support faculty research innovation grants • Promote collaboration and communication across departments as teams roll out new initiatives • Ensure technical skill training opportunities for existing staff are plentiful and funded 	<p>Specific Goals</p> <ul style="list-style-type: none"> • Produce 5-year technology budget projection, based on input from the Operations and Engineering teams, to reflect needed changes to IT infrastructure supporting on-demand delivery of digital collections in bulk • Make a cross-team staffing plan for a new project to develop and implement standardized licenses and copyright statements for all online collection content • Find a reliable source for useful readings and latest news on AI, ethics, and licensing issues
<p>Tejaswini Sona is Associate University Librarian for Digital Scholarship and Technology at a mid-size elite research institution. Her department has a strong track record for executing grant-funded projects to create and publish online collections from digitized archives, books, and maps. When she earned her PhD in Medieval Studies, TEI had just come on the scene; she feels like she has been playing catch-up with digital humanities ever since.</p>	<p>“Lately the biggest problem we have is keeping up with the demand by faculty for our top-notch course-driven sessions and advanced workshops. A great problem to have!”</p> <p>“The team working on the Library’s ‘digital collections 2.0’ initiative is struggling to establish and articulate criteria for data quality. What does ‘good enough’ look like? How is success measured? How much work is required to re-process legacy collections for computational access? I hear these questions coming up at professional conferences, too.”</p> <p>“After a recent conversation with a program officer at a major foundation about potential collaborative projects we are exploring, I have new perspective on certain communities’ sensitivities to having their historical material open broadly for computational work. Careful framing of our approach in this space is critical.”</p>	

Photo credit: CC BY 2.0 hfrost@stanford.edu

Collections as Data Persona

persona /ˌpɜːˈsɒnə/: an archetypal user


Software Developer	Motivations	Specific Goals
 <p>Javier Hawkins is a software developer at a large public university in the Pacific Northwest. He is part of a digital humanities team responsible for assisting faculty and students with their research. Javier began his career at a large Silicon Valley startup and gained deep experience in software development, especially front-end applications. He began working in academia because of his interest in open-source software development.</p>	<ul style="list-style-type: none"> Partner with faculty, students, and university staff to build products and tools that interact with data to produce new research Share code and tools developed for local projects with other developers Collaborate with other library staff to build resources that support and preserve digital humanities work 	<ul style="list-style-type: none"> Discover and search data resources with a documented license to reuse the data Gain recognition within the university community that digital humanities work should be part of data science and research software engineering initiatives Develop applications that remix and expose data in novel ways, particularly through data processing and visualization techniques Partner with outside developers to produce and enhance open source applications
<p>"Digital humanities at its best is collaborative. It is good for me to connect with researchers to really understand the goals of their project. But sometimes it feels like all I do is sit in meetings and never have time to focus on actually developing something. It is difficult to find the right balance."</p> <p>"Everyone seems to be building an API these days. An API is nice, but not necessary. Providing data in a standard documented format that I can query with the same code and toolset is more important than the delivery method."</p> <p>"I know there's a ton of data out there that could be remixed in interesting ways, but it's so hard to find. As a result, most of our projects are forced to create the data set as well."</p>		

Photo credit: Daniel Foster, CC BY-NC-SA 2.0

Collections as Data Persona
 persona /i pər'sɒnə/: an archetypal user


Postdoctoral Researcher	Motivations	Specific Goals
 <p>Anna Pernod landed a postdoc position at an interdisciplinary humanities center at a large private university in the US where she teaches one class per semester. The rest of her time is spent researching, preparing articles for publication and working on turning her dissertation into her first book. She is French and completed her Ph.D. in Comparative Literature at Princeton. She is on the job market hoping to secure a tenure track teaching and research position at an American or European university.</p>	<ul style="list-style-type: none"> • Take advantage of new digital tools and methods in order to expand on work done as part of their dissertation. • Raise academic profile and impress potential employers with demonstrable experience doing cutting edge research. • Expand professional network to include others in their field also using computational methods. 	<ul style="list-style-type: none"> • Create a corpus on francophone poetry from France and its former colonies that spans from the 16th century to the present. • Use text analysis software to study patterns of word use as well as content across the entire corpus. • Develop smaller scale text analysis projects that would be appropriate for undergraduate literature classes. • Gain the technical skills to do this work independently.
<p>“Everything I do right now is with an eye on the job market. It is really my top priority. Many job ads are looking for ‘Digital Humanities’, but my education did not fully prepare me to do highly technical work. I’m looking for digital projects that are meaningful but also technically easy enough for me do.”</p> <p>“I went to a talk on campus about a text mining project that was really interesting, but it sounded like the researcher worked with a lot of people in the library to make it happen. I don’t know who to ask at our library.”</p> <p>“My limited experience with text analysis makes me worry a bit about the quality of the data that is available. Some digital collections I use look great, but the OCR text behind the images is sometimes very poor. I need to be able to tell how reliable a data set is so I can decide if it is useful.”</p>		

Photo credit: Cécile Lapoivre on [flickr](https://www.flickr.com/photos/celap/), CC BY

Collections as Data Persona

persona / 1 per_s0ne/: an archetypal user


Public Policy Data Analyst	Motivations	Specific Goals
 <p>Ed Kishiyama is a data analyst at a major non-government organization (NGO) based in the San Francisco Bay Area that focuses on immigration and related public policy issues. Inspired by his grandparents' experience at an internment camp for Japanese in the 1940s, Ed earned his master's degree in Political Science and decided to pursue a career helping immigrant communities. Now, on top of his day job, he teaches "Intro to Data and Data Science" as adjunct faculty for a Library and Information Science program at the state university.</p>	<ul style="list-style-type: none"> • Desire to actively influence the introduction and implementation of modernized laws that support immigrant rights at the federal, state and local level • Finding new data, both current and historical, that fill information gaps in the agency's ongoing research in core policy areas • Pass on his experience and skills to the next generation of information professionals 	<ul style="list-style-type: none"> • Develop more robust data management and documentation practices for his team, including tracking provenance, research method, version history and processing history • Write narrative and budget for a funding proposal to a private foundation to collaborate with a local university on a data archive for the network of NGOs in North, Central and South America • Evaluate the latest data visualization software package options
<p>"I spend a lot of time assessing the quality of data that we obtain. It's tedious at times, but ensuring that we are working with accurate data collected by authentic sources via rigorous -- or least reasonably documented! -- methodologies is fundamental to the success and impact of our work."</p> <p>"There's interesting research coming out of universities that leverages computational techniques to analyze historical data extracted from scanned records and to tie it to current trends. I try to stay on top of what the leading academic centers or institutes are covering. I figure it is key to growing my network."</p> <p>"I emphasize with my students the important role that libraries play in making data findable and helping people to find it, to serve as a data connector. If you can't find it, it doesn't matter how good the data is! For one assignment, they have to trace a dataset and its downstream use, how it is processed, republished, documented, and cited. An eye-opening exercise!"</p>		

Photo credit: SITCON on [flickr](#), CC BY-SA 2.0

Collections as Data Persona

persona /ˌpɜːrˈsɒnə/: an archetypal user


High School Teacher	Motivations	Specific Goals
 <p>Matthew Miller is a math teacher at a suburban high school in the southeast and has been working for seven years. He teaches mostly advanced math classes in statistics and calculus, though his schedule changes from year to year. He especially enjoys working with students completing the International Baccalaureate diploma; they are required to complete a senior project working with real-world data.</p>	<ul style="list-style-type: none"> Expand opportunities for students to develop applied math skills through computational analysis Incorporate 21st century digital technologies and methodologies into the classroom without needing to learn a programming language Build innovative lesson plans that meet the rigorous research requirements of the IB diploma program 	<ul style="list-style-type: none"> Create lesson plans that teach not only how to process and manipulate data, but how to find data on the open web and evaluate its authenticity/reliability Discover reliable big data sources students can easily understand, use, and manipulate for class projects Develop citation strategies for class projects to simplify and standardize the evaluation of source data sets and project results Partner with local resources (internal or external) to help students build technical proficiency and data literacy
	<p>"Despite their digital reputation, high school students tend not to have a high level of technical proficiency beyond using apps like Snapchat, Instagram, and Twitter. I wish I didn't have to spend class time teaching Microsoft Excel."</p> <p>"My school system recently purchased the G Suite for Education, which gives students and teachers access to a suite of Google collaborative tools. Are there any published lesson plans that make use of these tools?"</p> <p>"In my statistics classes, I incorporate real world datasets whenever I can into assignments. My students need the most help not with the math, but with finding and evaluating data sources they find on the internet."</p>	

Photo credit: SiYH on [flickr](#), CC BY-SA 2.0

Appendix 4: 50 Things



Want to support collections as data at your institution, but not sure how to begin? Drawing on what we learned from engaging with practitioners and researchers throughout the [Always Already Computational](#) project, the project team compiled a list of 50 Things you can do to get started. 50 Things is intended to open eyes, stimulate conversation, encourage stepping back, generate ideas, and surface new possibilities. If any of that gets traction, then perhaps you can make the case for investing in collections as data at your institution in a meaningful, if not systematic, way.

Our best advice: start simple and engage others in the process. You may find some activities listed here are already underway!

About this publication: 50 Things was published in October 2018 under a CC BY-NC-SA 4.0 license.

1. Know how optical character recognition (OCR) output is produced in your digitization workflows. What software is used? What formats are created? What levels of accuracy are produced? Where is it stored? Is it available for user download?
2. Create an inventory of full-text collections managed by your institution. Document rights status, license status, discoverability, and downloadability. Ask the question: are we offering optimal access for computational use of the full-text? How can we make it better?
3. Migrating a legacy digital collection to a new system or platform? Take the opportunity to make the content accessible to researchers that have computational projects in mind.
4. Interview the archivist, librarian, or curator responsible for a digital collection to document data provenance and decisions made in the course of collection processing and digitization. Work to make this information publicly available.
5. Inventory your data holdings. Just make a simple list. And then commit to keeping it up to date, and watch it grow.
6. Add new fields to the collection management database to indicate and describe data components.
7. Survey your digital collections to identify characteristics -- good metadata, open access, good OCR, high usage, relevance to a high-profile academic program or research area at the institution

-- which lend themselves to high impact as data.

8. Recognize and identify the things you need to do differently than have been done for physical collection objects.
9. Find out if your digital collection database or access platform has an API available for querying by the public. If it does not, see if it is possible to develop one. If it does, determine if it is actively used. If it is actively used, see if you can reach out to users and ask about their usage!
10. Talk to a colleague responsible for systems that provide networked access to digital collections about possible approaches to facilitate download of collection data in bulk.
11. Add a terms of use to your archival finding aids.
12. Read the language of your organization's collection deed of gift or purchase agreement to evaluate whether it allows for providing access to collection content in the form of data.
13. Review your digital collections metadata and evaluate the rights statements and license statements in terms of consistency and clarity. Are you able to adopt rightsstatements.org?
14. Socialize Collections as Data as something that can be supported by units and staff across the library. Identify some champions across the organization and people who have skills or position to do the work.
15. Talk to people responsible for research data management to encourage planning for data preservation and other considerations that make it possible for others to reuse the data in the future.
16. Review your institution's mission statement or strategic plan documentation, and consider if and how Collections as Data activities are aligned with and support it.
17. Share sample projects with community partners to give them an idea of how their collections can be used and be relevant to new ways of conducting scholarship.
18. Network with people who work with data and have the skills or knowledge you need to get your work done.
19. Identify barriers and limitations to what services you can offer support, and talk with colleagues about creative, feasible solutions to overcome them.
20. Publish or present on "Wikidata for librarians," including case studies of libraries working with Wikidata to expand discovery of collections.

21. Read up on IIIF (for example, check out this [useful tutorial](#)) and determine what hurdles to implementation exist at your institution. Then talk to relevant folks about what it would take to overcome them.
22. Read the resources in the Always Already Computational project's [Zotero library](#).
23. Develop a workshop focused on the use of data in and about collections; shop it to department faculty and incorporate it into research orientations for faculty and students.
24. Mentor a liaison interested in learning a data science skill who is well positioned to identify datasets and data support needs amongst their researchers.
25. Conduct user testing of your library's main discovery environment, with the goal of understanding how easy or hard it is for a researcher to find the available data collections.
26. Develop a portal page with a site map specifically for discovering collections at your institution available for computational use and related support services.
27. Begin tracking demand for and use of data in and about your collections.
28. For a collection that cannot be made available openly on the web, investigate if your organization is able to support mediated access to the data, such as through an offline or encrypted workstation.
29. Prepare and provide datasets that are intentionally useful, in terms of size and complexity, for teaching in semester- or quarter-long classes.
30. For classes that draw directly on library collections and generate data, ask the students to submit their data products back to library, through the institutional repository. Normalize the process of giving back and augmenting the collections with data. This may work particularly well for collections that are institutionally or regionally focused.
31. Identify a faculty member who does computational analysis for their own research and find a way to transfer or replicate the tools and approaches they use to apply them to a library collections-as-data use case.
32. If you offer an API to your repository, evaluate the public-facing documentation to see if it is clear, current, accurate, and discoverable by researchers.
33. Publish documentation about how to find, use, and interpret collections as data in multiple places including blogs, README files, and LibGuides.

34. A dataset should always be accompanied by a README plain text file that documents basic, important information about the data. Make READMEs part of your data documentation practice. Develop one or more template to that can be used by librarians and researchers.
35. Make an effort to make existing OCR output generated from past scanned text collections projects more available for computational analysis, such as through bulk download.
36. When planning your next digitization project, incorporate additional steps for preparing content files, OCR or transcription text, and metadata for bulk access. Document the key issues and decision points you encounter as you evolve and expand your digitization workflows.
37. Talk to colleagues involved in taking in deposits to your institutional repository or research data repository about a process for encouraging and accepting contributions back from users of data in your collections.
38. Gain the support of administration by following and supporting the work of third-party research groups like OCLC that help bolster and highlight the trends in the development of collections as data.
39. Provide a resource that shows a data user how to cite a dataset, and that shows a data creator how to format a preferred citation for an original dataset and a derivative dataset.
40. Ask a subject specialist at your institution if faculty or students are requesting data about or derived from library collections.
41. Take a public services librarian, curator, or archivist out for coffee to talk about collections as data. Ask what they are hearing from faculty, students, and other users of collections about computational use and which collections have potential for taking action to lower barriers to computational use.
42. Investigate how your library is collecting, managing, and making email archives accessible. Consider whether a collections as data approach will serve your institution's goals.
43. Start small. Start with a research question, and choose projects that have promise to be generalizable for use by future scholars such that the investment is worth the level of commitment. No one-offs!
44. Start with a prototype or proof of concept. It's fine if your Collections as Data project does not integrate with institutional repository or formalized infrastructure.
45. Collaborate with subject specialists or instruction librarians to ask scholars about interest in computational data in and about collections. Compile their ideas to make a case, and build a

team for the next opportunity to pursue one of them.

46. Be thoughtful and strategic about allocating scarce resources to collection digitization projects. Consider prioritizing projects that produce outcomes that are reusable (derivative datasets) and repeatable (processes, tools, workflows) that can benefit your department and your users again and again.
47. Explore what it would take for your organization to contribute subject data to Wikidata, drawing on a local collection and then incorporating the Wikidata links into your local discovery environment.
48. Test how data gathered in a crowdsourcing project can be associated with the existing source object data and can also serve as stand-alone dataset.
49. Use your favorite search engine to find information about APIs provided by museums and read about the various ways that data about museum collections can be analyzed to discover new insights.
50. Keep tabs on the projects emerging in the [Collections as Data: Part to Whole project](#), funded by the Mellon Foundation. They are bound to point a way forward for us all!

Appendix 5: Collections as Data Methods Profiles

CAD Methods Profiles are designed to help people who work in libraries, archives and museums gain a better understanding of common research methods that make use of cultural heritage collections for computational analysis. Of course, these descriptions are simplified versions of the methods, and are described mostly in the context of their implications for the creation, description, packaging, or distribution of collections as data. Profiles should be used in the context of the principles articulated in the Santa Barbara Statement on Collections as Data.

Text Mining

Laurie Allen and Scott Enderle, University of Pennsylvania

1. What is it?

Looking for patterns in text. Generally, text mining is done on a corpus of texts rather than a single text. Finding and assembling a corpus that is appropriate to the research needs of a project can be one of the trickiest and most time consuming things that a researcher does when approaching a project. There is not currently an agreed upon standard for describing or sharing text corpora, though there are a variety of guides to finding them, and vendors who sell access to text that researchers can assemble to create a corpus.

See a few definitions and links:

- Drucker, Johanna. Data Mining and Text Analysis - Introduction to Digital Humanities. Accessed August 27, 2018.
- Underwood, Ted. Seven Ways Humanists Are Using Computers to Understand Text. The Stone and the Shell (blog), June 4, 2015.

2. Who uses it?

Text mining is used across humanities disciplines (notably language and literature departments, and history) and in the social sciences, especially political science, communications, and business. There are also text corpora used in machine learning applications as well as linguistics. Disciplinary uses of text mining vary both in method of analysis, and, importantly, in the kinds of texts included in the corpus of study. For example, a corpus of the front page articles of current major newspapers might be valuable to a political scientist, while a scholar of 19th C. English novels might want a corpus of literary reviews.

3. What form of data is most useful for this method?

Generally, researchers doing text analysis will want to use plain text (i.e. machine readable, but without markup) in large quantities. They will also need accompanying metadata at a variety of scales. That is, sometimes they'll want metadata at the book/article level, or at the collection level, and for some uses, it is helpful to have chapter or section level metadata. In linguistic uses, analyses of texts sometimes include annotations down to the specific phoneme level, which make linguistic corpora less widely produced by libraries/archives/museums.

4. What might researchers explore when they're text mining?

They might look for word frequency counts (how often is a particular word used) at the page, article/chapter, or volume level, or use those counts for further analysis. For that reason, a dataset of frequency counts, even in the absence of fulltext, is often useful, especially in cases where the full content of a corpus can not be made available because of copyright restrictions.

Researchers often look for patterns in the data as they relate to features in the metadata (for example, how does the frequency of a word in texts change over time). Reliance on both the metadata about each text and the text themselves makes it important for researchers to know about large inconsistencies in the data or metadata quality. For example, if the OCR quality is inconsistent across a collection, it is very useful to include standard metadata about OCR quality for each text, if it is known. Or, if cataloging or metadata creation practices changed over time, those changes should be noted so that researchers can account for those changes in their analyses.

In some cases, people are interested in locations of words on pages (If an OCR program has included information about bounding boxes, it would be nice to have multiple versions – one with bounding boxes, and the other without).

5. Common tools used for text mining

Most people who do text mining are using scripting languages like Python or R.

Beyond that, there are a few other tools, useful for analysis and teaching like:

- [Voyant](#)
- [AntConc](#) - (See also Heather Froehlich's [AntConc lesson on Programming Historian](#))
- [Topic Modeling Tool](#)
- [Mallet](#)

6. Things to look out for when preparing collections for text mining

Copyright: This is a big one, for obvious reasons. Where fulltext can not be provided, some libraries provide wordcounts or other analytics about the texts.

Documentation of text and metadata: Multiple versions of texts can be a big source of frustration or confusion in text analysis. For example, a series of reports might have the same first page, which is duplicated across all reports. Flagging those kinds of duplications can be valuable in helping researches cut the preparation time to making a corpus usable.

7. Examples of this method in use

Underwood, Ted, David Bamman, and Sabrina Lee. "The Transformation of Gender in English-Language Fiction." *Journal of Cultural Analytics*, 2018. <https://doi.org/10.22148/16.019>.

Barron, Alexander T. J., Jenny Huang, Rebecca L. Spang, and Simon DeDeo. "Individuals, Institutions, and Innovation in the Debates of the French Revolution." *Proceedings of the National Academy of Sciences*, April 17, 2018, 201717729. <https://doi.org/10.1073/pnas.1717729115>.

8. Examples of collections optimized for this use

"Documenting the American South: DocSouth Data." Accessed August 27, 2018. <https://docsouth.unc.edu/docsouthdata/>.

Chronicling America: <https://chroniclingamerica.loc.gov/>

La Gaceta De La Habana: <https://merrick.library.miami.edu/cubanHeritage/cubanlaw/lagaceta.php>

Network Analysis

1. What is it?

Network analysis supports quantitative and qualitative study of relationships between entities. Entities can be people, places, or things. Network analysis is especially helpful for studying multiple levels of complex systems.

A few resources and links:

“Network Analysis: Lesson Directory.” *Programming Historian*.
<https://programminghistorian.org/en/lessons/?topic=network-analysis>

Easley, David, and Jon Kleinberg. *Networks, Crowds, and Markets: A Book by David Easley and Jon Kleinberg*. Accessed May 14, 2019.
<https://www.cs.cornell.edu/home/kleinber/networks-book/>.

Locke, Brandon. *Humanities Data Curation Record. Network Graphs and Network Analysis*. 2017. Reprint, Data Praxis, 2018. <https://github.com/datapraxis/hdcr>.

2. Who uses it?

Network analysis is used across a wide range of communities with some variation in terminology based on discipline. While social network analysis is popular, network analysis is also used to study physical infrastructure, e.g. transmission of energy through an electrical grid, or the flow of traffic. It can also be used for fictional characters in plots. In business network analysis it is used to study how organizations form, how money transfers from one place to another. It is also used, famously, in recommendation engines.

3. What form of data is most useful for it?

Researchers need relational information for network analysis, which can be found in many datasets. However, not all networks are useful for analysis, so there can be a fair amount of exploration in finding network datasets. The most basic forms of data for network analyses simply require that each record includes two entities and a relationship. For example, a simple spreadsheet with many rows and three columns. For each row: one person (entity) sent a letter (relationship) to another person (entity), or one publication (entity) was authored (relationship) by a person (entity). Other data can become part of network analysis as well, but the simplest notion of the network simply requires entities and relationships.

4. What data features might researchers explore?

After establishing whether network analysis is the right method, researchers might explore the size of a particular network, either by counting the number of nodes (entities) or number of edges (relationships). They might ask what is the percent of the network that is isolated from the rest? They may also look at network level measurements - who is most central, who are the most important conduits? What are the people places or things that have easiest access to outer bounds of network? They may look at the clustering coefficient – do relationships in the network tend to clump together or are they fairly diffuse?

5. Common Tools

Palladio <http://hdlab.stanford.edu/palladio/> (for very lightweight exploration of networks, designed for historical data)

Cytoscape <http://www.cytoscape.org/>

Gephi <https://gephi.org/>

NodeXL <https://www.smrfoundation.org/nodexl/>

Pajek <http://mrvar.fdv.uni-lj.si/pajek/>

6. Examples of this method in use

Warren, Christopher N., Daniel Shore, Jessica Otis, Lawrence Wang, Mike Finegold, and Cosma Shalizi. "Six Degrees of Francis Bacon: A Statistical Method for Reconstructing Large Historical Social Networks." *Digital Humanities Quarterly* 010, no. 3 (July 12, 2016).

Moravec, Michelle. "Network Analysis and Feminist Artists." *Art@s Bulletin* 6, no. 3 (November 30, 2017). <https://docs.lib.purdue.edu/artlas/vol6/iss3/5>.

White, Howard D., and Katherine W. McCain. "Visualizing a Discipline: An Author Co-Citation Analysis of Information Science, 1972–1995." *Journal of the American Society for Information Science* 49, no. 4 (1998): 327–55.

[https://doi.org/10.1002/\(SICI\)1097-4571\(19980401\)49:4<327::AID-ASIA>3.0.CO;2-4](https://doi.org/10.1002/(SICI)1097-4571(19980401)49:4<327::AID-ASIA>3.0.CO;2-4).

Bibliography of Historical Network Research <http://historicalnetworkresearch.org/bibliography/>

7. Examples of collections optimized for this use

The following sources provide directories of network data:

"CASOS Tools: Network Analysis Data | CASOS." <http://casos.cs.cmu.edu/tools/data2.php>.

"Index of Complex Networks." Index of Complex Networks. <http://icon.colorado.edu/>.

"Stanford Large Network Dataset Collection." <http://snap.stanford.edu/data/index.html>.

Visualization - Thomas interested in this, planning to try and chat with Lauren Klein

Appendix 6: National Forum Position Statements

March 2017

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 certainly informed the work of the forum, and consequently the iterative community based development of project outcomes.

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, "The 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-Forsen) 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 researchers 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 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);

- o 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;
- o 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 through 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 into 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.

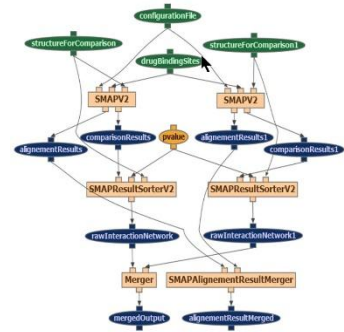
Drucker, Johanna. 2014. *Graphesis: Performative Approaches to Graphical Forms of Knowledge Production in the Humanities*. Cambridge: Harvard University Press.

Assessing Data Workflows for Common Data 'Moves' Across Disciplines

Alan Liu, University of California Santa Barbara

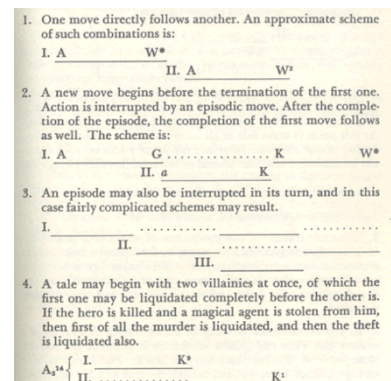
In considering how library collections can serve as data for a variety of data ingest, transformation, analysis, replication, presentation, and circulation purposes, it may be useful to compare examples of data workflows across disciplines to identify common data "moves" as well as points in the data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

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](https://doi.org/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 Neatrou, 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

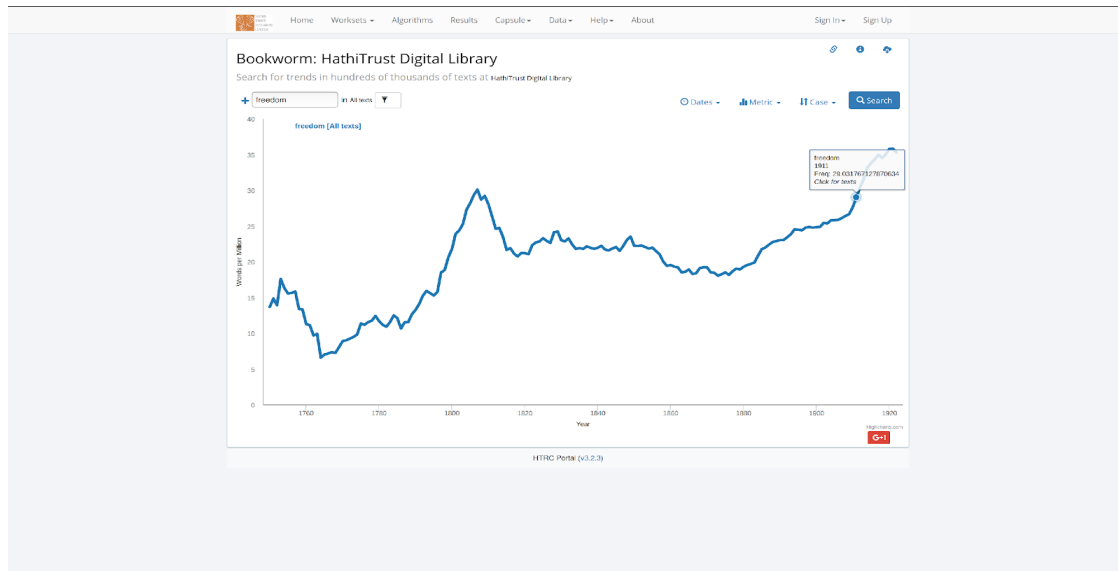
How can we integrate generations of high-quality, professionally-created metadata with electronic versions of the object itself? Particularly when copyright comes into play, we can't simply hope for openness; and there's a steep trade-off between the thoroughness of a well-thought-out standard and a 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.



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 for 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 world 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_Archives_Unleashed.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.



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]



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/>

Appendix 7: Forum Summaries

Forum 1: March 1-3, 2017 | Santa Barbara, California

The first forum was a gathering of key stakeholders, practitioners, thought leaders, and scholars currently working with collections as data. Each participant was asked to prepare a position statement in advance of the forum to help frame the discussion. Forum sessions were a mixture of group discussions, presentations, and small group work using human centered design techniques. Activities were designed to document current practice, surface problems, and generate new ideas and approaches for collections as data work. Although crafting a joint framework and strategic direction for collections as data was an initial goal of the forum, this was ultimately proved not to be achievable because of the multiplicity of techniques, approaches, and user needs for collections as data. Instead, forum participants crafted the Santa Barbara Statement, which represented a consolidation of the major themes of the forum. These included the complexity of the collections as data landscape, particularly the wide range of consumers and use cases; questions of scalability; open access solutions; ethical concerns; and partnerships.

Agenda

March 1

8:30 Breakfast

8:45 Welcome & Introductions

9:15 Project scope overview

Thomas Padilla

9:45 Project Outcomes--focused group discussion

10:15 Break

10:30 Collections as Data Panel -- Existing implementations

Miriam Posner (UCLA), Harriett Green (UIUC), Tim Sherratt (University of Camberra), Mia Ridge (British Library), Jefferson Bailey (Internet Archive), Gabrielle Foreman (University of Delaware)

11:45 Idea Generation

Discussion: *You each came with a set of collections as data related ideas, expressed in part through your position statements. You are in a group of people with a range of experiences. During this time we would like you to work to align your experiences to generate ideas that hold the potential to push collections as data work forward. We ask that you focus your discussion on enumerating as many ideas as possible. We do not expect you to create detailed roadmaps whereby these ideas might be pursued. This conversation is purely geared toward getting as many of our ideas to the surface as possible.*

12:00 Working Lunch

1:30 Sharing

2:00 Break

2:15 Play it Out
Discussion: *How might some of the ideas you generated be implemented?*

3:45 Break

4:00 Sharing

4:30 Reflection

Discussion: *This afternoon you spent time reflecting on your collective position statements and discussing ideas that push collections as data work forward. We now ask you to spend a few moments in your focused group to critique these all of the ideas that were generated. What do you think are particularly good or useful ideas? What might be easy to implement? What problems and pitfalls exist?*

5:00 Set Stage for Day 2

March 2

8:30 Breakfast

9:00 Gather Data

Activity: *In this exercise, you will rely on each other as a sort of “focus group” to gather a set of data about how you engage with collections as data. Please answer the following questions as a group, recording your answers in this document. You may choose which set of questions are most relevant to your perspective in creating/manipulating/consuming collections as data. Some groups may choose to answer from multiple perspectives. We will build upon this data the remainder of the day.*

10:00 Break

10:15 Story Generation

Activity: *Using the data gathered earlier this morning from all of the groups and your own personal experiences, write 2-3 user stories.*

11:45 Lunch

1:00 Story Review and Critique

Activity: *Examine and refine the use cases generated by another group.*

2:00 Prototyping

Activity: *Using the best or most interesting ideas from the story generation activity, design a product, system, service or curriculum, etc., that meets the needs of one or more people you choose from the stories. You will be presenting your prototype idea using a Concept Poster. Be sure to consider the effect of your solution on other stakeholders to demonstrate viability and impact.*

3:00 Break

3:15 Share prototypes

4:00 Discussion: Implications for Libraries

5:00 Review of Day

March 3

- 8:30** Breakfast
9:00 Discussion: Absences
9:30 Statement Creation
10:45 Break
11:00 Engagement
11:45 Closing remarks

Attendees

John Ajao

*Director, Systems and Repository Operations
University of California Santa Barbara*

Jefferson Bailey

*Head of Web Archiving Programs
Internet Archive*

Alex Chassanoff

*Software Curation Postdoctoral Fellow
Massachusetts Institute of Technology*

Tanya Clement

*Assistant Professor of Information
University of Texas Austin*

P. Gabrielle Foreman

*Professor of English and Black American Studies
University of Delaware*

Daniel Fowler

*Developer Advocate
Open Knowledge Foundation*

Harriett Green

*English and Digital Humanities Librarian
University of Illinois at Urbana Champaign*

Jennifer Guiliano

*Assistant Professor of History
Indiana University-Purdue University Indianapolis*

Matthew Miller

*Head of Semantic Applications & Data Research
New York Public Library*

Anna Neatrou

*Metadata Librarian
University of Utah*

Miriam Posner

*Digital Humanities Coordinator
University of California Los Angeles*

Sheila Rabun

*Community and Communications Officer
Stanford University*

Mia Ridge

*Digital Curator
British Library*

Laila Sakr

*Assistant Professor of Film and Media Studies
University of California Santa Barbara*

Ben Schmidt

*Assistant Professor of History
Northeastern University*

David Seubert

*Curator of Performing Arts Collections
University of California Santa Barbara*

Julie Hardesty
Metadata Analyst
Indiana University

Tim Sherratt
Associate Professor of Digital Heritage
University of Canberra

Christina Harlow
Metadata Librarian
Cornell University

Hannah Skates Kettler
Digital Humanities Librarian
University of Iowa

Greg Jansen
Research Software Architect
University of Maryland

Timothy St. Onge
Cartographer
Library of Congress

Lisa Johnston
Research Data Management/Curation Lead &
Co-Director University Digital Conservancy
University of Minnesota

Santi Thompson
Head of Digital Research Services
University of Houston

Matthew Lincoln
Data Research Specialist
Getty Research Institute

Kate Zwaard
Head of National Digital Initiatives
Library of Congress

Alan Liu
Distinguished Professor of English
University of California Santa Barbara

Forum 2: May 7-8, 2018 | Las Vegas, Nevada

After spending a year at conferences, workshops, and seminars talking about what collections as data is, we held a second national forum focused the nuts and bolts of collections as data work, particularly how communities interested in getting started with collections as data work could move forward. The first day of the forum focused on current implementations and how a variety of consumers, from librarians to scholars to the general public, interacted with collections as data resources. This section of the forum was livestreamed and received over 400 live and subsequent views. As in the first forum, the variety of these collections as data implementations once again demonstrated that the collections as data landscape is complex and no one set of solutions will be feasible or even appropriate for everyone. Forum participants then focused on reality checks of *Always Already Computational* deliverables based on their own experiences with collections as data.

Agenda

Monday, May 7

8:30 Breakfast

8:45 Dean Welcome & Introductions

9:00 Project Update

Thomas Padilla

9:30 Panel 1: Who is Collections as Data for?

Who is Collections as Data for? Building on Principle 5, the forthcoming version of the [CaD Santa Barbara Statement](#) will assert that "Collections as data designed for everyone serve no one."

How has your work with CaD been forged around specific people, whether those represented in the collections, built into the design of the dataset, or reflected in your own teaching and/or learning? What work have you done to match CaD with populations?

Dot Porter (UPenn), Shawn Averkamp (NYPL), Bergis Jules (UC Riverside)

10:30-10:45 Break

10:45 Panel 2: What is the coolest thing about your Collections as Data work?

What is the coolest thing about your collections as data work? Tell us why you became involved with this work and what motivates your continued dedication or interest. We'd like to show our attendees the spirit and possibilities of collections as data work.

Micki Kaufman (CUNY), Inna Kouper (Indiana), Greg Cram (NYPL), Laurie Allen (UPenn)

11:45-12 Break

12:00 Panel 3: How have you implemented Collections as Data?

Viewers of our livestream are likely interested in how they might participate in or grow collections as data. How have you started, shifted, or institutionalized collections as data? How do you see this work aligning with your institutional/organizational mission? What surprised you about the process, and what do you plan or hope to do next?

*Meghan Ferriter (LOC), Mary Elings (UC Berkeley), Helen Bailey (MIT),
Veronica Ikeshoji-Orlati (Vanderbilt)*

1:00 Lunch

2:00 Introducing the Guide

2:30 Reality Check on Project Deliverables -- group-based discussion and activities

All

5:00 Break for dinner - *On your own*

Tuesday, May 8

8:30 Breakfast

9:00 Future Directions: Moving Stuff Forward -- group-based discussion

All

11:45 Wrap up

Thomas Padilla

12:00 End - *Box lunch provided*

Attendees

Elvia Arroyo Ramirez

*Assistant University Archivist
University of California Irvine*

Shawn Averkamp

*Manager of Metadata Services
New York Public Library*

Helen Bailey

*Engagement Data Engineer
Massachusetts Institute of Technology*

Alex Chassanoff

*CLIR/DLF Postdoctoral Fellow in Software
Curation
Massachusetts Institute of Technology*

Kalani Craig

*Clinical Assistant Professor, Department of
History
Co-Director, Institute for Digital Arts &
Humanities
Indiana University*

Greg Cram

*Associate Director of Copyright and Information
Policy
New York Public Library*

Mary Elings

*Head of Technical Services
The Bancroft Library, University of California
Berkeley*

Meghan Ferriter

*Senior Innovation Specialist
Library of Congress*

Micki Kaufman

City University of New York

Inna Kouper

*Assistant Scientist, School of Informatics,
Computer, and Engineering
Assistant Director, Data to Insight Center
Indiana University*

María Matienzo

*Collaboration & Interoperability Architect
Stanford University*

Jake Orlowitz

*Head of The Wikipedia Library
Wikimedia Foundation*

Sarah Patterson

*Lecturer, Department of English
University of Massachusetts Amherst
Co-Founder, Colored Conventions*

Dot Porter

*Curator of Digital Research Services University of
Pennsylvania*

Chaitra Powell

*African American Collections and Outreach
Archivist
University of North Carolina Chapel Hill*

Chela Scott Weber

*Director of Library and Collections
California Historical Society*

Devin Higgins

*Digital Library Programmer
Michigan State University*

Hannah Scates Kettler

*Digital Humanities Research and Instruction
Librarian
University of Iowa*

Veronica Ikeshoji-Orlati

*CLIR Postdoctoral Fellow
Vanderbilt University*

Laura Wrubel

*Software Development Librarian
George Washington University*

Bergis Jules

*University and Political Papers Archivist
University of California Riverside*

Appendix 8: Conference engagements, 2017-2018

Conferences as a way to expand conversation beyond the two national forums. Limited money, chose to spend it by hosting mini-forums with user groups not at the national forum. More gathering of use cases and critique of our assumptions. Re-emphasized diversity of experience, capacity, and needs.

2017.

LDCX (March 27-29, Stanford, California)

- Thomas Padilla, Hannah Frost, “Supporting End User Computation / Use of Collections” (1 hour unconference session)

Csvconf (May 2-3, Portland, Oregon)

- Laurie Allen (keynote)

Texas Conference on Digital Libraries (May 23-25, Austin, Texas)

- Sarah Potvin, “Almost Already Computational: An Update from the Library Collections as Data Effort” (poster)

Association of College and Research Libraries Digital Humanities Interest Group webinar (June 26, online)

- Thomas Padilla, “What Does it Mean: Library Collections as Data” (3 speakers, 60 minute panel)

American Library Association (June 26, Chicago, Illinois)

- Laurie Allen, “New Kinds of Collections: New Kinds of Collaborations,” on panel for “Creating the Future of Digital Scholarship Together: Collaboration from Within Your Library” (3 projects, 90 minute panel)

Society of American Archivists (July 23-29, Portland, Oregon)

- Alexandra Chassanoff, Thomas Padilla, and Elizabeth Russey Roke, “Open Forum - Always Already Computational: Collections as Data” (75 minutes)

Digital Humanities (August 7, Montreal, Quebec)

- Sarah Potvin, Thomas Padilla, Laurie Allen, Stewart Varner, “Shaping Humanities Data” (full-day preconference symposium)

DLF eResearch Network (August 9)

- Thomas Padilla, “Collections as Data” (60 minutes)

Digital Library Federation (October 23-25, Pittsburgh, Pennsylvania)

- Thomas Padilla, “Collections as Data: An Update” (7 minutes)

- Thomas Padilla, Laurie Allen, Stewart Varner, Elizabeth Russey Roke, Hannah Frost, Sarah Potvin, “Collections as Data Workshop” (2 hours)

Samvera Connect (November 9, Salt Lake City, Utah)

- Hannah Frost, “Collections as Data and Samvera” (50 minutes)

Coalition for Networked Information (December 11, Washington, DC)

- Thomas Padilla, Laurie Allen, Hannah Frost, “Always Already Computational: Collections as Data” (1 hour)

2018.

American Historical Association (January 4, Washington, DC)

- Laurie Allen, Stewart Varner, “Collections as Data,” in workshop on “Getting Started in Digital History 2018” (4 hour workshop)

National Institute for Computer-Assisted Reporting (March 10, Chicago, Illinois)

- Thomas Padilla and Laurie Allen, “Cultural heritage data? Computational use, needs, and opportunities” (60 minutes)

Digital Public Library of America Annual Members Meeting (March 13-14, Atlanta, Georgia)

- Elizabeth Russey Roke, “DPLA as Data: Collections as Data in Practice” (90 minute workshop)

LDCX (March 26-28, Stanford, California)

- Hannah Frost and Kate Lynch, “Collections as Data” (60 minutes)

Los Angeles Arts Datathon (April 27, Los Angeles, California)

- Thomas Padilla, “Collections as Data x Arts as Data” (keynote)

DH + Libraries, Sidney Harman Center for Polymathic Studies, University of Southern California (April 30, Los Angeles, California)

- Thomas Padilla, “On a Collections as Data Imperative” (60 minutes)

Society of American Archivists (August 12-18, Washington, DC)

- Elizabeth Russey Roke, “Collections as Data,” Electronic Records Section Meeting (45 minute discussion)

Open Repositories (June 4-7, Bozeman, Montana)

- Hannah Frost and Sarah Potvin (Moderators), Mark Jordan, Katherine Lynch, Helen Bailey, “Enabling Computational Access at Scale: Are Repositories Serving Collections-as-Data?” (45 minutes)

HILT (June 4-8, Philadelphia, Pennsylvania)

- Thomas Padilla and Mia Ridge, “Collections as Data” (Week long workshop)

Dariah Beyond Europe Workshop at Library of Congress (October 3-4, Washington DC)

- Laurie Allen, Stewart Varner “Collections As Data: Digital Collections for Emerging Research Methods.” (keynote + workshop 2 hours)

Digital Library Federation (October 15-17, Las Vegas, Nevada)

- Thomas Padilla, Stewart Varner, Hannah Frost, Elizabeth Russey Roke, Sarah Potvin, “Always Already Computational, Never Quite Automatic: Towards a Collections as Data Framework” (55 minutes)
- Sarah Potvin, Thomas Padilla, Santi Thompson, Liz Woolcott, Amanda Rust, Giordana Mecagni, “What would the ‘community’ think? Three grant-funded team reflect on defining community and models of engagement” (55 minutes)

Appendix 9: Digital Humanities 2017 preconference: Shaping Humanities Data

Description

How can cultural heritage institutions develop and provide access to collections that are more readily amenable to computational use? How does a movement toward thinking about collections as data prompt an opportunity to reframe, enrich, and/or contextualize collections in a manner that expands use while avoiding replication of bias inherent in collection practice? The **Collections as Data** project presents **Shaping Humanities Data** as a venue to explore these questions at [Digital Humanities 2017](#). Shaping Humanities Data features eleven talks and five demonstrations. Talks and demonstrations were solicited through a [CFP](#) and reviewed by an international program committee. The event also includes opportunities for discussion and workshopping [Collections as Data](#) frameworks. The workshop will inform the development of recommendations that aim to support cultural heritage collections as data efforts.

Schedule

August 7, 2017

9:30-10:00

- Introductions, Schedule, Project Update

10:00-10:50

- Reusable Computational Processing of Large-Scale Digital Humanities Collections (Marciano and Jansen)
- MARCing the Boundary: Reusing Special Collections Records through the Early Novels Database (Kashyap and Van Tine)
- Leveraging Core Data for the Cultural Heritage of the Medieval Middle East (Schwartz)

11:00-12:00

- Lessons learned through the Smelly London project (Leem)
- Historical Public Health Data Curation: Indiana State Board of Health Monthly Bulletin Project (Pollock and Coates)
- Javanese Theatre as Data (Varela)
- High Performance Computing for Photogrammetry Made Easy (Dombrowski, Gniady, Simpson, Meredith-Lobay)

1:00-1:45

- Using IIRF to answer the Data Needs of Digital Humanists (Di Cresce)
- Demonstrating A Multidisciplinary Collections API (Almas and Baumgardt)

1:45-2:45

- Collections as Data Workshopping

3:00-3:50

- Umbra Search as Data: A digital sandbox to cross the digital divide (Marcus)

- Audio Analysis for Spoken Text Collections (Clement and McLaughlin)

4:00-4:50

- Facilitating Global Historical Research on the Semantic Web: MEDEA (Tomasek and Vogeler)
- Mending the Vendor: Correction and Exploratory Augmentation of Collections as Data (Locke)
- Learning through Use: A case study on setting up a research fellowship to learn more about how one of our collections works as computationally amenable data (Severson and Vejvoda)
- Addressing Copyright and IP Concerns when using Text Collections as Data (Sensenev, Dickson, and Tracy)
- Libraries as Publishers of a New Bibliographical Unit (Claeyssens)

4:50-5:30

- Wrap-Up

Program Committee members

Harriett Green, University of Illinois at Urbana Champaign

Inna Kizhner, Siberian Federal University

Alberto Martinez, Colegio de México

Ian Milligan, University of Waterloo

Gimena Del Rio Riande, Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET)- University of Buenos Aires

Laurent Romary, Inria and DARIAH

Henriette Roued-Cunliffe, University of Copenhagen

Melissa Terras, University College London

Presentation abstracts

Demonstrating A Multidisciplinary Collections API

Bridget Almas and Frederik Baumgardt, Tufts University; Tobias Weigel, DKRZ; Thomas Zastrow, MPCDF
The Collections Working Group of the Research Data Alliance (RDA) is a multidisciplinary effort to develop a cross-community approach to building, maintaining and sharing machine-actionable collections of data objects. We have developed an abstract data model for collections and an API that can be implemented by existing collection solutions. Our goal is to facilitate cross-collection interoperability and the development of common tools and services for sharing and expanding data collections within and across disciplines, and within and across repository boundaries. The RDA Collections API supports Create/Read/Update/Delete/List (CRUD/L) operations. It also supports set-based operations for Collections, such as finding matches on like items, finding the intersection and union of two collections, and flattening recursive collections. Individual API implementations can declare, via a standard set of capabilities, the operations available for their collections. The Perseids

Project at Tufts University is implementing this API for its collection of annotations on ancient texts. We will review the model and the functionality of the API and demonstrate how we have applied it to manage Perseids humanities data. We will also provide examples of how it is being applied for collections of data in other disciplines, including Climate Computing and Geoscience. Finally, we will solicit feedback from the participants in the workshop on the API and model and its applicability for other collections of cultural heritage data.

Libraries as Publishers of a New Bibliographical Unit

Steven Claeysens, Koninklijke Bibliotheek

Large-scale digitisation of historical paper publications is turning libraries into publishers of data collections for machines and algorithms to read. Therefore the library should critically (re)consider 1) its new function as a publisher of 2) a new type of bibliographical content in 3) an exclusively digital environment. What does it mean to be both library and publisher? What is the effect of remediating our textual and audiovisual heritage, not as traditional bibliographic publications, but as data and datasets? How can we best serve our patrons, new and old, machines and humans?

In my talk I want to address these questions drawing on my background as a book historian specialized in Publishing Studies, and on my experience as the Curator of Digital Collections at the national library of the Netherlands (KB) responsible for providing researchers with access (Data Services) to the large collections of data the KB is creating.

At the KB we found there is no one-way solution to cater the needs of Digital Humanists. I will reflect upon their requirements by analysing the requests for data by Digital Humanists the KB received during the year 2016. What kind of data were they looking for? Why did they need the data?

I will identify both valuable as well as incompatible user requirements, indicating the conflicting expectations and interests of different disciplines and researchers. Therefore I argue that 1) a close collaboration between scholars and librarians is essential if we really want to advance the use of large digital libraries in the field of Digital Humanities, and 2) we need to carefully reconsider our role(s) as a library.

Audio Analysis for Spoken Text Collections

Tanya Clement and Steve McLaughlin, The University of Texas at Austin

At this time, even though we have digitized hundreds of thousands of hours of culturally significant audio artifacts and have developed increasingly sophisticated systems for computational analysis of sound, there is very little provision for audio analysis. There is little provision for scholars interested in spoken texts such as speeches, stories, and poetry to use or to even begin to understand how to use high performance technologies for analyzing sound. Toward these ends, we have developed a beginner's audio analysis workshop as part of the HiPSTAS (High Performance Sound Technologies for Access and Scholarship) project. We introduce participants to essential issues that DH scholars, who are often more familiar with working with text, must face in understanding the nature of audio texts such as poetry

readings, oral histories, speeches, and radio programs. First, we discuss the kinds of research questions that humanities scholars may want to explore using features extracted from audio collections—laughter, silence, applause, emotions, technical artifacts, or examples of individual speakers, languages, and dialects as well as patterns of tempo and rhythm, pitch, timbre, and dynamic range. We will also introduce participants to techniques in advanced computational analysis such as annotation, classification, and visualization, using tools such as Sonic Visualiser, ARLO, and pyAudioAnalysis. We will then walk through a sample workflow for audio machine learning. This workflow includes developing a tractable machine-learning problem, creating and labeling audio segments, running machine learning queries, and validating results. As a result of the workshop, participants will be able to develop potential use cases for which they might use advanced technologies to augment their research on sound, and, in the process, they will also be introduced to the possibilities of sharing workflows for enabling such scholarship with archival sound recordings at their home institutions.

Using IIF to answer the Data Needs of Digital Humanists

Rachel Di Cresce, University of Toronto

How can we provide researchers and instructors with seamless access to dispersed collections, controlled by their formats, frameworks and softwares, across cultural heritage organizations? How can we allow free movement of this data so it can be analyzed, measured and presented through different lenses? And how can we support this research without placing too high a technical burden on those institutions, especially those with limited resources? These questions have been at the centre of the University of Toronto's Mellon-funded project, Digital Tools for Manuscript Study, which aims at integrating the International Image Interoperability Framework (IIIF), based on Linked Data principles, with existing tools to improve the researcher's experience. Essentially, the project shifts focus away from the tool that makes use of the data onto the data itself as a research and teaching tool.

At the core of the project is working with humanists to understand how they conduct their research and what they need in order to do digital scholarship effectively. We identified, for example, strong needs for data portability, repository interoperability, and tool modularity in scholarly work. We make use of the IIIF data standard to support data portability, the Mirador image viewer for its suite of tools for image presentation and analysis and Omeka for its wide adoption among digital humanities scholars and cultural heritage organizations. In addition, we have developed a standalone tool set called IIIF To Go. This is a user-friendly IIIF start-up kit, designed to support both research and pedagogical uses. This talk will discuss our attempt to democratize an international standard by (1) embedding it in tools with wide traction and low entry barriers in the digital humanities and manuscript studies community (2) limiting the technical load required to make use of the standard and tools for instruction and research and (3) looking toward Linked Data at GLAM institutions.

High Performance Computing for Photogrammetry Made Easy

Quinn Dombrowski, University of California Berkeley; Tassie Gniady, Indiana University; John Simpson, Compute Canada; Megan Meredith-Lobay, University of British Columbia

Photogrammetry (generating 3D models from a series of partially-overlapping 2D images) is quickly gaining favor as an efficient way to develop models of everything from small artifacts that fit in a light box to large archaeological sites, using drone photography. Stitching photographs together, generating point clouds, and generating the dense mesh that underlies a final model are all computationally-intensive processes that can take up to tens of hours for a small object to weeks for a landscape to be stitched on a high-powered desktop. Using a high-performance compute cluster can reduce the computation time to about ten hours for human-sized statues and twenty-four hours for small landscapes.

One disadvantage of doing photogrammetry on an HPC cluster is that it requires use of the command line and Photoscan's Python API. Since it is not reasonable to expect that all, or even most, scholars who would benefit from photogrammetry are proficient with Python, UC Berkeley has developed a Jupyter notebook that walks through the steps of the photogrammetry process, with opportunities for users to configure the settings along the way. Jupyter notebooks embed documentation along with code, and can serve both as a resource tool for researchers who are learning Python, and as a stand-alone utility for those who want to simply run the code, rather than write it. This offloads the processing to the HPC cluster, allowing users to continue to work on a computer that might normally be tied up by the processing demands of photogrammetry.

MARCIing the Boundary: Reusing Special Collections Records through the Early Novels Database

Nabil Kashyap, Swarthmore College, and Lindsay Van Tine, University of Pennsylvania

In this presentation, Early Novels Database project (END) collaborators Nabil Kashyap and Lindsay Van Tine will offer perspectives on the possibilities and perils of reframing the special collections catalog as a collaborative datastore for humanities research. Among other activities, the END project includes curating records from regional special collections, developing standards for enhancing catalog records with copy-specific descriptive bibliography, and publishing open access datasets plus documentation. Work on END therefore excavates basic questions around what thinking through library holdings as data might actually entail. What ultimately constitutes "the data"? What do they do? For whom? Starting from Leigh Star's notion of the boundary object, this presentation explores the theory and praxis of MARC as a structure of knowledge that can allow "coordination without consensus."

The MARC records at the core of the END dataset, the result of meticulous work on the part of institutional catalogers, serve as "boundary objects"—that is, they serve as a flexible technology that both adapts to and coordinates a range of contexts. These contexts, in turn, can have very different needs and values, from veteran catalogers to undergraduate interns, special collections to open source repositories, and from projected to actual uptake and reuse of the data in classrooms and research.

These shifting contexts call into question just what the "data" is. It will look different to a cataloger, an outside funding organization, a sophomore, a programmer, or an 18th c. scholar. What might appear straightforward—creating derivatives, for example—instead reveals a host of issues. Transforming nested into tabular data brings to light frictions between disparate assumptions as to the unit of study, whether a work or volume or a particular copy. Privileging certain fields either effaces the specificity of

transcription or sacrifices discoverability. There is no transparent “data dump”; instead, every act of transformation reinscribes a set of disciplinary and institutional values. Viewing collections as data is as much about opening up data as about actively demonstrating and to an extent prescribing research possibilities.

Lessons learned through the Smelly London project

Deborah Leem, Wellcome Trust and University College London

I propose to present the intended aims of the Smelly London project; what we achieved; challenges we experienced working with digitised collections; and possible directions for further development. In order to increase the impact and value that cultural heritage digital collections can offer we believe that their online collections and platforms should be more amenable to emerging technologies and facilitate a new kind of research.

Wellcome Library – part of Wellcome – is one of the world’s major resources for the study of health and histories. Over the past few years Wellcome have been developing a world-class digital library by digitising a substantial proportion of their holdings. As part of this effort, approximately 5,500 Medical Officer of Health (MOH) reports for London spanning from 1848-1972 were digitised in 2012. Since September 2016 Wellcome have been digitising 70,000 more reports covering the rest of the United Kingdom (UK) as part of UK Medical Heritage Library (UKMHL) project in partnership with [Jisc](#) and the Internet Archive. However, no digital techniques have yet been applied successfully to add value to this very rich resource.

As part of the [Smelly London](#) project, the OCR-ed text of the MOH London reports has been text-mined. Through text mining we produced a geo-referenced dataset containing smell types for visualisation to explore the data. At the end of the Smelly London project the MOH smell data will also be available via other platforms and this will allow the public and other researchers to compare smells in London from the 19th century to present day. This has the further potential benefit of engaging with the public. However, cultural heritage organisation do not offer platforms that can help researchers share or communicate the data derived from digital collection use.

Mending the Vendor: Correction and Exploratory Augmentation of Collections as Data

Brandon Locke, Michigan State University

Like many university libraries, Michigan State received external hard drives filled with collections they held perpetual licenses to. Like many university libraries, those collections have mostly remained mostly unused since they’ve been acquired. The data required processing to make them usable, but without demand for specific data from scholars, there was little benefit or reason to make all of the data available.

In an effort to pilot a project to make this data more available and to promote use of the datasets, Brandon Locke (Director of LEADR), Devin Higgins (Library Programmer), and Megan Kudzia (Digital Scholarship Technology Librarian), embarked on a project to make the papers of Fannie Lou Hamer available for download. Hamer’s papers were chosen based on her historical stature and interest to

faculty and graduate students in the Department of History, and upon the relatively small size of the collection.

The original scope of the project was for Higgins and Kudzia to make the plain text files available for download by any MSU student, faculty and staff. LEADR staff would then experiment with different text and data mining tools to add metadata and create subsets and auxiliary datasets to accompany the collection.

After Higgins and Kudzia made the plain text files available to the campus community, the LEADR staff immediately encountered troubles with Named Entity Recognition. Upon inspection, the OCR on the files were far too flawed for any accurate text mining, and the entire collection had to be redone using the provided page images with close training and manual correction.

This talk will detail some of the shortcomings in the supplied data, discuss opportunities for experimental text and data mining to enhance and augment existing collections datasets, and engage in opportunities for collaborations between institutions in improving data quality.

Reusable Computational Processing of Large-scale Digital Humanities Collections

Richard Marciano and Greg Jansen, University of Maryland

The Digital Curation Innovation Center (DCIC) at the U. Maryland iSchool, officially launched the “DRAS-TIC” archiving platform at iPRES 2016, in Oct. 2016. This stands for Digital Repository At Scale That Invites Computation [To improve Collections], and is rolled out under a community-based open source license. The goal is to build out an open source platform into a horizontally scalable archives framework serving the national library, archives, and scientific management communities. As a potential scalable and computational platform for Big Data management in large organizations in the cultural heritage, business, and scientific research communities.

This digital repository framework can scale to over a billion records and has tools for advanced metadata extraction - including from images, file format conversion, and search within the records and across collections. The underlying software is based on the distributed NoSQL database, Apache Cassandra, created to meet the scaling needs of companies like Facebook. DRAS-TIC supports integration by providing a standard RESTful Cloud Data Management Interface (CDMI), a command-line interface, web interface, and messaging as contents are changed (MQTT). We are now exploring connecting DRAS-TIC with a graph database engine to support social network analysis and computing of archival and library collections.

We wish to demonstrate this environment with reusable clustering workflows for grouping digitized forms by their layout, a recurring use-case in many digital humanities projects. This is a preprocessing step that has the potential to lead to more accurate OCR of regions in images within digitized forms.

Umbra Search as Data: A digital sandbox to cross the digital divide

Cecily Marcus, University of Minnesota Libraries

Publicly launched in 2017, the University of Minnesota Libraries' [Umbra Search African American History](#) has been working with partners across the country—from the Digital Public Library of America to Yale University to Howard University—to facilitate digital access to African American cultural history. As more than a search tool, Umbra Search doesn't just bring together over 500,000 digital materials from 1,000 US libraries, archives and museums. It also promotes the use of these materials through programming with students, educators, scholars, and artists, and leads a massive digitization effort of African American materials to build out a national digital corpus of African American history. Now, Umbra Search is exploring what it means to share the Umbra Search digital corpus as a data set that helps to bridge the digital divide and promote digital literacy among underrepresented youth and kids of color. By packaging curated sets of Umbra Search data around thematic topics (as well as providing access to the whole of Umbra Search data) with accessible digital storytelling tools that allow students to make data their own, Umbra Search provides an introduction to digital storytelling and other digital humanities skills through the lens of African American history and culture. Umbra Search's national digital corpus provides a unique opportunity to engage students with STEAM activities and skill building with culturally relevant content that affirms African American history and culture. This talk discusses the rationale for developing a digital sandbox that provides libraries with a new model for activating primary source materials and digital collections—often considered to be among the more rarefied and inaccessible collections in libraries—and digital humanities tools in communities that may not regularly engage with archives, primary source digital collections, or digital humanities.

Historical Public Health Data Curation: Indiana State Board of Health Monthly Bulletin Project

Caitlin Pollock and Heather Coates, Indiana University-Purdue University Indianapolis

As digital scholarship librarians, enhancing open digital content to facilitate reuse is a key mission of our work. This talk will introduce the work of IUPUI librarians in curating the Indiana State Board of Health Monthly Bulletin (1899-1964). While in circulation, this resource was sent to all health officers and deputies in the state, plus individual subscribers. Physicians shared information about health and wellness, communicable diseases, patent medicines, food safety, and many other topics. As such, the Bulletin provides a unique historic portrayal of Indiana public health practice, fascinating images, and regular vital statistics from the early and mid-20th century. This project brings together the Ruth Lilly Medical Library and the IUPUI University Library to leverage librarian expertise in digital humanities, medical humanities, public health, the history of medicine, and data curation. Our initial focus is curating a 10-year span (1905-1914) of these bulletins in order to develop and refine processes that can be adapted for other digital collections. Our curation efforts focus on providing greater accessibility to students and scholars of Indiana and medical history, public health, and Hoosiers across the state. We are creating three types of products: TEI documents; geocoded citizens and professionals, community organizations and businesses, and buildings; and vital statistics data. Data dictionaries are being developed to support analysis of the vital statistics and to capture additional context about historic knowledge of disease and death. Project documentation will be developed to support exploration by the public and use by scholars and provide transparency with regards to the decisions made during curation. All products generated from the project, including protocols for curation, will be shared openly under a CC-BY license on platforms including Github and the TEI Archiving, Publishing and Access Service (TAPAS) Project.

Leveraging Core Data for the Cultural Heritage of the Medieval Middle East

Daniel L. Schwartz, Texas A&M University

I direct Syriaca.org, a core data project for Syriac history, literature, and cultures. Syriac is a dialect of Aramaic once spoken by populations across the Middle East and Asia. Syriac sources document key moments in the interaction of Judaism, Christianity, and Islam and offer unique perspectives on the history of the Middle East from the Roman period through Ottoman rule and into the tumultuous present in Iraq, Syria, and the Levant. Syriaca.org has built a core data infrastructure useful to any digital project in the field that is interested in incorporating our URIs for persons, places, works, manuscripts, etc. I would like to propose a 30-minute demonstration of three projects that highlight this utility. 1) SPEAR (Syriac Persons, Events, and Relations) is a digital prosopography that employs our core data model (URIs) to extract and encode data about persons, events, and relationships from primary source texts. The scale enabled by the digital allows extensive treatment of many subaltern groups usually left out of traditional print prosopography. TEI encoding and serialization into RDF allow for multiple ways to query and visualize this data. 2) The New Handbook of Syriac Literature is an open-access digital publication that will serve as both an authority file for Syriac works and a guide to accessing their manuscript representations, editions, and translations in digital and analog formats. Though still in development, this Handbook will more than double the number of works contained in the last publication to attempt something similar, Anton Baumstark's *Geschichte*, which is over 90 years old. The Handbook is part of Syriaca.org's efforts to produce reference resources that help overcome the colonial biases that informed Orientalist organization of the cultural heritage of the Medieval Middle East. 3) We are developing a URI resolver that any project in the field using our URIs can incorporate into their website to show users how many and what types of resources Syriaca.org has on the entities included in their data and to provide direct links to those resources.

Addressing Copyright and IP Concerns when using Text Collections as Data

Megan Senseney, Eleanor Dickson, and Daniel G. Tracy, University of Illinois

Open source text data mining tools such as Voyant and publicly-available services such as the HathiTrust Research Center (HTRC) have brought the potential of new research discoveries through computational analytics within reach of scholars. While the tools for mining and analyzing the contents of digital libraries as data are increasingly accessible, the texts themselves are frequently protected by copyright or other IP rights, or are subject to license agreements that limit access and use.

The HTRC recently convened a task force charged to draft an actionable, definitional policy for so-called non-consumptive use, which is research use that permits computational analysis while precluding human reading. This year, the HTRC released the task force's Non-Consumptive Research Policy, which is shaping revised terms of service and tool development within the HTRC. Building on the development of the HTRC's policy, our team is seeking to catalyze a broader discussion around data mining research using in-copyright and limited-access text datasets through an IMLS-funded national forum that will

bring together experts around issues associated with methods, practice, policy, security, and replicability in research that incorporates text datasets that are subject to intellectual property (IP) rights. The national forum aims to produce an action framework for libraries with recommendations that will include models for working with content providers to facilitate researcher access to text datasets and models for hosting and preserving the outputs of scholars' text data mining research in institutional repositories and databanks.

This short talk will describe the task force's work to establish a Non-Consumptive Research Policy for the HTRC and outline next steps toward building a more comprehensive research agenda for library-led access to the wealth of textual content existing just out-of-reach in digital collections and databases through the upcoming national forum.

Learning through use: A case study on setting up a research fellowship to learn more about how one of our collections works as computationally amenable dataset

Sarah Severson and Berenica Vejvoda, McGill University Library and Archives

McGill University Library and Archives recently completed a major project to retrospectively digitize all of the dissertations and theses in the our collections. Once these were added to the institutional repository, the metadata and full text of over 40,000 electronic theses and dissertations (ETD), from 1881 - present, became searchable using the traditional database structure of keywords and full text. With such a large and comprehensive corpus of student scholarship, we wanted to use this collection as our first foray into thinking about 'collections as data' and what kinds of research could be done if we opened up the entire raw, text corpus.

In order to encourage use and dialogue with the collection, the Library created a Computational Research Fellowship through an innovation fund. The fellowship call was left deliberately open in order to learn what people wanted to do with the collection and the only condition was that they share what they learned openly through presentations about their work and host any code in an open environment such as GitHub.

The selected fellow project will specifically utilize Python's Natural Language Toolkit and capitalize on using word2vec (a word embedding algorithm developed by Google), to build an application with a front-end, web-based interface that will allow researchers to examine how literary terms have changed over time in terms of usage and context. The project will also include a data visualization component using Plotly (a Python library) to promote interactive and visually meaningful data displays. More concretely, researchers will be able to enter a concept and a time-period of interest and visualize how the context of the concept has evolved over time. By way of example, the concept of "woman" shifts contextually between First-wave feminism and prior, as well as through subsequent waves of feminism. This presentation will look at how we are thinking of our ETD collection as a computationally amenable dataset; the computational fellowship as a means of engagement; and, what we hope to learn about the collection and future library text mining services and support.

Facilitating Global Historical Research on the Semantic Web: MEDEA (Modeling semantically Enhanced Digital Edition of Accounts)

Kathryn Tomasek, Wheaton College (Massachusetts); Georg Vogeler, Centre for Information Modeling - Austrian Centre for Digital Humanities, University of Graz

Social and economic historians have spent at least the past fifty years creating data sets well suited for analysis using post-WWII computational tools (SPSS/SASS). Contemporary efforts by such historians as Patrick Manning to aggregate data sets for human systems analysis demonstrate a desire to take advantage of the more recent tools represented by the semantic web. Both Tomasek and Vogeler have explored ontologies that can be integrated into the CIDOC-CRM family of event-based models and used for markup of digital scholarly editions of accounts, a genre of archival documents that support humanities research as well as social science research. This short paper offers a brief introduction to recommendations for producing digital scholarly editions of accounts that include references to a book-keeping ontology using the TEI attribute @ana. Vogeler has tested comparability of data across a small sample of such editions for which the references have been transformed into RDF triples. New editions are being added to those stored in the GAMS repository (Geisteswissenschaftliches Asset Management System) at the University of Graz between now and August 2017. We see these editions in sharp contrast to the example of “page-turning” simulations referenced in the cfp for the workshop: creating full digital scholarly editions of accounts using TEI, the book-keeping ontology, and RDF triples are an example of shaping humanities data for use and reuse by taking advantage of the affordances of the semantic web.

Javanese Theatre as Data

Miguel Escobar Varela, National University of Singapore

The [Contemporary Wayang Archive](#) is an archive of Indonesian theatre materials. The online portal's primary goal is to enable users to watch videos alongside transcripts, translations and scholarly notes. However, a new version currently under development will enable users to query the archival materials via APIs. The first API will be directed at linguistic queries from the transcript and translation corpus. The goal is to enable data-driven investigations of the ways Javanese and Indonesian are used in the performances. Although these languages are widely spoken (Indonesia is the fourth most populous country in the world and Javanese is its most widely spoken regional language), there are almost no machine-readable resources in these languages that can be used in digital humanities and computational linguistics research projects. A second API is aimed at video processing applications. The API will serve video-frame-level data that can be used to interrogate and visualize the collection in new ways. We believe that most theatre projects in DH remain heavily focused on textual data or on numerical data such as revenue numbers, cast sizes and collaboration networks. However, we believe that video processing offers a rich and yet untapped avenue for inquiry [1]. We aim to encourage further research into this area via our video processing API. This talk will briefly outline the objectives and history of CWA, our goals for the future and the technical and intellectual property rights challenges that we face.

References: [1] Escobar Varela, M and G.O.F. Parikesit, 'A Quantitative Close Analysis of a Theatre Video Recording' in *Digital Scholarship in the Humanities* (forthcoming), doi:10.1093/llc/fqv069

