Collections as Data: Part to Whole
Final Report

---

1. **Date Submitted:** August 31, 2020

2. **Recipient organization:** University of Pittsburgh

3. **Project team:**

    - Tyrica Terry Kapral (Project Lead), Humanities Data Librarian, University of Pittsburgh
    - Aaron Brenner, Associate University Librarian for Digital Scholarship & Creation, University of Pittsburgh
    - Matthew J. Lavin, Assistant Professor of Humanities Analytics, Data Analytics Program, Denison University (Fall 2020–present); Clinical Assistant Professor of English and Director of Digital Media Lab, University of Pittsburgh (Fall 2015–Spring 2020)
    - Gesina Phillips, Digital Scholarship Librarian, University of Pittsburgh

4. **Project title:** From Collection Records to Data Layers: A Critical Experiment in Collaborative Practice (operational short title: "CaD@Pitt")

5. **Project summary:**

The CaD@Pitt project is based in the University of Pittsburgh Library System and aims to increase the visibility and discoverability of library collections, make library collections data accessible for computational use, and enable scholars to extend/enrich collections data with critical, research-driven layers of additional data. This project is targeted toward library workers, scholars (researchers, students, and instructors), and constituents of the public who are seeking to implement similar endeavors at their own institution, acquire and generate collections data, or develop/incorporate collections data-centered curricula to teach critical and computationally minded data practices.

The CaD@Pitt project includes archival collections and catalog items in the Library's general, special, and distinctive collections. Our project also enables researchers to curate new collections by selecting items from these collections. While our project is intended to support any and all library collections, it has prioritized those reflecting the perspectives of underrepresented groups, which are often given inadequate attention for structural or systemic reasons. These collections have strong potential to tell as-yet untold stories and can increase

institutional commitment to bringing underrepresented histories to light. Our target collections feature the following:

- the voices of African Americans, American Labor Unionists, American left-wing organizations, Latinx communities, the LGBTQ community, and feminists;
- a diverse array of serials (e.g., journals, magazines, newspapers, newsletters) and ephemera (e.g., broadsides, flyers, cartoons).

Serials and ephemera are foregrounded in this project because, historically, the conditions of their production and reception often made serials and ephemera more desirable venues for specific counterpublics than monograph publication.

Our project takes a "data layers" approach, which diverges from monolithic data paradigms (i.e., singular, non-interpretive, and exhaustive). Instead, it presents an interface comprising data from multiple sources that vary in encoding scheme, granularity of description, and completeness/richness. This approach liberates data creation and curation from the expectation of perfection or singularity of authority, and it allows data to be enriched, augmented and interpreted incrementally over time through a layering process. It also provides a practical and low-barrier entry point for scholars to create datasets, or data layers, based on their research priorities and to share those layers. Our project model specifies three types of data layers:

- **source data**: snapshots of library collections metadata files in their original source format (e.g., MARC21, EAD, DC, MODS, RELS-EXT/RDF);
- **base layers**: curated datasets (CSV files) derived and simplified from the source data layers; and,
- **extension layers**: scholar-created datasets or outputs that enrich/augment library collections data (i.e., source data and base layers).

Scholars can download prepared data layers from the repository, request source data and base layers for library collections not already represented in the repository, and share their own extension layers in the repository. Our project serves as *a* model for how other libraries can share collections data, and our repository can support data layers for collections from other libraries, though local and (inter)national element sets and controlled vocabularies may be used differently across institutions.

The base layers are created using a Python program that extracts and transforms source data into a flat data model (i.e., CSV). This data munging/wrangling process involves mapping the base layer elements to their respective nodes in source data XML files (i.e., EAD, MODS, MARCXML, or RELS-EXT/RDF), selecting only these mapped nodes for extraction, inputting (sometimes processed) values into rows and columns, and outputting one or two CSV files for collection-level and/or item-level data. It also automatically creates folders for output files.

Finally, this project facilitates scholars' engagement with and enrichment of library collections data through five instructional modules that 1) orient learners to collections as data and as products of curation and 2) teach critical and computationally minded data practices:

- **Develop a Custom Collection**: Create a collection, drawing from any combination of the Library's general, special, distinctive and/or archival collections;
- **Design a Layer**: Propose a data collection plan for a dataset (extension layer), based on a custom or pre-existing collection, that answers a research question or meets a particular need;
- **Critique a Layer**: Critique the utility, feasibility, and ethicality of an extension layer;
- **Implement a Layer**: Implement an extension layer by entering data into a spreadsheet;
- **Visualize a Layer**: Visualize collections data using a visualization tool.

These modules have been designed as a sequence but may be used for individual lessons and otherwise modified to suit varying contexts.

6. Activities:

Activities proposed in application	**Status of proposed activities: complete, incomplete, canceled**	**Explanation of variances with proposed activity**
Extract collections data from containing systems, determine flattened data model for base layer, and transform data into CSV files	complete	Task that was not specified in the proposal: Develop Python scripts to transform data from hierarchical data model (XML) to flattened data model (CSV).
Create online repository	complete	
Upload CSV files to online repository for open access	complete	We have also uploaded source data to the Data Layers Repository (GitHub).

Prepare web application and instructional documentation for students to use to create data layers	canceled	Due to the cancelation of the pilot course we intended to serve as a pedagogical laboratory for our project, we were unable to implement all of the proposed activities with students. Instead, Lavin directed 2 undergraduate students in independent research projects using data layers and computational methods. We also developed 5 instructional modules and collaborated with two other instructors to integrate 1-2 modules in their courses, which entailed introducing collections as data concepts; discussing data ethics; and proposing, critiquing, implementing, and/or visualizing extension layers. Instead of the web application, we used Google Sheets and Forms. We did not test a "best practices" document with students, but we have completed the document based on course integration experiences.
Work with students to develop extension layers, generate data visualizations, reflect on process with suggestions for possible collections of interest for future classes, and develop/refine "best practices" documentation for creating extension layers.	incomplete	
Work with students to publish their data layers and documentation	canceled	
Develop workflows for incorporating enrichments from the students' data layers back into library collections data	incomplete	The team has brainstormed possible methods for integrating data enrichments into the Library's digital repository and catalog, but further development has been stalled due to the Library's integrated library system (ILS) migration project.
Enable robust use of data layers by identifying potential common models for publishing and using	complete	Much of this activity has been incorporated in the overall project documentation. While we have not created

extension layers in online repositories; creating methods for using fields and records from previously created data layers and joining fields between datasets; and developing and publishing usage documentation for users		generalized methods for joining fields between datasets, the project enables open access to previously created data layers and merging data layers via unique identifiers as foreign keys, and the Data Layers Repository includes an example of merging fields between two datasets.
Share project work with local and wider DH community	incomplete	While one of the project team members was able to present about the project work at a national conference (DLF2019), the team as whole was unable to present at the conference to which our proposal has been accepted (Keystone DH 2020), as it was canceled due to the COVID-19 pandemic.

7. Additional project changes:

Activities	**Rationale + Impact**	**Date approved**
Added Gesina Phillips to the project team as the Teaching and Learning Coordinator.	Phillips's contribution to the work of the project has been considerable and essential to the development of the instructional modules and their implementation in courses, the datathon, and library-based projects.	2019-12-05
Hosted datathon, a crowdsourced data entry event, on February 27, 2020. During the event, we introduced participants to the	This event allowed us to test the "Implement a Layer" instructional module as well as implementation of modules in settings outside	2020-02-28

concept of collections as data participants contributed to extension layers (subject tags and geotags).	the classroom.	

8. Observations, insights, and new understandings:

The principles and practices of the data layers model diverge from those of libraries, and this kind of collections as data project requires reconciling some major differences. Libraries historically prioritize making resources available to patrons, whereas our project aims to make the metadata for these resources available to users in computation-ready forms. Currently, users cannot export collections data from the Library's digital repository, partly because there is no technological infrastructure in place to support this and partly because the metadata quality does not, for the most part, meet the standards of a more monolithic data paradigm. Thus, only library specialists with access to the administrative interface for the Library's installation of Islandora (created by Library IT, namely Willow Gillingham) can export data. At the beginning of the CaD@Pitt project, Voyager was the Library's integrated library system (ILS), which allowed patrons to bulk export binary MARC records easily; with the ILS migration to Ex Libris Alma-Primo in Summer 2020, which promotes discovery of and access to resources, this ease of access was lost. One of our project's next steps will be to determine the best way to provide access to catalog collections data and to establish a workflow.

Libraries often make resource lists available in formats that privilege reading and printing, like LibGuides and document files (e.g., PDF, MS Word, images scans of documents), rather than computation, processing, manipulation, etc. In the CaD@Pitt project, part of addressing this issue entailed manually transferring data from LibGuides and document files to spreadsheets, which enabled us to programmatically extract identifiers for exporting records. Depending on the number of documents from which data must be transferred, this process may require more programmatic means and/or time. Further, our project privileges encoded and tabular data (in our data layers) to enable computational methods. The datasets we have created using LibGuides and PDFs for special collections can now be shared with their curators for more robust usage.

Because making digital surrogates of archival materials available is often prioritized over generating rich and consistent metadata—a time- and labor-consuming task—we had to grapple with developing base layers that do not reliably provide many data points that users want or expect, such as genre/subject data. That said, there is a considerable range in the quality of descriptive metadata (e.g., richness, granularity, and completeness) across archival collections,

due to changing technologies, tools, scale, standards, professional practices, etc. A few, more specific factors that affect our project include:

- priorities and circumstances during the time the collection was processed/digitized (e.g., quick turnaround, specific user-driven goals);
- differences in metadata entry practices across units or from project-to-project and changes over time (e.g., practices before and after the establishment of our library's Metadata and Discovery unit; the guidelines for describing photographs differ from those of other types of objects due to curatorial decisions);
- the type of objects being described (e.g., a book may provide more and different information than, say, an event flyer);
- differences between originating software and their organization of data (e.g., Archivists' Toolkit, ArchivesSpace, Islandora).

In developing our base layer metadata element sets, we strove to balance between incorporating data points that are common denominators across collections and including data points that we expect to be useful for research. Nevertheless, the base layers for many, if not most, collections are incomplete to some extent.

In addition to the incongruities between the principles and practices of our data layers model and that of libraries, we found that implementing our collections as data project was a bigger and more challenging undertaking than we anticipated, largely due to complexities in developing our base layers, the numerous objectives of the project (described in the Project Summary) and the expertise required to achieve them, the nature and vast variety of our collections data, the complexity of the code required to process the data, and the challenges of implementing the instructional modules in various contexts.

A large part of our project was conceptualizing the "base layer." In choosing a tabular data model for our base layer, our goal is to provide access to collections data in a form that is not only computation-ready but also familiar/accessible to most of our users, like the scholars who have been creating spreadsheet datasets for our library collections over the past five years. Other institutions may opt for different data models and principles in their projects, including hierarchical data models like TEI and (semantic) network models like linked open data. Though our base layer centers flat data, our data layer concept is rather data-model agnostic and presupposes that extension layers may take other data models or forms, such as TEI- or RDF-based data layers, geospatial data layers (e.g., vector or raster data), or any number of data outputs and visualizations. Tabular data was the lowest-barrier implementation for our project; however, determining the scope of our base layers still proved challenging. Initially, we intended to create base layer metadata element sets that included data points providing contextual data that we, as educators and curators, feel is necessary for critical/ethical/responsible use of the data. For example, what if we included an element for the names of people who digitized the object to

acknowledge the labor(ers) behind making the resource available, even if that meant that most of these fields may be left empty because institutions don't maintain a record/memory of this kind of information? Ultimately, we decided that the base layer should provide data points that *are* collected by the library, amenable to computational research (i.e., relatively standardized), and common enough to warrant cross-collection analysis. Opportunities for critical reflection lie in data documentation, discussions surrounding the data (e.g., in instructional sessions), and extension layers. For instance, students' attempts to visualize the creation of collection materials across time raised the question of how to proceed when the creation date, a seemingly straightforward data point, is not known/exact (e.g., a date range) or when there is conflicting information. Others taking on similar collections as data projects may settle on other resolutions, depending on their values, objectives, audiences, and resources.

Furthermore, we had to consider how many of the viable data fields to include in our base layers and how much to process/clean the source data. We pared down the data elements in the base layers from those in the source data to avoid representing an overwhelming amount of information and to include data points that we expect to be most useful for scholars in general. However, metadata elements that have been excluded in the base layer may be incorporated in an extended version of the base layer (i.e., extension layer) by request; in other words, scholars may customize the base layer to include one or more additional elements from the source data. Because we want our data layer model to be flexible, extensible, and user-driven, we may modify the base layer according to scholars' usage patterns. Regarding data processing, we have determined not to clean the source data for base layers at this time and to only process the data in order to format concatenated values; data cleaning is reserved for extension layer projects because we, as a project team and an institution, do not have the capacity to clean the records for all collection items. Base layers may be updated to incorporate cleaned data. So far, the project team has pursued extension layer projects to clean/enter data values for title, creator, date, and genre elements. We look forward to working with the Metadata and Discovery unit to update the collection records in our Library's digital repository as well.

At the beginning of our project, we knew that there would be different base layer metadata element sets for archival materials, serials, and monographs; yet, it became more complicated once we began developing our base layers. Our project brings together collections metadata from finding aids for non-digitized and (partially or entirely) digitized archival collections, metadata files created during digitization/ingestion of archival collection items, and catalog records. This entails reconciling numerous disparities between collection item metadata in our project's base layers. Our archival collection base layers utilize EAD files derived from finding aids, which describe not only the collection but the finding aid itself, as well as MODS files that describe collection items. Thus, archival collections in our project are described at the finding aid level, collection level, and item level, resulting in two CSV files (with the finding aid and collection levels combined). There is no item-level base layer for non-digitized items

because only digitized items receive MODS files, but scholars can still create extension layers for analog items. The project team originally planned to include a series-level base layer for archival collections, but due to the limited number of data points for series-level description, we decided to subsume it in the item-level base layer. We had also planned to include technical metadata in the item-level base layer, such as digitization dates and file information (e.g., name, format, size), but we realized that this description is not at the item level but the image level; for example, technical metadata describes the files for each page of a book, not the book itself. We are excluding technical metadata from the base layer at this time, but it is possible for scholars to request this data, and we may figure out a way to aggregate this data for digital items comprising multiple images. Catalog (MARC) records tend to be richer and more consistent than our records for digitized archival collection items—though there are digital collection items that were previously cataloged, the metadata records for which are drawn from MARC records. On the other hand, digitized serials in archival collections are described at the issue level, while serials in the catalog are primarily described at the title level; in other words, description for serials tend to be more granular for digitized archival collections. To address these discrepancies, our project employs three major types of base layer with the following variations:

- **Archival collections**: digitized (finding aid/collection-level, and item-level base layers) or non-digitized (finding aid/collection-level base layer);
- **Serial collections**: from the catalog (item-level base layer) or from the digital repository (finding aid-level, collection-level, and item-level base layers);
- **Monograph collections**: from the catalog (item-level base layer) or from the digital repository (finding aid-level, collection-level, and item-level base layers).

Our data layer model is relational, using primary and foreign keys (unique identifiers) to establish relationships between base layers in multi-level sets and item records across data layers. Development of the base layer may be more or less complex in projects at other institutions, depending on the range of collection types and available source data.

This kind of project requires a range of expertise and competencies in the following: researchers' collections-based needs and interests, collections metadata (e.g., creation practices, standards/schemas, encoding, vocabularies), local library/archival collections (e.g., curatorial knowledge), local library/repository system(s), GitHub (or a similar platform), scripting in Python (or another language that supports text processing and working with XML), tabular data, pedagogy and instruction, and project management. While the project team received some much-appreciated assistance from experts in the Library's Metadata and Discovery unit (Head, Mike Bolam), Archives and Special Collections department (Digital Collections Coordinator, Kristin Britanik), IT department (System Developer, Willow Gillingham), and from a graduate student assistant (Evelyn Chan) in organizing the datathon, all of the project's roles and responsibilities were distributed between the four project team members. Given the time-frame

of the grant period, it was quite time-consuming and labor-intensive to complete the work of the project. Forming a larger team with well-dispersed and -coordinated responsibilities is ideal for this kind of project.

Understanding and evaluating our collections metadata required countless hours of examining metadata files and documenting findings; reading documentation for metadata standards/schemas, controlled vocabularies, encoding schemes, and Islandora; and discussing the nature of our local metadata files, metadata creation practices with Bolam and Britanik. Thanks to an XSLT script created by Bolam for XML-to-TSV conversions, we were able to learn XSLT well enough to modify the initial script to analyze metadata files. Programmatic metadata analysis is essential to this project because it enables us to better understand the nature of and patterns across our collections records, such as missing element values or all subelements and attribute values for elements currently in use. Moreover, this is the only way to properly document the base layers in our data dictionary and application profiles, which *describe* how data has already been input rather than *prescribe* how data should be input. In the process of analyzing the metadata, we have been able to share findings with Bolam and Britanik, enabling the resolution of various metadata issues; hence, the CaD@Pitt project has established a mutualistic relationship with the Metadata and Discovery unit and the Archives and Special Collections department. This project has engaged with a wide range of metadata standards/schemes, including EAD, MODS, MARC, Dublin Core (DC), RDF (RELS-EXT, RELS-INT) and FITS (TECHMD)—quite the alphabet soup. We have chosen to use EAD, MODS, and RELS-EXT for archival collections and MARC for catalog records because they offer the richest data for our collections items; we convert binary MARC records to MODS files to simplify the data munging process in creating base layers (same scripting as for archival collections). Projects teams comprising metadata experts and/or working with fewer standards/schemas will likely proceed through the metadata evaluation process more quickly/efficiently. In any case, findings will impact the development of the base layer(s) and munging script(s).

Writing efficient, reusable code has been a challenge unto itself because of the idiosyncrasies in XML files, even when they are ostensibly in the same format and among very similar collections. We have been developing a Python program to extract base layer values, but we found a need to write numerous “ad hoc” rules to deal with potential inconsistencies in XML files, and we have often had to rely on our script to do more data processing than we would like (such as splitting a pair of values encoded as one string, or using a regular expression to extract a substring with the desired data). Initially, we were working in a Jupyter notebook that could be easily run by scholars but, as our code grew in complexity, it became unwieldy. So we are currently using command line arguments and two Python scripts, one for the functions and one (config file) for the dictionaries of data fields that are passed to the functions. We ultimately settled on an approach that 1) converts XML of BeautifulSoup objects; 2) uses BeautifulSoup

select() statements to extract data; and 3) calls custom functions when a straightforward select() statement is insufficient.

The appeal of this approach was that we initially wrote a generic function to call any select() statement. All select() statements could then be specified in a separate file. In theory, updating the code would be as simple as updating the config file. Over time, however, it's become clear that specialized checks and queries are more commonly needed than we anticipated. For example, when extracting data for one or more 'creator' fields, our script uses a select() statement to select all 'name' elements that are children of 'mods' elements. A custom function then finds the 'name' element's 'role' child element, and looks for a 'roleTerm' grandchild element equal to the value 'creator'. If the value is a match, all 'namePart' element values are returned and concatenated (comma separated) for the final value of the 'creator' column. Long-term, we hope that more consistency will be established upstream of our script. If we can encode XML files more consistently (or even a handful of fields more consistently) our extractor will have a better chance of remaining relevant. However, we also need to be prepared to adapt our current code more effectively to newly discovered idiosyncrasies.

Engaging in various instructional contexts throughout our project has been challenging but enlightening. Due to the cancellation of the class in which we expected to step through the modules in sequence, we were unable to test the modules with a cohort of students over the course of a semester. However, in Fall 2019 and Spring 2020, we were able to teach versions of the "Design a Layer," "Critique a Layer," "Implement a Layer," and "Visualize a Layer" modules through class visits (thanks to Amy Murray Twyning and Daniel Libertz), and we used a version of the "Implement a Layer" module for our datathon. During the class visits (two for each course) and homework assignments, students actively engaged with concepts related to data ethics and potential bias, which was particularly in line for classes that had already encountered works like Safiya Umoja Noble's *Algorithms of Oppression* or Cathy O'Neill's *Weapons of Math Destruction*. The students tended to propose sophisticated research ideas, from sentiment analysis to programmatically measuring images or blank space on a page. Although students ran up against questions of scale and feasibility while discussing the implementation of such projects, their enthusiasm tended to lead to constructive class sessions. This testing was invaluable for the creation and revision of the modules, but did not allow for continuous engagement with learners or gathering feedback about the modules as a whole.

Fortunately, in Summer 2020 we were able to engage with three library colleagues (Sandy Buehner, Diane Hughes, and Ed Lewis) and one MLIS graduate student intern (Kahlila Chaar-Pérez, who also consulted on modules under development/revision) as they tested the modules throughout their own research processes. Their projects have resulted in the development of two collections, with two extension layers in the works: a 19th-century juvenile literature collection and a standardized genre extension layer; 2) an Afro-Latinx collection and research-driven extension layer. This allowed us to test the first four modules sequentially and

make adjustments to their flow and clarity. Simultaneously presenting to our learners the concept of collections as data, the multiple objectives and strategies of our project, and the tasks that they would be undertaking during introductory sessions proved too overwhelming for them. We are, accordingly, scaling back the content for introductory sessions. Ultimately, we discovered that the modules can be implemented in several contexts and with many different audiences. Though they were initially designed for classroom engagements, we found that they can be used effectively as a framework for orienting people who intend to create extension layers, as in the cases of our tester-researchers. In their current form, this is perhaps not an obvious fit for the modules; future revision may lead to a more streamlined model—still based on the modules—for orienting researchers. We also found the "Implement a Layer" module worked well on its own as a framework for the datathon, gesturing toward other possible implementations of collections as data projects like ours in co-curricular and community settings.

9. Repository-backed final deliverables:

- **CaD@Pitt project website**: https://CaDatPitt.github.io
- **CaD@Pitt GitHub repositories**: https://github.com/CaDatPitt
    - Data layers repository: https://github.com/CaDatPitt/data-layers
        - Collections data and datasets
        - Data extraction/transformation scripts
    - Data visualization demos repository: https://github.com/CaDatPitt/visualization-demos
        - Sample data visualizations of data layers
    - Documentation repository: https://github.com/CaDatPitt/documentation
        - Project documentation
    - Instructional modules repository:
        - Materials for and artifacts from CaD@Pitt instructional modules

10. Publications, articles, presentations, and news articles related to the grant:

- Luster, Dominique, Samantha Ticknor, Tyrica Terry Kapral, and Obden Mondesir. (2019). *Collections as Data: The Outtakes*. Panel session at the 2019 DLF Forum. https://dlfforum2019.sched.com/event/S2US/m2e-collections-as-data-the-outtakes.

11. Expected post-grant activities:

The project team is looking forward to several post-grant activities. In addition to continuing to develop and maintain the Cad@Pitt GitHub repositories, we plan to establish a more robust use model by developing the following:

- public-facing mechanisms (e.g., workflows, forms) for users to request/acquire library collections data and contribute data layers to the repository; and,
- a dynamic user interface that will better facilitate discovering, exploring, and generating collections datasets (e.g., enabling tasks such as searching/filtering/faceting by fields of the datasets, generating data visualizations, analyzing data layers).

We will also continue with curricular development and course integrations of instructional modules, including collaborating with the University of Pittsburgh's Humanities Engage program on such endeavors as the program's opportunities for developing collections-based modules for undergraduate courses. With the conclusion of the ULS's ILS migration, we will pursue developing and implementing workflows for incorporating enrichments from extension layers back into the library catalog and digital repository, coordinating with Library units such as the Metadata and Discovery, IT, and Archives & Special Collections. Finally, we anticipate submitting one or more papers for conference presentations or publication, regarding lessons learned, the data munging and coding aspects of the project, the data layers model, and/or collections-focused analyses enabled by the project's data layers.