DHQ: Digital Humanities Quarterly
2020
Volume 14 Number 2

Open Data in Cultural Heritage Institutions: Can We Be Better Than Data Brokers?

S.L. Ziegler <sziegler1_at_lsu_dot_edu>, Louisiana State University Libraries

Abstract

Treating collections in cultural institutions as data encourages novel approaches to the use of historic collections. To reframe collections as data is to focus on how digitized collection material, collection metadata, and transcriptions can be used and reused for various types of computational analysis. Scholars active in the field of digital humanities have long taken advantage of computational data. This paper focuses on the work of cultural heritage institutions, which are increasingly offering collections as data. This paper outlines the collections as data project and examines specific examples of cultural institutions active in this space. The paper then details the practices of data brokers, and explores how the data broker model can frame the use of data in cultural heritage institutions. In closing a number of experiments are described that might help mitigate the harm that data in cultural institutions might cause. As we create and share data, can we be sure we are better than data brokers?

Introduction: Collections as Data in Cultural Institutions

Recently, increased attention has been paid to data in cultural institutions.[1] In both 2016 and 2017, the Library of Congress hosted conferences on the use of library collections as data [Library of Congress 2016] [Library of Congress 2017]. Also in 2016, the Institute of Museums and Library Services (IMLS) funded a two-year project, Always Already Computational (AAC), “lead to the creation of a framework to support library collections as data, the identification of methods for making computationally-amenable library collections more discoverable, use cases and user stories for such collections, and guidance for future technical development” [IMLS 2016]. In addition to workshops and meetings, the AAC[2] team compiles information on like-minded projects, and has released “The Santa Barbara Statement on Collections as Data”, a document of guiding principles for treating collections as data [AAC 2018b]. 1

The AAC project has done a great deal of important work in bringing together a wide variety of practitioners and examples and for this reason situating an exploration of data in cultural heritage institutions within the framework of the collections as data conversation is beneficial. As a catalyst within a wider world of data-oriented endeavors in cultural institutions, the AAC project has opened new avenues of investigation and has amplified the need for collaboration among institutions and practitioners. It is becoming increasingly common to see issues related to data in special collections libraries appearing in syllabi, library strategic goals, and position papers [Lied Library]. “The growing interest in collections as data,” writes Chela Scott Weber in a recent OCLC Research Position Paper, “means we must collaborate with colleagues in scholarly communications, data services, and elsewhere across the library to grapple with what computational access to our collections might look like” [Weber 2017]. 2

Collections as data explores an expansive definition of data. “To see collections as data begins with reframing all digital objects as data,” Thomas Padilla writes. “Data are defined as ordered information, stored digitally, that are amenable to computation. Wax cylinders, reel to reel tape, vellum manuscripts, websites, masterworks, musical scores, social media, code and software in digital collections are brought onto the same field of consideration” [Padilla 2017]. In addition to digital collections being reconceptualized as data, the metadata — such as titles, descriptions, dates — can also be rethought as data. “Data as well as the data that describe those data,” explains the Santa Barbara Statement, “are 3

considered in scope. For example, images and the metadata, finding aids, and/or catalogs that describe them are equally in scope. Data resulting from the analysis of those data are also in scope" [AAC 2018b].

4 In many ways treating collections as data eases some barriers to sharing data. However, collections as data is not the same as open data. Open data has few, if any restrictions on use and reuse [Open Knowledge International n.d.]. "Accessibility and reusability," write Koster and Woutersen-Windhouwer, "do not require collections and objects to be freely available, modifiable and shareable with free tools," as the open definition requires. "Some metadata or objects will be copyright protected, have privacy issues or local law issues" [Koster 2018]. When we rethink collections as data, collections are usually easier to share, however, there are still many reasons that the data might not be open.[3]

5 In November 2019 the AAC grant project came to an end and was succeeded by a new phase, Collections as Data: Part to Whole ("Part to Whole" n.d.). Collections as Data: Part to Whole fosters "the development of broadly viable models that support implementation and use of collections as data" by funding project teams that "will develop models that support collections as data implementation and holistic reconceptualization of services and roles that support scholarly use" ("Part to Whole" n.d.). Even beyond the AAC project the number and scale of cultural heritage collections available as data continues to increase. In 2019, the Library of Congress, with funding from the Andrew W. Mellon Foundation, launched the Computing Cultural Heritage in the Cloud (CCHC) project, to "pilot ways to combine cutting edge technology and the collections of the largest library in the world, to support digital research at scale" [Library of Congress 2019]. Also in 2019, the National Library of Scotland launched the Data Foundry Website to present, "collections as data in a machine-readable format, widening the scope for digital research and analysis" [National Library of Scotland 2019].

Data in Cultural Heritage Institutions

6 The AAC team has compiled a number of Facets, or case studies that draw attention to many ways cultural institutions are creating and using data.[4] These examples include the use metadata, digital facsimiles, and structured transcriptions. The metadata examples show new forms of access and engagement with collections, "allowing people to creatively re-imagine and re-engineer our collection in the digital space" [Newbury and Fowler n.d.]. The Carnegie Museum of Art, for example, makes available "data on approximately 28,269 objects across all departments of the museum[:] fine arts, decorative arts, photography, contemporary art, and the Heinz Architectural Center" [Carnegie n.d.]. Released as part of the 120th anniversary celebration of the museum, the data promotes the central mission of the museum. "The case to provide the public increased access to museum data was not a difficult one at the Carnegie Museum of Art," explain the authors of the data, "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" [Newbury and Fowler n.d.].

7 In addition to sharing metadata, many libraries and museums are allowing full access to digital facsimiles. The Getty, for example, "makes available, without charge, all available digital images to which the Getty holds the rights or that are in the public domain to be used for any purpose. No permission is required" [Getty Trust n.d.]. As another example, OPenn, at the University of Pennsylvania, "contains complete sets of high-resolution archival images of manuscripts … along with machine-readable TEI P5 descriptions and technical metadata. All materials on this site are in the public domain or released under Creative Commons licenses as Free Cultural Works" [University of Pennsylvania n.d.].

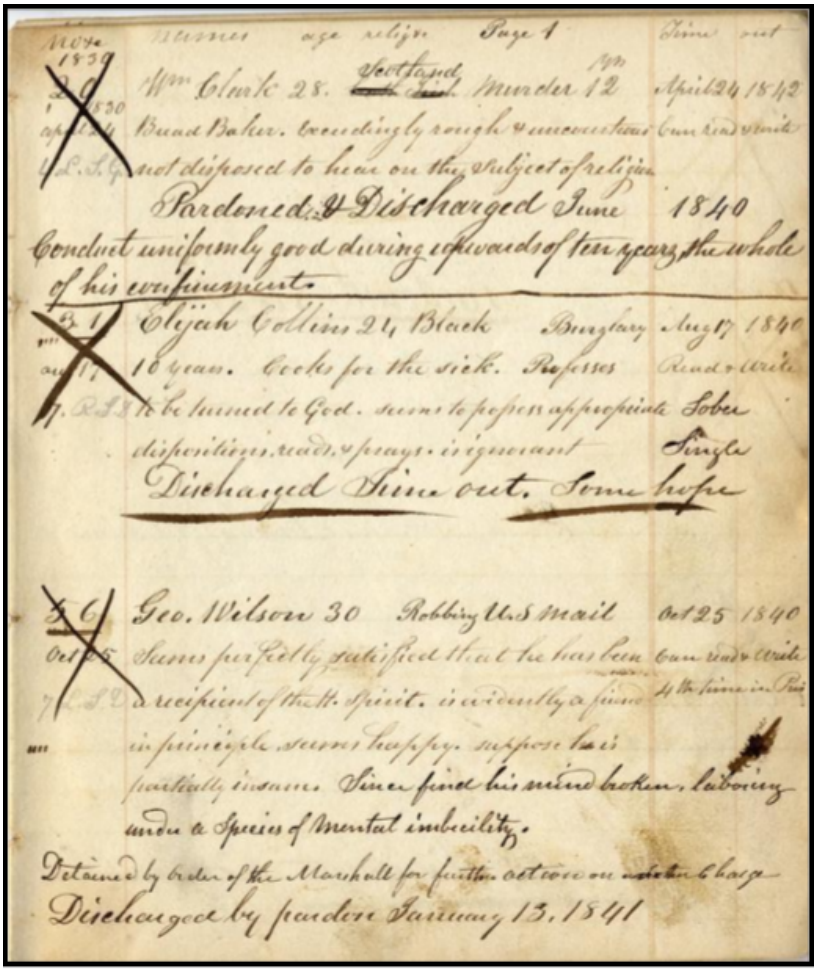
8 Beyond metadata and digital facsimiles, collections reformatted as structured data is also a growing trend in cultural institutions. This often involves transcribing text and applying some type of structure. Haverford College Libraries reformatted collections into TEI-encoded text for the projects "Beyond Penn's Treaty" [Zarafonetis and Horowitz n.d.] and "Ticha: A Digital Text Explorer for Colonial Zapotec" [Lillehaugen and Zarafonetis n.d.]. Reformatting the collections as structured data allows for enhanced linkages between items as well as customized interfaces. The Museum of Modern Art (MOMA) structured exhibition data in a database, and shared as a CSV (Lill n.d.). This project has resulted in a rich history of past MOMA exhibitions.

9 A fuller exploration of a particular open data project will help highlight both the process and possible pitfalls of collections as data work in cultural heritage institutions. The American Philosophical Society Library (APS) digitized,

reformatted as data, and opened historic prison records. The APS holds three admission books created by the Eastern State Penitentiary.[5] The books, covering the years 1830-1850 (with a gap in the 1840s), hold information on each prisoner, including name, age, the crime(s) for which they had been found guilty, the sentence, and often a note on when they were freed (or died). Also included, though less consistently, is gender, race, and religious affiliation. Additionally, notes from the moral instructor appear for each record; similar to contemporary prison parsons, moral instructors were non-denominational religious authorities and they recorded a paragraph-length note on each inmate which details the religious education of the prisoner as well as other biographical elements.

In 2015, the team[6] digitized the admission books. At this point, the scanned images of the pages of the admission books could be treated as data. As data, they are open to computational analysis. Just as digital representations of paintings allow the application of computer models to extract information and recombine it to find new patterns and propose linkages between pieces,[7] the pages of the prison admission books can be treated as data. For example, by analyzing the pages as data, one could ask if the handwriting changes based on the positive or negative attributes ascribed to the prisoner.

10



Sample Image of Prison Admission Book showing typical entry. Each entry begins with prisoner's name, age, race (usually if non-white), location of birth (usually if outside of US), and offense

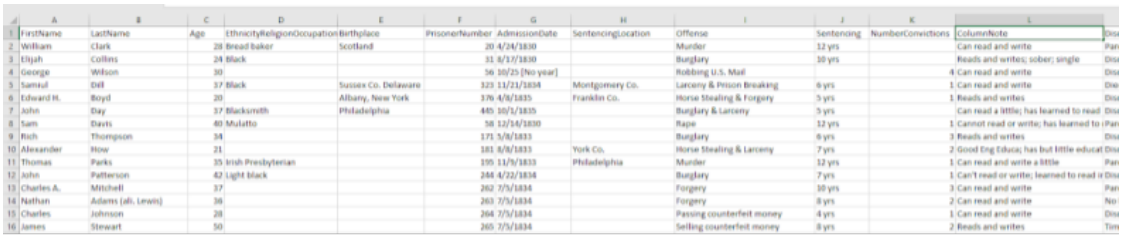
Figure 1.

Taking the process one step further, in 2016 and 2017 the team transcribed the content into spreadsheets.[8] The library published the data as CSV files in a variety of outlets, including the APS Github repository [APSa n.d.] and the Magazine for Early American Datasets [Ziegler and Ziogas 2016]. The workflow was simple: each transcriber had two computer monitors, one to see the digitized page and the other to type into a spreadsheet. The information on the pages are mostly consistent. For example, the first line of each entry usually includes the prisoner's name, age, crime for which he or she is incarcerated, race (usually if non-White), and location of birth (if outside of US). The repetition and

11

consistency enabled the information to be captured with few issues.

However, the process is not always obvious, and choices do matter. For example, the handwriting can often be difficult to read, and conventions about how to depict abbreviations, spelling errors and indecipherable words had to be established.[9] Sometimes racial categories are mentioned, such as *black*, or *light black*. Sometimes, on the same place on the page, religions are mentioned, such as *Catholic*. Sometimes the birth location is listed near the same place on the page, such as “Irish”. It is easy to imagine that the categories of race, religion, and birth location were doing interpretive work for both the creator of the records and the intended audience. Jacksonian America was a time of great anxiety of immigration and saw a large number of race riots [Feldberg 1980]. The transcription team needed to make decisions about how to group these pieces of information. Records are not neutral; the process of transcription is not neutral.[10] 12



	A	B	C	D	E	F	G	H	I	J	K	L	
1	FirstName	LastName	Age	EthnicityReligionOccupation	Birthplace	PrisonerNumber	AdmissionDate	SentencingLocation	Offense	Sentencing	NumberConvictions	ColumnNote	
2	William	Clark	28	Bread baker	Scotland	20	4/24/1830		Murder	12 yrs		Can read and write	Pan
3	Elijah	Collins	24	Black		31	8/17/1830		Burglary	10 yrs		Reads and writes; sober; single	Dis
4	George	Wilson	30			56	10/25 [No year]		Robbing U.S. Mail		4	Can read and write	Dis
5	Samuel	Dill	37	Black	Sussex Co. Delaware	323	11/21/1834	Montgomery Co.	Larceny & Prison Breaking	6 yrs	1	Can read and write	Dis
6	Edward H.	Boyd	20		Albany, New York	376	4/8/1835	Franklin Co.	Horse Stealing & Forgery	5 yrs	1	Reads and writes	Dis
7	John	Day	37	Blacksmith	Philadelphia	445	10/1/1835		Burglary & Larceny	5 yrs		Can read a little; has learned to read	Dis
8	Sam	Davis	40	Mulatto		58	12/14/1830		Rape	12 yrs	1	Cannot read or write; has learned to r	Pan
9	Rich	Thompson	34			171	5/8/1833		Burglary	6 yrs	3	Reads and writes	Dis
10	Alexander	How	21			181	8/8/1833	York Co.	Horse Stealing & Larceny	7 yrs	2	Good Eng Educa; has but little educat	Dis
11	Thomas	Parks	35	Irish Presbyterian		195	11/5/1833	Philadelphia	Murder	12 yrs	1	Can read and write a little	Pan
12	John	Patterson	42	Light black		244	4/22/1834		Burglary	7 yrs	1	Can't read or write; learned to read a	Dis
13	Charles A.	Mitchell	37			262	7/5/1834		Forgery	10 yrs	3	Can read and write	Pan
14	Nathan	Adams (ali. Lewis)	36			263	7/5/1834		Forgery	8 yrs	2	Can read and write	No:
15	Charles	Johnson	28			264	7/5/1834		Passing counterfeit money	4 yrs	1	Can read and write	Dis
16	James	Stewart	50			265	7/5/1834		Selling counterfeit money	8 yrs	2	Reads and writes	Tim

Spreadsheet data created from digitized admission book, showing how text was divided and columns were titled.

Figure 2.

From the point of view of the APS Library data initiatives team, the CSV files themselves are the finished product. The Open Data Initiative of the APS Library aims to increase access to computational data by identifying material in the collection that “is conducive to being reconfigured as datasets” [APSb n.d.]. In the case of this project, however, we built example visualization tools as a means of promoting the use of the data. In late 2017, the library launched a series of visualization tools grouped together under the title “Eastern Apps: Visualizing Historic Prison Data” [APSc n.d.]. A suite of three visualization apps, “Eastern Apps” serves as an example of the use of collections as data for digital humanities projects, and has served to help the APS Digital Scholarship Center promote the use of other datasets to digital humanists [Ziegler and Marti 2019].[11] 13

As the examples in this section highlight, collections as data in cultural institutions can take many forms. Metadata enables the re-imagining of art collections, as in the case with the Carnegie Museum of Art. Digitized objects allow use and reuse for both scholarly and imaginative purposes. Collection content can be transcribed as structured open data allowing for analysis, visualizations, access through innovative interfaces, and new types of research. However, there is always interpretive work that needs to be done when creating datasets from cultural heritage material. 14

Data Brokers

In June 2016, HBO’s *Last Week Tonight With John Oliver* purchased nearly 15 million dollars of debt for 60 thousand dollars. The purchase was part of an exposé on the debt buying business, what John Oliver called “a grimy business.” A large part of the griminess of the business is the ability to buy and sell personal information about individuals. Names, addresses, social security numbers and other information is passed from one buyer to the next, often emailed in spreadsheets. The buyers hope to pressure the named individuals into paying the debts, creating a profit for the debt buyer. While the *Last Week Tonight* exposé focused on debt buying. The business of buying and selling personal information extends much further. 15

In March 2014, CBS’ *60-Minutes* aired an episode on data brokers. “Every piece of data about us now seems to be worth something to somebody,” said Tim Sparapani during the show. “And lots more people are giving up information about people they do business with, from state Departments of Motor Vehicles, to pizza parlors” [CBS 2018]. The data from different sources are gathered together to form dossiers. “The dossiers are about individuals,” one interviewee continued. “That's the whole point of these dossiers. It is information that is individually identified to an individual or 16

linked to an individual" [CBS 2018]. This information can then be used to identify individuals with health issues, significant debt, and individuals who suffer from addictions. This data is then sold to potential employers, other data brokers and increasingly to credit monitoring companies and law enforcement.

17 In 2014, the FTC released a report on data brokers that highlighted how little we know about the data being collected, and how much data there is about us. Because most of the data gathered about us do not come directly from us, most people do not fully grasp the amount of information data brokers collect and sell. "Some of the information data brokers collect, like bankruptcy information, voting history, consumer purchase data, web browsing activities and warranty registrations" are gathered from other sources [Federal Trade Commission 2014]. This data is used to put us in categories that make the data easier to market to other companies. "Potentially sensitive categories include those that primarily focus on ethnicity and income-levels, a consumer's age, or health-related conditions like 'Expectant Parent,' 'Diabetes Interest,' and 'Cholesterol Focus'" [Federal Trade Commission 2014].

18 The practice of data brokerage is secretive, and there is often no way to appeal incorrect information. The profiles these companies assign to us are often incorrect. "In the world of data brokers, you have no idea who all has bought, acquired or harvested information about you, what they do with it, who they provide it to, whether it is right or wrong or how much money is being made on your digital identity," writes Kalev Leetaru, of his efforts to determine who is making money from his information. "Nor do you have the right to demand that they delete their profile on you" [Leetaru 2018]. In the case of Leetaru, the companies got many things wrong, including his age. In 2017 writer Caitlyn Renee Miller bought information about herself from a data broker only to find that "nearly 50 percent of the data in the report about me was incorrect" [Miller 2017].

19 In 2017, Equifax announced a data breach that allowed the personal information of 143 million people to be stolen. In 2018, Facebook announced that the data analysis company Cambridge Analytica used personal data in ways that easily match John Oliver's definition of grimy. The use of personal data by companies large and small to profit is an important backdrop against which to evaluate open data in cultural institutions. Equifax and Cambridge Analytica are not, technically, data brokers. The former is a credit monitoring company and the latter, before declaring bankruptcy [Confessore and Rosenberg 2018], was a data analysis firm. However, the role of data brokers, companies that buy, combine, and sell personal information to other companies, is instructive for our purposes, and the shades of differences can be grouped together for our purposes.

20 Data brokers, along with credit monitoring companies and data analytic companies, benefit from the information of other people. This information often harms individuals through the categories they create. Examples such as "'single mom struggling in an urban setting' or 'people who did not speak English and felt more comfortable speaking in Spanish' or 'gamblers'" [Naylor 2016]. Categories such as these, which are created, populated, and shared by data brokers, make life harder for individuals.[12] The data are also used to construct systems that target and harm individuals. Advances in predictive policing, civilian surveillance, and backlashes against activists are all the outcomes of systems built on data shared by and among companies representing us in categories we cannot appeal [Winston 2018] [Feldman 2018] [Waldman et al. 2018] [Levin 2018].

21 Data brokers offer an important example of one way of interacting with data. This is an example against which we should compare ourselves when we release data in cultural institutions and use data for digital humanities. We should always ask ourselves, are we better than data brokers?

Are We Better Than Data Brokers?

22 Data brokers profit from other people's information. Those described in their datasets often have no way of knowing how they are being represented, and have no way of questioning or correcting this representation. As data becomes more prevalent in cultural institutions, and many of us — through publishing papers and presenting on our work — benefit from data about other people, now is good time to evaluate ourselves in relation to data brokers. This section explores examples of harm done by institutions as they represent individuals and groups.

23 Identifying specific cases of harm can be difficult. For this reason, it is common to focus on groups that are historically

marginalized, and who have reason to be suspicious of their representation in mainstream culture. Anyone who has ever had a reason to fear categorization by a dominant culture can more easily understand the power of data. Many groups have reason to be suspicious. For the purposes of this paper, however, examples will focus on African American representations. This decision is meant to both draw attention to the unique position of the African American community as a marginalized group and to honor the important work of generations of scholars who struggle to educate the dominant culture about these and related dangers.

24 Are people harmed by the data that we have and share? It is not standard practice for cultural institutions to share social security numbers, credit card numbers, or other sensitive personal information in either physical records, digitized facsimiles, or datasets.[13] As such, there is a significant difference between us and the work done by data brokers. Even with standardized practices in place to protect individuals, cultural institutions, historically, have done a form of harm to groups though representation. If we do not take this seriously now, we are likely to compound the problem through our open data.

25 Writing about the role of the media in enforcing negative representations of African Americans, and thus the media's culpability in historic lynchings, Sherrilyn Ifill writes, "[t]he failure to report on the 'ordinariness' of the black people's lives … undermined the ability of whites to see their black neighbors, servants, and laborers as human beings" [Ifill 2007, 168]. She continues:

> Whites could at best ignore the conditions in which most blacks lived and at worst develop a sense that blacks did not lead normal lives in which education, work and family were paramount and central. Instead, blacks could be seen as 'other,' 'different,' not possessed of the same humanity as whites ... The complicity of ordinary whites, who stood and watched a lynching without interfering, was made possible by the dehumanizing choices the media made in their coverage of blacks. [Ifill 2007, 168]

And the representations of one group of another, in this case the representation of African Americans by predominantly Caucasian media institutions, continues to affect our society. "The lingering remnants of these dehumanizing portrayals of blacks in the media," writes Ifill, bringing the issue to the present, "have modern currency," including over-incarceration of African American males, hyper policing of black communities, and police brutality [Ifill 2007, 168–9]. Many of these issues are often ignored by white communities because of a history of media representation of African Americans.

26 The representation of one group by another group can range from obvious fiction to pretense of objective truth. "All groups tell stories," writes David Pilgrim, founding curator of the Jim Crow Museum, "but some groups have the power to impose their stories on others, to label others, stigmatize others, paint others as undesirables, and to have these social labels presented as scientific fact, God's will or wholesome entertainment" [Pilgrim 2017, 8]. The types of stories matter. It also matters which type of institution is telling the story. "When we watch movies or read novels," continues Pilgrim, "we know that they are stories; we identify the characters, follow the plot and anticipate the conclusion. But there are other stories that are not so easily identified — sometimes they masquerade as object, race-neutral truth" [Pilgrim 2017, 9]. To illustrate this point, Pilgrim investigates the use of pseudo-science to justify racist beliefs and actions. Scientific institutions were used, during slavery and Jim Crow, to promote and legitimize racism.

27 In what way are cultural institutions doing the same? The collecting practices of cultural institutions have long been marred by the racial bias of the archivists and curators who build collections. The decisions made about what is collected are colored by the opinions of those doing the collecting and this has tipped the scales on how African Americans are represented in archives. An overt example from the author's own institution is a case in point. In 1945, the LSU Department of Archives and Manuscripts (a pre-curser to the current Special Collections) was offered the opportunity to acquire the collection of African American bibliophile and book collector, Henry P. Slaughter. The quality of the collection was endorsed by archivist Herbert Kellar and book dealer Forest Sweet who wrote, "it is no sense a collection to be filed away - it is rather a collection to be worn out with legitimate use for what it can offer as a basis of study of the negro problem" [Sweet 1945].

Despite this endorsement, the University eventually passed on the acquisition on the advice of Archivist and History Professor, Edwin Davis:[14] “the collection has been selected with the plan to emphasize the Negro’s point of view of the race problem. If this is true it is my opinion that the collection will have a considerable amount, perhaps an appreciable percentage, of what might be termed weak material. I have however, no evidence for this opinion. It might prove to be a very valuable collection” [Davis 1946]. Davis, based only on his assumption that any collection that centers an African American perspective has minimal value, declined the collection.[15] Rarely does such documentation, but similar decisions are common throughout collecting institutions. 28

Decisions about what material cultural institutions collect have long term repercussions that are felt for generations. In March 2019, Tamara Lanier filed a lawsuit against Harvard University over ownership of a daguerreotype photograph of an enslaved ancestor named Renty and his daughter, Delia. The photographs were commissioned by Professor Louis Agassiz and utilized as evidence of inferiority of the African American race. The photographs were rediscovered in 1976 hidden away in the attic of a campus museum. Since then the University has loaned the photographs to other museums but also limited the use of the images by researchers due to their “sensitive” nature [Hartocollis 2019] [Schuessler 2019]. 29

What is collected matters, and so does its description. The role of the library catalog in reinforcing dominant points-of-view has been explored many times.[16] Melissa Adler, for example, draws attention to the “ways that sections of library classifications were constructed based on ideas about African Americans” [Adler 2017, 5]. Adler traces the creation of the library classifications as they map to racist ideas active in the late 19th and early 20th centuries. The library catalog freezes these ideas in place, Adler claims, and makes them look natural and obvious, similar to the manner of the pseudo-science Pilgrim describes. 30

The business of gathering, combining, and selling data is not new. However, the scope of surveillance and reach of the systems created with the data is unprecedented due to new tools and methods. In the same way, library catalogs have long represented groups of people in problematic ways. What is different is the new tools and methods we are using to promote the use and reuse of these descriptions and collections. The collections as data framework in cultural institutions carries with it the possibility for our descriptions of people to be shared, combined with other data, and used to negatively affect groups. The ability to exert “control over group and personal identity and memory,” writes Noble, “must become a matter of concern for archivists, librarians, and information workers” [Noble 2018, 172]. This is all the more urgent as we look toward a future of increased share-ability and data-oriented services. Now is the time to ask how we can be better than data brokers. 31

How Can We Be Better?

This section posits three possible directions that are still in the early phases of exploration. These possibilities are listed as sincere efforts to investigate possible steps toward a future in which my work with data feels less grimy. As white librarians, we work in a field that has long struggled to be inclusive to historically marginalized communities.[17] We hope to implement these steps as experiments at our current institution, and to join a community of practitioners exploring these topics in different settings. These experiments are meant to inform the practice of collections as data — as it relates to metadata, digital facsimiles, and structured data — as well as the practice of digital humanities. These approaches are informed by literature on empathy, data science, and critical librarianship. 32

Empathy and Description

The grimy business of data brokers is legal.[18] While it is important that cultural institutions follow laws when we rethink our collections as data, and when we use this data to build digital humanities projects, this alone will not be enough. For example, the Health Insurance Portability and Accountability Act of 1996 (HIPPA) provides privacy protection for health-related data (Health and Human Services 2015), and, as we saw above, it is established practice in cultural institutions not to share sensitive financial or personal information about people represented in the collection. However this is not enough. A thin legalistic understanding of what should and should not be shared will never be robust enough to ensure we do not harm people we represent. There are no laws, of course, against harmful library subject headers or 33

terminology.

Arguing for a move away from thin legalistic frameworks, Michelle Caswell and Marika Cifor explore the role of feminist ethics in reconceptualizing the role of archives and the people represented in our collections. This approach is applicable for all cultural institutions, collections as data work, and digital humanities projects that use this data. “In a feminist ethics approach,” they write, “archivists are seen as caregivers, bound to records creators, subjects, users, and communities through a web of mutual affective responsibility” [Caswell and Cifor 2016, 23]. By caring about the subjects of records — the people represented in the records, who are “counted, classified, studied, enslaved, traded as property, and/or murdered” [Caswell and Cifor 2016, 36] — we can make decisions about the records based on our empathy with the people described. Caswell and Cifor write: 34

> [A] feminist approach guides the archivist to an affective responsibility to empathize with the subjects of the records and, in so doing, to consider their perspectives in making archival decisions. This is in contrast to the dominant West mode of archival practice, in which archivists solely consider the legal rights of records creators … In the feminist approach, the archivist cares about and for and with subjects; she empathizes with them. [Caswell and Cifor 2016, 36]

Even in situations where no laws exist to stop the sharing of information, we can ask ourselves: What would the people described in the records think of this project or representation? If the people described in our data could see our project, would we be as eager to work on this project?

Asking For Help

If we would work differently when those represented in our work can see it, how could we ensure that this happens? The particular individuals described in datasets by cultural institutions are likely to be deceased, as is the case for the historic prison records described above. However, we can work with members of affected communities. 35

Safiya Noble, writing about the shortcomings of Google and other tech companies, calls for the combination of technology and critical studies. “We need people designing technologies for society to have training and an education on the histories of marginalized people, at a minimum,” Nobel writes, “and we need them working alongside people with rigorous training and preparation from the social sciences and humanities” [Noble 2018, 70]. Applying this idea to the cultural institutions creating and releasing data, and to digital humanist using the data, we could bring in experts in African American Studies, for example, if our data represents the African American community, or Women and Gender Studies, if applicable. The expertise from their subject could be brought to bear within the cultural institution creating the data and within the digital humanities group working with the data. In short, we can ask for help.[19] 36

Many cultural institutions are unlikely to be able to create new professional positions specifically for individuals trained in the histories of marginalized communities. However, it is critical that we find ways to pay the scholars whose help we seek. In academic libraries and archives this might take the form of graduate assistantships. Digital humanists will likely be similarly restricted. However, including these roles in project plans and grant applications is one way to normalize the process of asking for help. Money is not the only incentive; this could be a valuable means of exposing scholars to possible careers in cultural institutions and digital humanities. However, monetary compensation is an important part of letting scholars know we take their expertise seriously. 37

Taking Suggestions

As we have seen, it can be very difficult to know what information data brokers have about us, and to correct what is incorrect.[20] Is there any reason why cultural institutions cannot do better? We could include feedback channels for the systems we create, and we could make our decisions as transparent as possible. We can bring in experts in the history of marginalized communities, as discussed above. We can also take a step further by ensuring that anyone — those we identify as experts and invite into the process, and those we do not — has a chance to ask us about our decisions and suggest alternate practices. We could ensure that every data project includes feedback options for people to ask questions and make suggestions. After all, we will not be able to invite everyone to work with us, but we could try to 38

include everyone who is affected by our representations.

Explaining Ourselves

39 We can also include context in downloadable data. For example, the CSV files that constitute the core of the historic prison data project, described above, have the transcriptions of the admission books. No metadata is included in the download, and no context. Instead of simple files containing structured transcriptions, we could use data packets to bundle contextualizing information together with the transcriptions.[21] As we release collections as data we could standardize the inclusion of context; as we promote data from cultural heritage institutions, we could normalize the process of incorporating the context supplied within data packages.

40 Part of the context we could add would be explanations about decisions that we make while creating data. A danger of institutionally-generated collections as data is the perception of objectivity. Devon Mordell, writing about this danger, proposes we frame collections data work in cultural heritage institutions within a new paradigm, a collections-as-data paradigm, that considers both the conceptual and practical concerns related to the use of data in archives [Mordell 2019]. The benefits of framing a paradigm around collections-as-data, Mordell argues, is to “ensure that a social justice critique is maintained within” the emerging work related to collections as data [Mordell 2019, 147]. In using archival material as data, there is a risk that the archival holdings and descriptions of the holdings will look objective and natural, and the work of archivists and others to show how archival collections are never neutral and natural will be obscured. Mordell argues that “active participation and critical discourse” around the tools and practices is needed to ensure that new technologies reinscribe a false since neutrality [Mordell 2019, 156]. Following Mordell, we can include context that encourages critical discourse around our data.

41 These four directions are presented as a means to begin conversations about practical implementation toward the goal of ensuring that the work we do is better than data brokering. There is still much to learn, and the implementation of any of these proposed solutions will certainly open new possibilities, of both success and failure. Many of the issues we hope to address have taken decades, and sometimes centuries, to create. There will be no quick solution.

Conclusion

42 The examination of data brokers is meant to offer a warning to those of us working in cultural heritage institutions. However, it is not just one example taken at random. Rather it is a generative lens to view our work because we are always already existing as data in the world of data brokers. The power that data brokers have over us to collect, analyze, describe, and sell our data should bring into sharp focus the power that we have over those with whose data we work.

43 Thinking about collections as data framework creates an ideal moment for a reflection on the creation and use of data in cultural institutions and digital humanities. This reconceptionalization enables unprecedented access and interaction with collection material in libraries, archives, and museums, i“ncluding but not limited to text mining, data visualization, mapping, image analysis, audio analysis, and network analysis” [AAC 2018b].

44 This is also a moment to consider how we do not want to interact with data. Having a ready example against which to judge our behavior is useful, and data brokers provide a perfect use case. Companies that buy, combine, and resell personal data to the detriment of individuals and groups can be the example we need. Defining ourselves in contrast to data brokers also grants us the opportunity to reflect on the historically problematic aspects of cultural institutions’ descriptive practices.

Notes

[1] Cultural institutions, for the purpose of this paper, are libraries, archives and museums. These institutions collect, describe, and make available material of cultural significance such as works of art, historical documents, and published material.

[2] Because the phrase collections as data is used to refer to both the conceptual practice of rethinking the role of collection material and the team active in the IMLS grant, the name AAC will be used for the latter.

[3] The author is thankful to Thomas Padilla for reviewing an early draft of this section for accuracy; any remaining errors or misrepresentations are the fault of the author.

[4] “A facet documents a collections as data implementation. An implementation consists of the people, services, practices, technologies, and infrastructure that aim to encourage computational use of cultural heritage collections.” see “A Release and a Call - Collections as Data Facets”. Always Already Computational - Collections as Data. Accessed May 17, 2018. https://collectionsasdata.github.io/facets/.

[5] Historically important for many reasons, the Eastern State Penitentiary championed the standardized use of “cellular isolation” in which each inmate spends all of his or her time in a cell by themselves. For more information, see the APS finding aid: https://search.amphilsoc.org/collections/view?docId=ead/Mss.365.P381p-ead.xml, and the website of the Eastern State Historic Site: https://www.easternstate.org.

[6] The digitization team for this project included Grace DiAgostino and Bayard Miller.

[7] On the uses of digitized art as data, see: “The Next Rembrandt”. The Next Rembrandt. Accessed May 29, 2018. https://www.nextrembrandt.com. Hristova, Stefka. “Images as Data: Cultural Analytics and Aby Warburg’s Mnemosyne”. *International Journal for Digital Art History* 0, no. 2 (October 18, 2016). https://doi.org/10.11588/dah.2016.2.23489.

[8] The transcription team consisted of Michelle Ziogas, Bayard Miller, and Kristina Frey

[9] See, for example, the “Conventions Used” section of the Github Read Me file. “Historic Prison Data” American Philosophical Society Github Repository, https://github.com/AmericanPhilosophicalSociety/Historic-Prison-Data#conventions-used

[10] For concerns about relying on transcriptions, see James H. Merrell, “‘Exactly as they appear’: Another Look at the Notes of a 1766 Treason Trial in Poughkeepsie, New York, with Some Musings on the Documentary Foundations of Early American History”, *Early American Studies* 12, no. 1 (Winter 2014): 202-237. See also Jacqueline Wernimont’s thoughts on the role of tabular data and the loss of understanding: https://digital-frontiers.org/conference/2017/texas-digital-library-closing-keynote-address-counting-dead-quantum-media-and-how-we, as well as Foreman, P. Gabrielle, and Labanya Mookerjee, “Computing in the Dark: Spreadsheets, Data Collection and DH’s Racist Inheritance” in *Always Already Computational Position Statements*, https://collectionsasdata.github.io/resources/, accessed 5/18/2018

[11] A growing selection of open data sets can be found on the APS Library Open Data page, “Open Data | APS Digital Library”. Accessed June 5, 2018. http://diglib.amphilsoc.org/data.

[12] Quantifying the harm is difficult due to the secretive nature of data transfer and use. The United Kingdom-based non-profit Privacy International is one of many drawing attention to this harm. “Video: What Is Data Exploitation?” Privacy International. Accessed May 29, 2018. http://privacyinternational.org/video/1626/video-what-data-exploitation.

[13] See, for example, “Questions and Answers on Privacy and Confidentiality”. *Text. Advocacy, Legislation & Issues*, May 29, 2007. http://www.ala.org/advocacy/privacy/FAQ. Software for digital preservation systems, such as ePadd for email, are also built with these issues in mind, see “EPADD User Guide 5.0”. Google Docs, May 2017. https://docs.google.com/document/d/1ZMuWU0z-IVsk80_IUEYMfVrwfCsS1bp0sjL28GBGcMU

[14] Herbert A. Kellar was the Director of the McCormick Historical Association and a founding member of the Society of American Archivists; Edwin Davis was the first archivist hired to manage the LSU Department of Archives and Manuscripts

[15] The collection is currently housed at the Atlanta University Center Robert W. Woodruff Library Archives Research Center. See:http://findingaids.auctr.edu/repositories/2/resources/62. The author wishes to thank Jenny Mitchell, Head of Manuscript Processing at LSU Libraries, for this example.

[16] See, for example, Berman, Sanford.Prejudices and Antipathies : A Tract on the LC Subject Heads Concerning People. McFarland & Co., 1993. Olson, Hope. “Mapping Beyond Dewey’s Boundaries: Constructing Classificatory Space for’ Marginalized Knowledge Domains”. *Library Trends* 47, no. 2 (Fall 1998): 233–54. Berman, Sanford. *Prejudices and Antipathies: A Tract on the LC Subject Heads Concerning People*. Adler, Melissa. “Classification Along the Color Line: Excavating Racism in the Stacks”. *Journal of Critical Library and Information Studies* 1, no. 1 (January 29, 2017). https://doi.org/10.24242/jclis.v1i1.17. “View of Engaging an Author in a Critical Reading of Subject Headings”. Accessed May 22, 2018. http://libraryjuicepress.com/journals/index.php/jclis/article/view/20/12.

[17] See, by way of introduction, Hudson, David James. “On Diversity as Anti-Racism in Library and Information Studies: A Critique”. *Journal of Critical Library and Information Studies* 1, no.1 (2017). DOI: 10.24242/jclis.v1i1.6. jesus, nina de. “Locating the Library in Institutional

Oppression – In the Library with the Lead Pipe". Accessed May 30, 2018. /2014/locating-the-library-in-institutional-oppression/. Galvan, Angela. "Soliciting Performance, Hiding Bias: Whiteness and Librarianship – In the Library with the Lead Pipe". Accessed May 30, 2018. /2015/soliciting-performance-hiding-bias-whiteness-and-librarianship/.

[18] At the time of this writing, the European Union's General Data Protection Regulations (GDPR) is just coming into effect, and while companies large and small that collect data are adjusting some behaviors the outcome on the core practice of data brokerage in the United States is yet to be determined. Ong, Thuy. "Microsoft Expands Data Privacy Tools Ahead of GDPR". The Verge, May 24, 2018. https://www.theverge.com/2018/5/24/17388206/microsoft-expand-data-privacy-tools-gdpr-eu. Tiku, Nitasha. "How Europe's New Privacy Law Will Change the Web, and More". WIRED, March 19, 2018. https://www.wired.com/story/europes-new-privacy-law-will-change-the-web-and-more/.

[19] It's worth noting here that this can easily go awry. There's a history of white people asking black people to explain why racism. For example, soon after the 2016 presidential election,Slate held a forum for African American writers to reflect on an increased demand of this type.Bouie, Jamelle, Gene Demby, Aisha Harris, Tressie McMillan Cottom, and Chau Tu. "I'm Not Your Racial Confessor". *Slate*, December 6, 2016. http://www.slate.com/articles/news_and_politics/politics/2016/12/the_black_person_s_burden_of_managing_white_emotions_in_the_age_of_trump.html. See also: Johnson, Theodore R. "How Black Writers Can Help White Readers". *The New Republic*, December 29, 2016. https://newrepublic.com/article/139541/black-writers-can-help-white-readers.

[20] Even attempts at transparency by data brokers often prove to be misleading. See: Insider, Business. "The Site That Shows You All The Personal Data Advertisers Have On You Isn't Entirely Accurate". *Business Insider*. Accessed May 25, 2018. http://www.businessinsider.com/acxiom-about-the-data-problems-2013-9. Singer, Natasha. "Acxiom Lets Consumers See Data It Collects". *The New York Times*, September 4, 2013, sec. Technology. https://www.nytimes.com/2013/09/05/technology/acxiom-lets-consumers-see-data-it-collects.html. Breen, Bant. "AboutTheData: Data Collection Gone Wrong". *Huffington Post* (blog), September 26, 2013. https://www.huffingtonpost.com/bant-breen/aboutthedata-data-collect_b_3998252.html.

[21] See, for example, Walsh, Paul, and Rufus Pollock. "Data Package". *Frictionless Data*. Accessed May 29, 2018. https://frictionlessdata.io/specs/data-package/. For a discussion on using data packets in cultural institutions, Averkamp, Shawn. "Data Packaging Guide", May 14, 2018. https://github.com/saverkamp/beyond-open-data.

Works Cited

AAC 2018a Always Already Computational - Collections as Data. 2018." Always Already Computational". Accessed June 5, 2018. https://collectionsasdata.github.io/

AAC 2018b Always Already Computational - Collections as Data. 2018. "The Santa Barbara Statement on Collections as Data Version 2". Accessed June 5, 2018. https://collectionsasdata.github.io/statement/

APSa n.d. American Philosophical Society. n.d. "Historic Prison Data". https://github.com/AmericanPhilosophicalSociety/Historic-Prison-Data

APSb n.d. American Philosophical Society. n.d. "Open Data". Accessed June 5, 2018. http://diglib.amphilsoc.org/data

APSc n.d. American Philosophical Society. n.d. "Eastern Apps: Exploring Historic Prison Data". Accessed June 5, 2018. https://diglib.amphilsoc.org/labs/eastern-apps/

Adler 2017 Adler, Melissa. "Classification Along the Color Line: Excavating Racism in the Stacks". *Journal of Critical Library and Information Studies* 1, no. 1 (January 29, 2017). https://doi.org/10.24242/jclis.v1i1.17.

CBS 2018 CBS News. 2018. "The Data Brokers: Selling Your Personal Information". Accessed June 5, 2018. https://www.cbsnews.com/news/the-data-brokers-selling-your-personal-information/.

Carnegie n.d. Carnegie Museum of Art. n.d. "Carnegie Museum of Art's Collection Dataset". Accessed June 5, 2018. https://github.com/cmoa/collection

Caswell and Cifor 2016 Caswell, Michelle and Marika Cifor. 2016. "From Human Rights to Feminist Ethics: Radical Empathy in the Archives", *Archival Science* 81 (Spring 2016): 23-43.

Confessore and Rosenberg 2018 Confessore, Nicholas, and Matthew Rosenberg. 2018. "Cambridge Analytica to File for Bankruptcy After Misuse of Facebook Data". *The New York Times*, May 3, 2018, sec. U.S. https://www.nytimes.com/2018/05/02/us/politics/cambridge-analytica-shut-down.html.

Davis 1946 Davis, Edwin. Letter to Guy Lyle. 12 January 1946. LSU Department of Archives and Manuscripts records,

A1506, University Archives, LSU Libraries Special Collections, Baton Rouge, La.

Federal Trade Commission 2014 Federal Trade Commission. 2014. “FTC Report Examines Data Brokers”. Accessed June 5, 2018. https://www.consumer.ftc.gov/blog/2014/05/ftc-report-examines-data-brokers.

Feldberg 1980 Feldberg, Michael. 1980. *The Turbulent Era: Riot & Disorder in Jacksonian America*. New York: Oxford University Press.

Feldman 2018 Feldman, Noah. 2018. “The Future of Policing Is Being Hashed Out in Secret”. *Bloomberg*, February 28, 2018. https://www.bloomberg.com/view/articles/2018-02-28/artificial-intelligence-in-policing-advice-for-new-orleans-and-palantir.

Getty Trust n.d. J. Paul Getty Trust. n.d. “Open Content”. Accessed June 5, 2018. https://www.getty.edu/about/whatwedo/opencontent.html.

HHS 2015 Health and Human Services. 2015. “Health Information Privacy”. Accessed June 5, 2018. https://www.hhs.gov/hipaa/index.html

Hartocollis 2019 Hartocollis, Anemona. “Taking On Harvard Over Rights to Slave Photos”. New York Times, 21 Mar. 2019, p. A1(L). Gale In Context: Biography, https://link.gale.com/apps/doc/A579490887/BIC?u=lln_alsu&sid=BIC&xid=c62f7942. Accessed 5 Dec. 2019.

IMLS 2016 Institute of Museum and Library Services. 2016. “LG-73-16-0096-16”. Accessed June 6, 2018. https://www.imls.gov/grants/awarded/lg-73-16-0096-16.

Ifill 2007 Ifill, Sherrilyn A. 2007. *On the Courthouse Lawn: Confronting the Legacy of Lynching in the Twenty-First Century*. Boston: Beacon Press.

Koster 2018 Koster, Lukas and Saskia Woutersen-Windhouwer. “FAIR Principles for Library, Archive and Museum Collections: A Proposal for Standards for Reusable Collections”. *The Code4Lib Journal*, May 7, 2018. http://journal.code4lib.org/articles/13427.

Leetaru 2018 Leetaru, Kalev. 2018. “The Data Brokers So Powerful Even Facebook Bought Their Data - But They Got Me Wildly Wrong”. *Forbes*, April 5, 2018. https://www.forbes.com/sites/kalevleetaru/2018/04/05/the-data-brokers-so-powerful-even-facebook-bought-their-data-but-they-got-me-wildly-wrong/#92a30923107a.

Levin 2018 Levin, Sam. 2018. “Black Activist Jailed for His Facebook Posts Speaks out about Secret FBI Surveillance”. *The Guardian*, May 11, 2018, sec. World news. http://www.theguardian.com/world/2018/may/11/rakem-balogun-interview-black-identity-extremists-fbi-surveillance.

Library of Congress 2016 Library of Congress. 2016. “Collections as Data 2016”. Accessed June 5, 2018. http://digitalpreservation.gov/meetings/dcs16.html.

Library of Congress 2017 Library of Congress. 2017. “Collections as Data IMPACT 2017”. Accessed June 5, 2018. http://digitalpreservation.gov/meetings/asdata/impact.html.

Library of Congress 2019 Library of Congress. 2019. “Library Receives $1M Mellon Grant to Experiment with Digital Collections as Big Data”. Accessed 7 Nov. 2019. https://www.loc.gov/item/prn-19-098/library-receives-1m-mellon-grant-to-experiment-with-digital-collections-as-big-data/2019-10-04/.

Lied Library Lied Library. “Collections as Data National Forum 2”. Filmed May 7-8 at the University of Nevada Las Vegas. Accessed June 5, 2018. https://www.youtube.com/watch?v=ENaPV2XmO9I.

Lill n.d. Lill, Jonathan. “The Museum of Modern Art Exhibition Index”. Accessed June 5, 2018. https://collectionsasdata.github.io/facet12/.

Lillehaugen and Zarafonetis n.d. Lillehaugen, Brook and Michael Zarafonetis. “Ticha: A Digital Text Explorer for Colonial Zapotec”. Accessed June 5, 2018. https://collectionsasdata.github.io/facet12/.

Miller 2017 Miller, Caitlyn Renee. 2017. “I Bought a Report on Everything That’s Known About Me Online”. *The Atlantic*, June 6, 2017. https://www.theatlantic.com/technology/archive/2017/06/online-data-brokers/529281/.

Mordell 2019 Mordell, Devon. 2019. “Critical Questions for Archives as (Big) Data”. *Archivaria*, vol. 87, no. 0, May 2019, pp. 140–61.

National Library of Scotland 2019 National Library of Scotland. 2019. “Launch of Data Foundry Website”. https://www.nls.uk/news/archive/2019/09/data-foundry. Accessed 7 Nov. 2019.

Naylor 2016 Naylor, Brian. 2016. “The Data Brokers: Selling Your Personal Information”. CBS News, July 11, 2016. https://www.cbsnews.com/news/the-data-brokers-selling-your-personal-information/.

Newbury and Fowler n.d. Newbury, David, and Daniel Fowler. n.d. “Carnegie Museum of Art Collection Data”. Accessed June 5, 2018. https://collectionsasdata.github.io/facet2/.

Noble 2018 Noble, Safiya Umoja. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: New York University Press.

Open Knowledge International n.d. Open Knowledge International. n.d. “The Open Definition - Open Definition - Defining Open in Open Data, Open Content and Open Knowledge”. Accessed June 5, 2018. https://opendefinition.org/.

Padilla 2017 Padilla, Thomas. 2017. “On a Collections as Data Imperative”. UC Santa Barbara. Retrieved from https://escholarship.org/uc/item/9881c8sv.

Pilgrim 2017 Pilgrim, David. 2017. *Watermelons, Nooses, and Straight Razors: Stories from the Jim Crow Museum*. Oakland, CA: PM Press.

Schuessler 2019 Schuessler, Jennifer. “Portraits Of Slaves Require Moral Lens”. *New York Times*, 23 Mar. 2019, p. C1(L). Gale In Context: Biography, https://link.gale.com/apps/doc/A579778374/BIC?u=lln_alsu&sid=BIC&xid=613dc968. Accessed 5 Dec. 2019.

Sweet 1945 Sweet, Forest. Letter to Edwin Davis. 7 January 1945. LSU Department of Archives and Manuscripts Records, #A1506, University Archives, LSU Libraries Special Collections, Baton Rouge, La.

University of Pennsylvania n.d. University of Pennsylvania. n.d. “OPenn: Read Me”. Accessed June 5, 2018. http://openn.library.upenn.edu/ReadMe.html.

Waldman et al. 2018 Waldman, Peter, Lizette Chapman, and Jordan Robertson. 2018. “Palantir Knows Everything About You”. *Bloomberg*, April 19, 2018. https://www.bloomberg.com/features/2018-palantir-peter-thiel/.

Weber 2017 Weber, Chela Scott. “Research and Learning Agenda for Archives, Special, and Distinctive Collections and Research Libraries”. OCLC Research, 2017. https://doi.org/10.25333/c3c34f.

Winston 2018 Winston, Ali. 2018. “A Pioneer in Predictive Policing Is Starting a Troubling New Project”. *The Verge*, April 26, 2018. https://www.theverge.com/2018/4/26/17285058/predictive-policing-predpol-pentagon-ai-racial-bias.

Zarafonetis and Horowitz n.d. Zarafonetis, Michael and Sarah M. Horowitz. “Beyond Penn’s Treaty”. Accessed May 5, 2018. https://collectionsasdata.github.io/facet11/.

Ziegler and Marti 2019 Ziegler, S. L. and Steve Marti. 2019. “A Hidden Gem Become a Fertile Mining Ground: Historic Prison Admission Books and Data-Driven Digital Projects”. *The Pennsylvania Magazine of History and Biography*. Vol. 143: 3. Pp. 363-373.

Ziegler and Ziogas 2016 Ziegler, Scott and Michelle Ziogas. 2016. “Eastern State Penitentiary Admission Book B”, - . 21. Philadelphia, PA: McNeil Center for Early American Studies [distributor], 2016. https://repository.upenn.edu

Ziegler n.d. Ziegler, Scott. “American Philosophical Society Open Data Projects”. Accessed June 5, 2018. https://collectionsasdata.github.io/facet4/.