2022

Just Because We Can Doesn't Mean We Should: On Knowing and Protecting Data Produced by the Jewish Consumptives' Relief Society

Jack M. Maness
University of Denver

Kim Pham
University of Denver

Just Because We Can Doesn’t Mean We Should: On Knowing and Protecting Data Produced by the Jewish Consumptives’ Relief Society

Publication Statement

This article was originally published as:

Maness, J. & Pham, K. (2022). Just because we can doesn't mean we should: On knowing and protecting data produced by the Jewish Consumptives' Relief Society. Humanist Studies & the Digital Age 7(1). http://dx.doi.org/10.5399/uo/hsda/7.1.8

Humanist Studies & the Digital Age, 7.1 (2022)
ISSN: 2158-3846 (online)
http://journals.oregondigital.org/hsda/
DOI: 10.5399/uo/hsda/7.1.8

Just Because We Can Doesn't Mean We Should: On Knowing and Protecting Data Produced by the Jewish Consumptives' Relief Society

Jack Maness and Kim Pham

Abstract: A recent project at the University of Denver Libraries used handwritten text recognition (HTR) software to create transcriptions of records from the Jewish Consumptives' Relief Society (JCRS), a tuberculosis sanatorium located in Denver, Colorado from 1904 to 1954. Among a great many other potential uses, these type- and hand-written records give insight into the human experience of disease and epidemic, its treatment, its effect on cultures, and of Jewish immigration to and early life in the American West. Our intent is to provide these transcripts as data so the text may be computationally analyzed, pursuant to a larger effort in developing capacity in services and infrastructure to support digital humanities as a library, and to contribute to the emerging HTR ecosystem in archival work.

Just because we can, however, doesn't always mean we should: the realities of publishing large datasets online that contain medical and personal histories of potentially vulnerable people and communities introduce serious ethical considerations. This paper both underscores the value of HTR and frames ethical considerations related to protecting data derived from it. It suggests a terms-of-use intervention perhaps valuable to similar projects, one that balances meeting the research needs of digital scholars with the care and respect of persons, their communities, and inheritors, whose lives produced the very data now valuable to those researchers.

Handwritten Text Recognition Technology and Archival Research

Historically, archival research has been a time-intensive endeavor. The sheer volume of manuscripts, photographs and other objects and artifacts in libraries, archives and museums is great enough to present significant challenges to many researchers. The challenge is compounded, of course, by the need to physically access materials that have never been digitized and rendered in machine-readable formats. In addition, limited resources in archival organizations mean many physical archives are only processed with minimal descriptive metadata: one might find, for example, from electronic records that the papers of a certain person or organization are in a particular archive, or even that the papers from some decade or another are in exact boxes, but frequently one may not know from these electronic records precisely what papers are in what boxes until reviewing them, much less what is written in those papers, or by whom. The result is that tedious and costly physical analysis remains the only access point to many archival materials.

The inherent and long-standing reality of difficulty in access to physical archives calls for the development of new technologies that ease access and make possible new forms of computational analysis of those documents. In some scientific disciplines these materials are considered "heritage," "dark," or "analog" data that are "at risk" and in need of "rescuing" (Interna'tl Science Council). Projects that make such data more readily available in the digital age are even considered for an international award in geoscience research

(Internat'l Earth Data Alliance). In all disciplines, the humanities included, untold discoveries await such projects and the technologies and infrastructures required by them, discoveries that necessitate physical documents not only be made electronic, but also that their contents, in the case of written language, be made readable by machines.

Such technical innovation has been underway for more than a generation. For many decades, primarily type-written documents have been "rescued," as it were, using optical-character-recognition (OCR). The history of OCR, in fact, can be understood as having origins two centuries ago in work to aid the visually-impaired, and that "by the end of the second decade of the 19th century . . . experimentation with OCR had convincingly demonstrated the feasibility of optically scanning printed materials, converting that material into electronic code signals, and subsequently encoding the electronic code signals" (Schantz 3).

Today, because large volumes of printed material have been, simply stated as a verb, "OCR'd," the technology "pervades contemporary humanities research and teaching . . . [and] make[s] it easy to forget that we are all engaged in relatively sophisticated modes of machine reading each time we use a digital archive" (Cordell and Smith 9). OCR is an imperfect technology, however, standards for its implementation are lacking, and its very improvement introduces opportunities for OCR'd documents to be reprocessed, yielding better and new machine-readable characters. Indeed, as Cordell and Smith conclude in their report, OCR improvement warrants collective research and standardization, due to its imperfection but also its nearly ubiquitous presence in the online services of libraries, archives, and museums.

Researchers do not currently enjoy the ubiquity of machine reading for handwritten textual materials, however, primarily for the obvious reason: handwriting is more varied and less standardized than machine-typed language, and therefore more difficult to encode into machine-readable formats. OCR technology can and does encode some handwritten text, but the nature of handwriting requires yet more sophisticated technologies that leverage artificial intelligence and machine learning, allowing algorithms to generalize character recognition enough that it can "read" across hands, but not generalize so much that it confuses characters and languages to the point of uselessness. Deploying HTR across a variety of archival infrastructures would vastly improve access to archival information and would also limit the need for human intervention to transcribe handwritten materials. Human transcription is a costly and sometimes ethically fraught process in its own right, including the use of unpaid labor, lack of diversity in volunteers, and potential reinforcement of the continued genderization of library, archival, and museum work (Orlowitz; Mayer).

While human transcription most certainly introduces elements of bias and other problematic issues, it is also vitally important to note that machine transcription does not escape them. HTR relies in some cases on "the imposition of a uniform transcription scheme [that] obscures regional, temporal, and ethnic varieties [of a language]. . . and creates an artificially homogeneous outlook," because ultimately machine facilitated transcription "[relies] on the language data provided by the transcriber [which] reproduce[s] the transcriber's linguistic bias in the output" (Kirmizialtinx and Wrisley 14). In addition to homogenizing languages and dialects, there is also a very basic availability bias inherenet in automated transcription, as machines can only trascribe characters and languages upon which they have been trained: it is somewhat safe to presume the HTR ecosystem being developed is almost certainly biased toward English and other European languages, for example.

Nevertheless, HTR and its application in humanist studies is no longer a promise but a reality. While the "first driving force behind handwritten text classification was for digit classification for postal mail," (Balci et

al. 1) it has become more widely used in banking (Ghosh et al.) to emerging applications in analyzing historical documents (Romero et. al.). And, as machine learning and neural networking advances, HTR technology "has the potential to transform access to our written past for the use of researchers, institutions and the general public [and] . . . can extend the existing research infrastructure of the archives, libraries and humanities domain" (Muehlberger et al. 955). Like OCR'd characters, these encoded handwritten texts will one day pervade digital archives, being readable and analyzable by humans and machines alike. And these encoded texts can and will be analyzed and mined in any number of ways.

Simply put, in the coming decades HTR promises to revolutionize methodologies in humanities disciplines that analyze archival materials, much as OCR has in recent decades. Large troves of transcribed materials will be available not only for search-and-retrieval purposes, but for large-scale textual analyses.

The Jewish Consumptives' Relief Society

Our project applies HTR to the records of the JCRS collection, an important archive included in the Beck Archives of Rocky Mountain Jewish History. The records of this institution, particularly the robust patient files, which include application forms and extensive correspondence, document an important era in the development of tuberculosis treatment in America and the history of organizations and the lives of individuals who built Jewish life in the Rocky Mountain region.

Tuberculosis, or "consumption" as it was also known, was the leading cause of death in late nineteenth century and early twentieth century America. No accepted standard for tuberculosis treatment prevailed in the early years, but by 1880 medical opinion emphasized fresh air for respiratory ailments. Because of its high altitude and dry and sunny climate, Colorado became so popular a treatment destination for tuberculosis victims it soon earned the nickname of "The World's Sanatorium."

A small group of Jewish working-class immigrants in Denver founded the JCRS because no publicly supported institution existed at the time. A non-sectarian organization that treated patients in all stages of the disease, over the next fifty years the JCRS provided its services free of charge to over 10,000 patients. Two notable East European Jewish physicians led the JCRS for extended periods: Dr. Charles Spivak, who served as the executive secretary until his death in 1927, and Dr. Philip Hillkowitz, who served as president until his death in 1948 (Abrams). JCRS records are among the most heavily-used collections in DU's archives, and dozens of handwritten letters and records have been digitized and made available through a digital repository for many years.

These letters, from the mundane to the poignant, detail stories that describe the immigration, growth and development of Colorado's Jewish community; of philanthropy and disease; of household and institutional finances; the art, literature and science of the time; and of daily life in this and similar institutions, where Americans and new immigrants to the nation lived, loved, suffered, and died in the tens-of-thousands. The collection is used in several courses at DU, ranging from sociology to history, wherein students might, for example, extract demographic data to explore diseases of the urban poor, or use genealogical methodologies to track prevailing immigration patterns and compare them to modern day issues of immigration. These learning opportunities contributed in part to several DU faculty members winning a national award for the use of primary sources in teaching from the in 2018 (Center for Research Libraries).

In addition to teaching, scholars have also used the collection in their published research to demonstrate modern clinical laboratory practices through the gathering of autopsy statistics (Wright and Abrams), to

foreground historical issues of immigration and the rise of Jewish hospitals in the U.S. through correspondence (Kraut), to examine Yiddish-language letters, poetry, and stories in tuberculosis sanatoriums (Gilman), and to examine daily life in tuberculosis sanatoriums and personal experiences with the disease (Rothman).

In these examples, students and researchers who used these materials only have the option to study a selection of individual records. Access to these records is provided in the form of visits to the archive or by browsing digital records available through our Digital Collections @ DU platform, https://specialcollections.du.edu/. Researchers often opt to visit the archive for an assignment or for research and record data using their laptop or mobile device from these records, effectively transcribing the same number of documents for their personal use. These users are limited in this respect to studying the few records that they were able to transcribe. Automatic transcription will not only ease these uses, but potentially lead to others, particularly in that transcription at scale allows for computational analysis of larger datasets.

Collections as Data Project

The opportunity to apply to the *Collections as Data: Part to Whole* initiative, supported by the Mellon Foundation, arose while transcription methods were being explored at DU Libraries. This multi-institutional initiative, centered at the University of Nevada at Las Vegas, provided funding to two project cohorts over the course of sixteen months that proposed to explore technical, organizational, and ethical issues related to turning archival collections into data. DU Libraries committed staff and technical resources to investigating and applying HTR to noteworthy collections in our archives as a pilot project, with the hopes of long-term adoption of these tools at scale. Because researchers and students often repeat the process of transcribing and collecting data from JCRS materials, a significant barrier to deep forms of analysis and potential creative forms of engagement with students and faculty, this archive was chosen for the pilot.

The handwritten documents in the JCRS records represent a valuable source not only to teaching and research in domains that use archival materials, but of data that should contribute to the ongoing development of HTR algorithms and technologies: training the models necessary to transcribe JCRS materials involved several hands, forms with both type and hand-written text, and quality of scans. The project contributes, then, to both scholarship that may use these data, but also digital library collections services as a discipline in itself. Another benefit of the project includes the creation and publication of these data as a way to preserve the information contained in historical texts at a greater scale than is possible through archival description alone. It also facilitates skills necessary in archival organizations to responsibly transform already digitized materials into machine-usable data. Essentially, *Collections as Data* allowed DU Libraries not only to produce and preserve a useful dataset, but to build internally the necessary skills and infrastructure for HTR work, even as it contributed to the larger infrastructure undergirding it, and to explore the ethical issues associated with both.

Six discrete phases guided the project: 1) Initiation/Planning; 2) Data Preparation; 3) Constructing an HTR pipeline; 4) Producing Collections as Data; 5) Building Delivery Platform; and 6) Service Design. The first phase involved training and planning on how to use the HTR platform *Transkribus,* technology developed with support from the European Union, which was used to transcribe the documents and produce the data (READ COOP). An Ethics Advisory Board (EAB) was also formed in the Initiation phase, which would become critically important as services were designed in the final stage.

The Data Preparation phase included generating ground truth and training data for the model that would ultimately auto-transcribe the collection. This work required manual transcription of documents, a necessary

step in that it provides correct data for machine-learning models so they may improve recognition and apply that "learning" to the entire collection (Figure 1). Quality assurance (QA) scripts were also created to assist in this process.

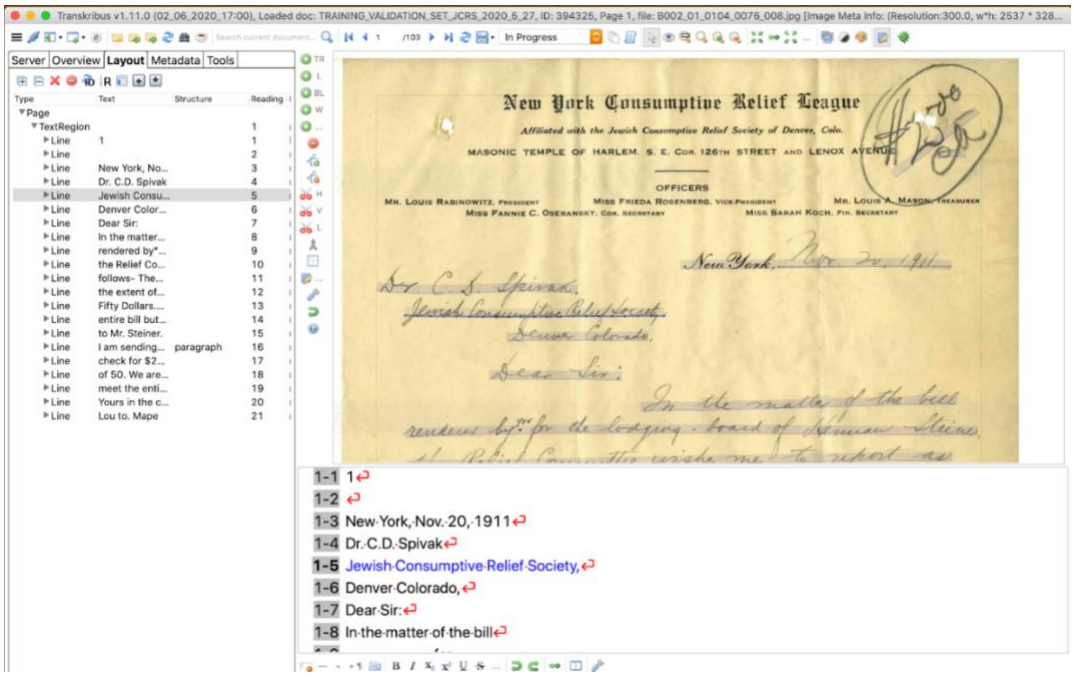

Figure 1. Editing and transcribing a document to generate ground truth in the *Transkribus* platform. Note the line-by-line transcription below the image.

The HTR pipeline phase entailed iterative training of the model in Transkribus, on both ground-truth and training data, to achieve a character error recognition rate (CER, the percentage of characters encoded incorrectly) low enough that the transcripts would be legible by humans. In this phase, every time the HTR pipeline is run in Transkribus, more data is produced, which is then manually corrected and added to the ground truth and training data; subsequent runs of the HTR pipeline then yield improved results. Two months into the project the CER was above 50%, and at the end of 18 months the goal of < 10% CER was accomplished (Figure 2). Research with Transkribus demonstrates that a CER of more than 10% renders automated transcriptions less useful research resources, simply because correcting errors would be more time consuming than manual transcription (Muehlberger et al. 962).

Time since project start	# Words of training data	CER on validation
2 months	1666 words	45%
3 months	2852	29.3%
6 months	6356	24.2%
9 months	11042	24.8%
12 months	19261	15.5%
18 months	40912	9.4%

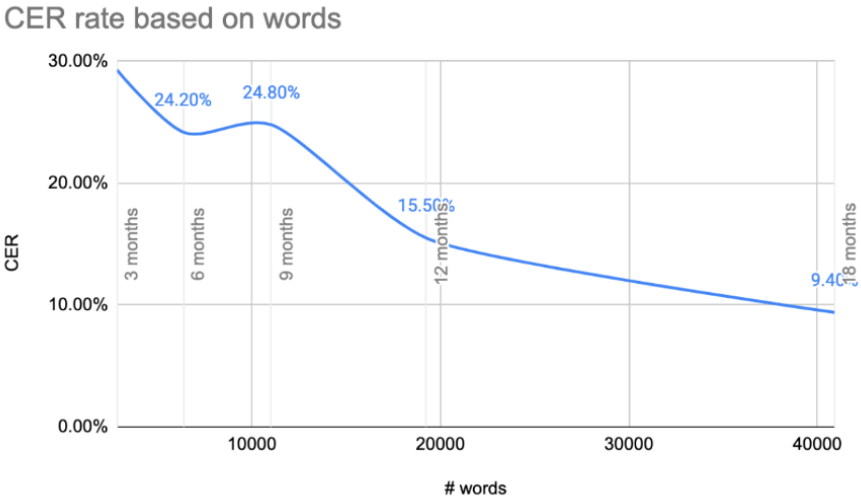

Figure 2. The process of training a machine learning model on the JCRS collection. The character error rate reached 9.4% after 18 months and over 40,000 words in the training data. Improvements began to plateau at this point and the model was sufficient to begin automated transcription.

Barbara Statement on Collections as Data" (Padilla). This document suggests ten principles upon which cultural heritage institutions should rest their access policies and services for computationally-analyzable data, noting that "ethical concerns are integral to collections as data," that "a commitment to openness" should be considered parallel to "care must be taken to comply with legal requirements, cultural norms, and the values of vulnerable groups" (Padilla et al. 2). Indeed, the second Santa Barbara Principle is: Ultimately, the project demonstrated this work is possible, sustainable, and valuable. But serious ethical considerations were also addressed in the final phase of the project.

Ethical Considerations

One of the evaluation criteria for the *Collections as Data: Part to Whole* call for proposals was whether the project "demonstrates commitment to developing and implementing processes for addressing complex ethical issues inherent to engagement with cultural heritage data, and the needs of marginalized and underrepresented communities" (Part to Whole). This criterion is to some degree an outgrowth of previous work of the *Always Already Computational – Collection as Data* project, funded by the Institute for Museum and Library Services (IMLS), from which arose the "Santa Barbara Statement on Collections as Data" (Padilla). This document suggests ten principles upon which cultural heritage institutions should rest their access policies and services for computationally-analyzable data, noting that "ethical concerns are integral to collections as data," that "a commitment to openness" should be considered parallel to "care must be taken to comply with legal requirements, cultural norms, and the values of vulnerable groups" (Padilla et al. 2). Indeed, the second Santa Barbara Principle is:

> "**Collections as data stewards are guided by ongoing ethical commitments.** These commitments work against historic and contemporary inequities represented in collection scope, description, access, and use. Commitments should be formally documented and made publicly

> available. Commitment details will vary across communities served by collections but will share common cause in seeking to address the needs of the vulnerable. Collection stewards aim to respect the rights and needs of the communities who create content that constitute collections, those who are represented in collections, as well as the communities that use them" (Padilla et al. 3).

This principle, applied to JCRS materials, particularly patient records published as data, illuminates a variety of serious ethical questions that need answering. Legal requirements, cultural norms, and the values of a potentially vulnerable group—in this case the Jewish community—were taken into consideration and resulted in a terms-of-use approach that differentiates uses and leverages the EAB, comprised of experts in several domains and members of the Rocky Mountain Jewish community.

Legal Requirements

Two legal requirements are at issue in determining access to JCRS materials. The first is the Health Insurance Portability and Accountability Act (HIPAA) of 1996 (United States Congress). This Act, and perhaps more to the point, the resulting "Privacy Rule" that was finalized by the Department of Health and Human Services in May of 2003, governs access and protection of certain health information (Department of Health and Human Services). This Rule defines both "Covered Entities," which are both individuals and organizations to whom the Act applies, and "Protected Health Information (PHI)," data about which the Act applies. Though the analysis can be somewhat complicated, with respect to JCRS materials held by the University of Denver Libraries, the Libraries are not a "Covered Entity" (Department of Health and Human Services) and JCRS records are not PHI, as the individuals to whom these records pertain are all deceased more than fifty years prior to the Libraries providing access to them (Department of Health and Human Services). HIPPA, then, does not restrict the libraries from providing access to information in the JCRS collection.

A second legal consideration is whether these data are subject to Institutional Review Board (IRB) protocols, which oversee the ethical conduct of research involving human subjects per federal regulation. Human subjects are defined by the code as "a living individual about whom an investigator (whether professional or student) conducting research obtains: (1) Data through intervention or interaction with the individual, or (2) Identifiable private information" (Code of Federal Regulations). Again, JCRS records are not subject to these regulations as they all involve deceased individuals.

There are then no known legal barriers to the libraries providing public access to JCRS records, and for this reason they have been available in part as scanned images for many years in various online repositories and have contributed to research and genealogy works.

Simply put, we can provide access to these materials. But that does not always mean we should.

Ethics Advisory Board

To address ethical issues related to "cultural norms" and the "values of vulnerable groups," per the *Santa Barbara Statement*, this project seated an EAB. Comprised primarily of members of the Jewish community in Denver, including many who work with the Rocky Mountain Jewish Historical Society, the EAB includes a physician, an attorney, an historian, a medical-ethicist, and a librarian with a Juris Doctorate. Once the project was funded, the EAB was presented the project design and objectives, primers on the ethical considerations of relevance, and asked to help construct a model by which access to JCRS data would be provided.

First and foremost, the EAB was advised that fundamentally new questions may arise from fact that HTR would enable the collection to be computationally analyzed. One concern grew from what was at the time

a very recent (and which of course remains an unspeakably horrific) attack on a synagogue in the United States. The newly public availability of this dataset, pertaining to a community vilified at points throughout history during the outbreak of disease (Schaub), potentially presents opportunities for bad-faith users that, while strictly possible from simply scanned images, is likely too painstaking to be of great concern. Perhaps a different access policy should be applied to transcripts as an entire dataset than the images and individual transcripts presented alongside them.

The EAB was also presented approaches adopted by projects with somewhat overlapping issues, in addition to summaries of the *Santa Barbara Statement* and the work that led to its formation. Its very formation was inspired, in fact, in conversations with a project team at the University of California San Francisco, whose work on "No More Silence – Opening the Data of the HIV/AIDS Epidemic," seeks to provide "200,000 pages of textual AIDS/HIV historical materials which have been digitized as part of various digitization projects . . . [and] will extract unstructured, textual data from these materials using Optical Character Recognition (OCR) and related software" (UCSF). This project of course shared similarities with respect to medical information in a potentially vulnerable community.

There are significant differences between these projects in that *No More Silence* entails some records that include PHI and are thereby subject to HIPPA. Their approach utilizes a confidentiality agreement as a prerequisite to view some records, wherein researchers agree not to disclose any PHI in their work in a legally-binding document that indemnifies the institution in any future litigation. The EAB did not feel such a strict access measure would be necessary for JCRS materials, even if not legally required, but was interested in an agreement of some sort that would commit researchers to ethical use.

Another project that informed conversations with the EAB was "Digital Histories of Eugenic Sterilization: Developing a Multi-Modal Prototype and Best Practices for Sensitive Health Data." This project sought, in part, to propose "guidelines for best practices around the digital uses of protected health information" (Stern and Wernimont 2) and found there is "disagreement across disciplinary fields about what constitutes appropriate use of sensitive materials" (4). Stern and Wernimont further framed the disagreement as "a generally unresolvable tension between 'the right to know' and the 'need to protect'" (4), concluded "the notion of universal best practices is problematic when dealing with such information" (11) and that "it is imperative for archive-makers to engage with impacted communities so that digital materials are meaningful for them" (14). EAB discussions regarding the JCRS materials echoed many of these themes, including disciplinary disagreements that led to a desire to somehow address the tension between "knowing" and "protecting."

Similar to *No More Silence,* the *Digital Histories* project included PHI and is further subject to "pertinent California Code that [states] all documents 75 years and older are completely open access and can be used in their entirety and documents less than 75 years old should be de-identified" (4). The project ultimately produced *Eugenic Rubicon*, "a digital resource that draws from and complements the demographic and social science research on eugenic sterilization in California" (Wernimont). This interactive digital resource includes records regarding victims of this state-sanctioned, forced sterilization, and manages the "knowing" versus "protecting" tension (and complies with relevant legislation) by redacting individual's names, openly exploring ethical issues in its text, and including and contributing to calls for redress to survivors of this "deplorable human rights abuse" (L.A. Times). Quite recently California has followed the lead of other states in providing reparations to survivors (Morris).

Finally, the EAB discussed published principles that guide work of the "Colored Conventions Project," a comprehensive and broad initiative that curates, shares, exhibits, and incorporates into teaching and research documents from "Colored Conventions," gatherings that were "held throughout the antebellum period and continuing for 30 years beyond the Civil War," and which "offered opportunities for free-born and formerly enslaved African Americans to organize and strategize for racial justice" (Colored Convention Project). The principles helped the EAB focus its efforts to manage the tension between "knowing" and "protecting" with respect to JCRS records as data, particularly the fifth:

> We affirm the role of Black people as data creators and elevate the ways in which Black conventions generated data and statistics to advance, affirm and advocate for Black economic and organizational success and access. We also recognize that data has long served in the processes and recording of the destruction and devaluation of Black lives and communities. We seek to avoid exploiting Black subjects as data and to account for the contexts out of which Black subjects as data arise. We seek to name Black people and communities as an affirmation of the Black humanity inherent in Black data/curation. We remind ourselves that all data and datasets are shaped by decisions about whose histories are recorded, remembered, and valued (Colored Convention Project).

There was a desire among the EAB and the project team to, in a similar fashion, elevate the ways in which the JCRS advanced and affirmed the right to health care of the poor, immigrants, and the marginalized, and the philanthropy and beneficence of the primarily Jewish founders and caretakers of the sanitarium. There was a desire to continue providing access to documents that lend themselves to genealogical research (a strong cultural value in many communities, including the Jewish community), and a desire to humanize the individual patients, nurses, doctors and staff of the JCRS, who sought simply to, as the mission of the sanitarium declared, "alleviate human suffering." And there was also a desire to protect these individuals and their progeny from destruction and devaluation, the realities of which could be made more possible by materials published as data, even as these data may also contribute to humanization and "knowing."

Simply put, the EAB decided not only that we can, but that we should. And that we should explicitly ask users to consider these ethical questions as they engage in the data provided them.

A hybrid approach was agreed upon, one wherein individual transcripts would be provided fully open access alongside scanned images in order to aid genealogy, search and retrieval of individuals, but access to the full dataset of transcripts, where computational analysis is made eminently more possible, would be accompanied by a terms of use agreement.

Terms of Use Approach

Inspired by and borrowing from *Colored Convention Project Principles*, *No More Silence*, and *Digital Histories*, JCRS data users are now required to agree to the following terms before given access to the collection as data:

> "I affirm the role of JCRS patients and staff as data creators and will avoid exploiting and/or dehumanizing them by treating them simply as data.
>
> My research will, when possible and appropriate, account for the contexts surrounding the JCRS subjects as data arise. My work will recognize that all data and datasets are shaped by decisions about how histories are recorded, remembered, and valued.
>
> If the nature of my work is such that I am sharing the life stories and/or narratives of individuals in these data, and I can do so with no potential harm to their reputation or that of their ancestors, I will

> honor them by naming them. If the nature of my work is such that I am exploring large-scale patterns in the dataset, and naming individuals serves no specific research purpose, I will anonymize and/or redact names within the data.
>
> If I am publishing the results of research conducted with these data, I will, if possible and appropriate, include a note of recognition and/or gratitude in my publication. We suggest a version of:
>
> 'This work was made possible in part by the patients, staff, nurses, physicians, and community of the Jewish Consumptives' Relief Society (JCRS). The people who lived, worked, and died at the JCRS sought to relieve human suffering. I am grateful to them.'"

This hybrid approach, while by no means completely or perfectly resolving the tension of "knowing" and "protecting," allows for the continued and enhanced engagement of these materials as single documents, while also introducing a new level of "protecting" that asks researchers to reflect on the data, their context, the humans and their context, and the researcher themselves and their context.

Discussion

While the tension between "knowing" and "protecting" cannot be fully resolved, the JCRS project and its engagement with the EAB underscores that archivists engaging with community members is imperative and that "it is incumbent upon those creating newly public and accessible resources on such sensitive histories to address issues of possible individual or community harm" (Stern and Wernimont 12). Working with the EAB allowed the project team to address issues of possible harm introduced by the availability of transcripts as a dataset, while also facilitating the use of data in ways that honors cultural norms and the values of the community.

Certainly, there are distinct differences among the projects discussed with the EAB. Whereas one involves forced, unnecessary and destructive invasions of the body, the JCRS offered desired medical attention to those desperate for it: one institution harmed and prevented life, the other lengthened and affirmed it. But there are also similarities, not only in terms of types of data, but underlying purposes of their original creation. Three of these projects, in fact, document in some manner or another the work of oppressed communities advocating for themselves, whether it be after centuries of enslavement and ongoing oppression, or during epidemics that disproportionally affected vulnerable communities.

That the work of the JCRS was benevolent does not prevent its data from being malevolently used, however. How and why the data was originally collected and used can be divorced from how and why it is stored and used by people long after the fact. Indeed, the very vocabulary used in digital archival work suggests the records of people may reduce those people to mere things: "The lexicon of digital collections extends the freighted, fretted, relation of categorization and data collection, to Black subject and Black subjectivity . . . the term 'item,' like 'object,' again recalls the ways in which Black people appear/ed in public records –as items on manifests" (Foreman 11). That the JCRS categorized and inventoried humans not as property, but in fact for the purpose of healing them, does not prevent users today from using the very same data in order to treat them as less than human. Ultimately, JCRS records are often mere ledgers, of people and transactions, of diagnoses and treatments, of events and relationships among individuals. As data, their use is not determined by who produced the data, but by those who analyze it.

And as responsible stewards of these data, it is incumbent upon project teams to ask the questions, engage community members, and resist the urge to apply generic best practices to each project.

Conclusion

This paper does not suggest an approach that can be simply replicated. On the contrary, the tension between knowing and protecting, between a commitment to openness and to ongoing ethical considerations that may run contrary to full openness, is one that can only be managed via careful planning and communication. As technology continues to evolve and is applied to archival sciences, it is important it be mediated by humans. Humans, not only as data producers, users, or analysts, but as individuals who are also members of communities. The person and their social context –whether they are the studied or the studier– should be held in the forefront throughout. What technology makes possible legality will not always constrict: only the intentional application of professional ethics can manage the tension in this work. The *Santa Barbara Statements* provide an enormously helpful centerpiece to these considerations.

This project does, however, offer an approach, informed by those that preceded it, that can be of use to similar projects. Seating an advisory board, with various expertise and of members of potentially vulnerable communities, and sharing with them approaches developed by others, can be an effective way of developing culturally-sensitive principles, values, and terms of use that guide digital collections work and their resulting analyses. The lure of powerful new technologies such as HTR are strong and scholars cannot divorce themselves from those they study, as harm is not always readily apparent.

As the humanities, archives and library studies, and other disciplines embark on quests to "rescue" archival data for modern use, as in general they should, they ought also to be careful to ask on a case-by-case and continual basis not only whether they should, but if so, how.

Acknowledgement

This work was made possible in part by the patients, staff, nurses, physicians, and community of the Jewish Consumptives' Relief Society (JCRS). The people who lived, worked, and died at the JCRS sought to relieve human suffering. We are grateful to them.

We are also thankful to our colleague Dr. Jeanne Abrams, Professor, Center for Judaic Studies and University Libraries; Director, Rocky Mountain Jewish Historical Society; and Curator, Beck Archives of Rocky Mountain Jewish History, University of Denver. For decades Dr. Abrams has been a careful steward and analyst of the JCRS and its records. Without her career-long dedication and expertise there would simply be no collection to transcribe, nor, perhaps, any awareness of its value.

Our work was also made possible by the *Collections as Data: Part to Whole* project and the support of the Andrew W. Mellon Foundation.

Works Cited

UCSF Archives & Special Collections. *UCSF Archives & Special Collections awarded $99,325 LSTA grant for textual data extraction from historical materials on AIDS/HIV*. 16 August 2018. https://blogs.library.ucsf.edu/broughttolight/2018/08/16/ucsf-archives-special-collections-awarded-99325-lsta-grant-for-textual-data-extraction-from-historical-materials-on-aids-hiv/. 16 August 2021.

Abrams, Jeanne. *Dr. Charles David Spivak: A Jewish Immigrant and the American Tuberculosis Movement*. Boulder, CO: University Press of Colorado, 2009.

Balci, Batuhan, Dan Saadati, and Dan Shiferaw. "Handwritten Text Recognition Using Deep Learning." CS231n: Convolutional Neural Networks for Visual Recognition, Stanford University, Course Project Report. 2017. http://cs231n.stanford.edu/reports/2017/pdfs/810.pdf .

Center for Research Libraries. *Award for Teaching: "Unmediated Archives: Creating an Immersive Experience for Undergraduate Students Across the Disciplines"*. 2018. https://www.crl.edu/focus/article/12596. 13 September 2021.

Code of Federal Regulations. "45 CFR 46.102(f)." *PROTECTION OF HUMAN SUBJECTS*. n.d. https://www.hhs.gov/ohrp/sites/default/files/ohrp/policy/ohrpregulations.pdf.

Colored Convention Project. *About the Colored Conventions*. n.d. https://coloredconventions.org/about-conventions/. 20 August 2021.

—. *Colored Convention Project Principles*. n.d. https://coloredconventions.org/about/principles/. 20 August 2021.

Cordell, David A. Smith and Ryan. "A Research Agenda for Historical and Multilingual Optical Character Recognition." 2019. *http://hdl.handle.net/2047/D20298542*. 19 July 2021.

Department of Health and Human Services. 2003. *Summary of the HIPPA Privacy Rule*. https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html. 16 August 2021.

—. *Covered Entities and Business Associates*. n.d. https://www.hhs.gov/hipaa/for-professionals/covered-entities/index.html. 16 August 2021.

—. *Health Information of Deceased Individuals*. n.d. https://www.hhs.gov/hipaa/for-professionals/privacy/guidance/health-information-of-deceased-individuals/index.html. 16 August 2021.

Foreman, P. Gabrielle and Labanya Mookerjee. "Computing in the Dark: Spreadsheets, Data Collection and DH's Racis Inheritance." *Always Already Computatoinal: Collections as Data National Forum Position Statements* (2017): 11-12. doi:10.5281/zenodo.3066161.

Ghosh, Rajib and Panda, Chinmaya and Kumar, Prabha. "Handwritten Text Recognition in Bank Cheques." *2018 Conference on Information and Communication Technology (CICT)*. 2018. 1-6.

Gilman, Ernest B. *Yiddish Poetry and the Tuberculosis Sanatorium: 1900-1970*. Syracuse University Press, 2014.

International Earth Data Alliance. *International Data Rescue Award in the Geosciences*. n.d. https://www.elsevier.com/awards/international-data-rescue-award-in-the-geosciences. 30 July 2021.

International Science Council Committee on Data, Data at Risk Task Group. *Data at Risk* . n.d. https://codata.org/initiatives/task-groups/previous-tgs/data-at-risk/. 30 July 2021.

James R. Wright, Jr and Jeanne Abrams. "Philip Hillkowitz The "Granddaddy of Medical Technologists" and Cofounder of the American Society for Clinical Pathologists and the Jewish Consumptives' Relief Society." *Arch Pathol Lab Med* (2018): 127-138.

Kirmizialtin, Suphan and David Wrisley. "Automated Transcription of Non-Latin Script Periodicals: A Case Study in the Ottoman Turkish Print Archive." *arxiv preprint* (2020). https://arxiv.org/abs/2011.01139.

Kraut, Alan. "“No Matter How Poor and Small the Building”: Health Care Institutions and the Jewish Immigrant Community." Yvonne Yazbeck Haddad, Jane I Smith, Jane I. Smith, John L. Esposito. *Religion and Immigration: Christian, Jewish, and Muslim Experiences in the United States*. n.d.

Los Angeles Times Editorial Board. "Editorial: California needs to do more than apologize to people it sterilized ." *Los Angeles Times* 21 January 2017. https://www.latimes.com/opinion/editorials/la-ed-eugenics-california-20170122-story.html.

Mayer, Allana. "Crowdsourcing, Open Data and Precarious Labour." 2016. *Model View Culture*. https://modelviewculture.com/pieces/crowdsourcing-open-data-and-precarious-labour. 16 August 2021.

Morris, Amanda. "'You Just Feel Like Nothing': California to Pay Sterilization Victims." *The New York Times* 11 July 2021. https://www.nytimes.com/2021/07/11/us/california-reparations-eugenics.html.

Muehlberger, G., Seaward, L., Terras, M., Ares Oliveira, S., Bosch, V., Bryan, M., Colutto, S., Déjean, H., Diem, M., Fiel, S., Gatos, B., Greinoecker, A., Grüning, T., Hackl, G., Haukkovaara, V., Heyer, G., Hirvonen, L., Hodel, T., Jokinen, M., Kahle, P. "Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study." *Journal of Documentation* (2019): 954-976.

Orlowitz, Jake. "The Crowdsourcing Fallacy." 2017. *Medium*. https://medium.com/a-wikipedia-librarian/the-crowdsourcing-fallacy-efe510c6f509. 16 August 2021.

Padilla, Thomas, Allen, Laurie, Frost, Hannah, Potvin, Sarah, Russey Roke, Elizabeth, & Varner, Stewart. "Santa Barbara Statement on Collections as Data --- Always Already Computational: Collections as Data." 2019. https://doi.org/10.5281/zenodo.3066209.

Part to Whole, Collections as Data. *Call for Proposals*. 2018. https://collectionsasdata.github.io/part2whole/cfp/. 16 August 2021.

Pham, Kim. *University of Denver Collections as Data*. 2020. https://wikis.du.edu/display/libpub/Collections+as+Data.

—. *University of Denver Collections as Data - Entire Dataset JCRS 2020_8_30*. 2020. https://doi.org/10.5281/zenodo.4150881.

READ COOP. *Transkribus: Where AI meets historical documents*. n.d. https://readcoop.eu/transkribus/. 24 September 2021.

Romero, V., Serrano, N., Toselli, A. H., Sánchez, J. A., & Vidal, E. "Handwritten text recognition for historical documents." *Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage*. 2011. 90-96.

Rothman, Sheila. *Living In The Shadow Of Death: Tuberculosis And The Social Experience Of Illness In America*. Basic Books, 1995.

Schantz, Herbert F. *The history of OCR, optical character recognition*. Manchester Center, VT: Recognition Technologies Users Association, 1982.

Schaub, Max. "Disease threat and the activation of antisemitism." 2020. https://osf.io/69avn/. 30 August 2021.

Stern, Alexandra Minna and Jacqueline Wernimont. "Digital Histories of Eugenic Sterilization: Developinga Multi-Modal Prototype and Best Practices for Sensitive Health Data." White Paper, Humanities

Collections and Reference Resources, Foundations Grant, Grant Number PW-234665-16. 2018. https://securegrants.neh.gov/publicquery/main.aspx?f=1&gn=PW-234665-16.

United States Congress. "Public Law 104-191." *HEALTH INSURANCE PORTABILITY AND ACCOUNTABILITY ACT OF 1996*. n.d. https://www.govinfo.gov/content/pkg/PLAW-104publ191/pdf/PLAW-104publ191.pdf.

Wernimont, Jacqueline and Alexandra Minna Stern. *Eugenic Rubicon.* n.d. https://scalar.usc.edu/works/eugenic-rubicon-/index. 20 August 2021.