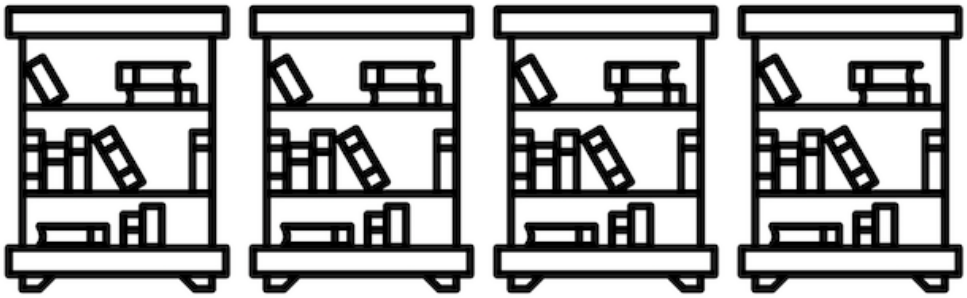
Always Already Computational: Collections as Data



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

Collections as Data Facets

August 2017 - August 2018

This publication is part of the Collections as Data Framework hosted at https://osf.io/mx6uk/.

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.

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 etheses. 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 etheses 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 ethesis 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 (etheses), 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 ethesis 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 etheses 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/cmoa/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 theyat 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 as 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 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 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 of 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 included 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 rysnc 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 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 Zarafonetis (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. Interlineal 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 as 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 isan 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 date-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.