Articles

From Digital Library to Open Datasets: Embracing a "Collections as Data" Framework

Rachel Wittmann, Anna Neatrour, Rebekah Cummings, and Jeremy Myntti

ABSTRACT

This article discusses the burgeoning "collections as data" movement within the fields of digital libraries and digital humanities. Faculty at the University of Utah's Marriott Library are developing a collections as data strategy by leveraging existing Digital Library and Digital Matters programs. By selecting various digital collections, small- and large-scale approaches to developing open datasets are explored. Five case studies chronicling this strategy are reviewed, along with testing the datasets using various digital humanities methods, such as text mining, topic modeling, and GIS (geographic information system).

INTRODUCTION

For decades, academic research libraries have systematically digitized and managed online collections for the purpose of making cultural heritage objects available to a broader audience. Making archival content discoverable and accessible online has been revolutionary for the democratization of scholarship, but the use of digitized collections has largely mimicked traditional use: researchers clicking through text, images, maps, or historical documents one at a time in search of deeper understanding. "Collections as data" is a growing movement to extend the research value of digital collections beyond traditional use and to give researchers more flexible access to our collections by facilitating access to the underlying data, thereby enabling digital humanities research.[1]

Collections as data is predicated upon the convergence of two scholarly trends happening in parallel over the past several decades.[2] First, as mentioned above, librarians and archivists have digitized a significant portion of their special collections, giving access to unique material that researchers previously had to travel across the country or globe to study. At the same time, an increasing number of humanist scholars have approached their research in new ways, employing computational methods such as text mining, topic modeling, GIS (geographic information system), sentiment analysis, network graphs, data visualization, and virtual/augmented reality in their quest for meaning and understanding. Gaining access to high-quality data is a key challenge of digital humanities work, since the objects of study in the humanities are frequently not as amenable to computational methods as data in the sciences and social sciences.[3] Typically, data in the sciences and social sciences is numerical in nature and collected in spreadsheets and databases with the intention that it will be computationally parsed, ideally as part of a reproducible and objective study. Conversely, data (or, more commonly, "evidence" or "research assets") in the humanities is text- or image-based and is created and collected with the intention of close reading or analysis by a researcher who brings their subjective expertise to bear on the object.[4] Even a relatively simple digital humanities method like identifying word frequency in a corpus of literature is predicated on access to plain

Rachel Wittmann (rachel.wittmann@utah.edu) is Digital Curation Librarian, University of Utah. **Anna Neatrour** (anna.neatrour@utah.edu) is Digital Initiatives Librarian, University of Utah. **Rebekah Cummings** (rebekah.cummings@utah.edu) is Digital Matters Librarian, University of Utah. **Jeremy Myntti** (jeremy.myntti@utah.edu) is Head of Digital Library Services, University of Utah.

text (.txt) files, high-quality optical character recognition (OCR), and the ability to bulk download the files without running afoul of copyright or technical barriers.

As "The Santa Barbara Statement on Collections as Data" articulates, "with notable exceptions like the HathiTrust Research Center, the National Library of the Netherlands Data Services & API's, the Library of Congress' Chronicling America, and the British Library, cultural heritage institutions have rarely built digital collections or designed access with the aim to support computational use."[5] By and large, digital humanists have not been well-served by library platforms or protocols. Current methods for accessing collections data include contacting the library for direct access to the data or "scraping" data off library websites. Recently funded efforts such as the Institute of Museum and Library Services' (IMLS's) *Always Already Computational* and the Andrew W. Mellon Foundation's *Collections as Data: Part to Whole* seek to address this problem by setting standards and best practices for turning digital collections into datasets amenable to computational use and novel research methods.[6]

The University of Utah J. Willard Marriott Library has a long-running digital library program and a burgeoning digital scholarship center creating a moment of synergy for librarians in digital collections and digital scholarship to explore collaboration in teaching, outreach, and digital collection development. A shared goal between the digital library and digital scholarship teams is to develop collections as data of regional interest that could be used by researchers for visualization and computational exploration. This article will share our local approach to developing and piloting a collections as data strategy at our institution. Relying upon best practices and principles from Thomas Padilla's "On a Collections as Data Imperative," we transformed five library collections into datasets, made select data available through a public GitHub repository, and tested the usability of the data with our own research questions relying upon expertise and infrastructure from Digital Matters and the Digital Library at the Marriott Library.[7]

DIGITAL MATTERS

In 2015, administration at the Marriott Library was approached by multiple colleges at the University of Utah to explore the possibility of creating a collaborative space to enable digital scholarship. While digital scholarship was happening across campus in disparate and unfocused ways, there was no concerted effort to share resources, build community, or develop a multi-college digital scholarship center with a mission and identity. After an eighteen-month planning process, the Digital Matters pop-up space was launched as a four-college partnership between the College of Humanities, College of Fine Arts, College of Architecture + Planning, and the Marriott Library. An anonymous $1 million donation in 2017 allowed the partner colleges to fund staffing and activity in the space for five years, including the hire of a Digital Matters director tasked with planning for long-term sustainability.

The development of Digital Matters brings new focus, infrastructure, and partners for digital humanities research to the University of Utah and the Marriott Library. Monthly workshops, speakers, and reading groups led by digital scholars from all four partner colleges have created a vibrant community with cross-disciplinary partnerships and unexpected synergies. Close partnerships and ongoing dialogue have increased awareness for Marriott faculty, particularly those working in and collaborating with Digital Matters, of the challenges facing digital humanists and the ways in which the library community is uniquely suited to meet those needs. For example, a University of Utah researcher in the College of Humanities developed "Century of Black Mormons," a community-based public history database of biographical information and primary source documents on black Mormons baptized between 1830 and 1930.[8] Working closely with the Digital Initiatives librarian and various staff and faculty at the Marriott Library, they created an Omeka S site that allows users to interact with the historical data using GIS, timeline features, and basic webpage functionality.

HTTPS://DOI.ORG/10.6017/ITAL.V38I4.11101

INSTITUTION DIGITAL LIBRARY

The University of Utah has had a robust digital library program since 2000, including one of the first digital newspaper repositories, Utah Digital Newspapers (UDN, https://digitalnewspapers.org/). In 2016, the library developed its own digital asset management system using open-source systems such as Solr, Phalcon, and nGinx after using CONTENTdm for over fifteen years.[9] This new system, Solphal, has made it possible for us to implement a custom solution to manage and display a vast amount of digital content, not only for our library, but also for many partner institutions throughout the state of Utah. Our main digital library server (https://collections.lib.utah.edu/) contains over 765,000 objects in nearly 700 collections, consisting of over 2.5 million files. Solphal is also used to manage the UDN, containing nearly 4 million newspaper pages and over 20 million articles.

Digital library projects are continually evolving as we redefine our digital collection development policies, ensuring that we are providing researchers and other users the digital content that they are seeking. With such a large amount of data available in the digital library, we can no longer view our digital library as a set of unique, yet siloed, collections, but more as a wealth of information documenting the history of the university, the state of Utah, and the American West. We are also engaged in remediating legacy metadata across the repository in order to achieve greater standardization, which could support computational usage of digital library metadata in the future. With this in mind, we are working to strategically make new digital content available on a large scale that can help researchers discover this historical content within a collections as data mindset.

Leveraging the existing Digital Library and Digital Matters programs, faculty at the Marriott Library are in the process of piloting a collections as data strategy. We selected digital collections with varying characteristics and used them to explore small- and large-scale approaches to developing datasets for humanities researchers. We then tested the datasets by employing various digital humanities methods such as text mining, topic modeling, and GIS. The five case studies below chronicle our efforts to embrace a collections as data framework and extend the research value of our digital collections.

TEXT MINING MINING TEXTS

When developing the initial collections as data project, several factors were considered to identify the optimal material for this experiment. Selecting already digitized and described material in the University of Utah Digital Library was ideal to avoid waiting periods required for new digitization projects. The Marriott Library Special Collections' relationship with the American West Center, an organization based at the University of Utah with the mission of documenting the history of the American West, has produced an extensive collection of oral histories held in the Audio Visual Archive which have typewritten transcripts yielding high-quality OCR. Given the availability and readiness of this material, we built a selected corpus of mining-related oral histories, drawn from collections such as the Uranium Oral Histories and Carbon County Oral Histories. Engaging in the entire process with a digital humanities framework, we scraped our own digital library repository as though we had no special access to the back end of the system, developing a greater understanding of the process and workflows needed to build a text corpus to support a research inquiry. In this way, we extended our skills so that we would be able to scrape any digital library system if this expertise was needed in the future.

The extensive amount of text produced by the corpus of 230 combined oral histories provided ideal material for topic modeling. Simply put, "topic modeling is an automatic way to examine the contents of a corpus of documents."[10] The output of these models is word clouds with varying sizes of words based on the number of co-occurrences within the corpus; larger words indicate more occurrences and smaller ones indicate fewer. Each topic model then points to the most relevant documents within the corpus based on the co-occurrences of the words contained in that model. In order to create these topic models from the

corpus of oral histories, a workflow was developed with the expertise of the Digital Matters cohort, implementing MALLET for R script, using the LDA topic model style, developed by Blei et al.[11]

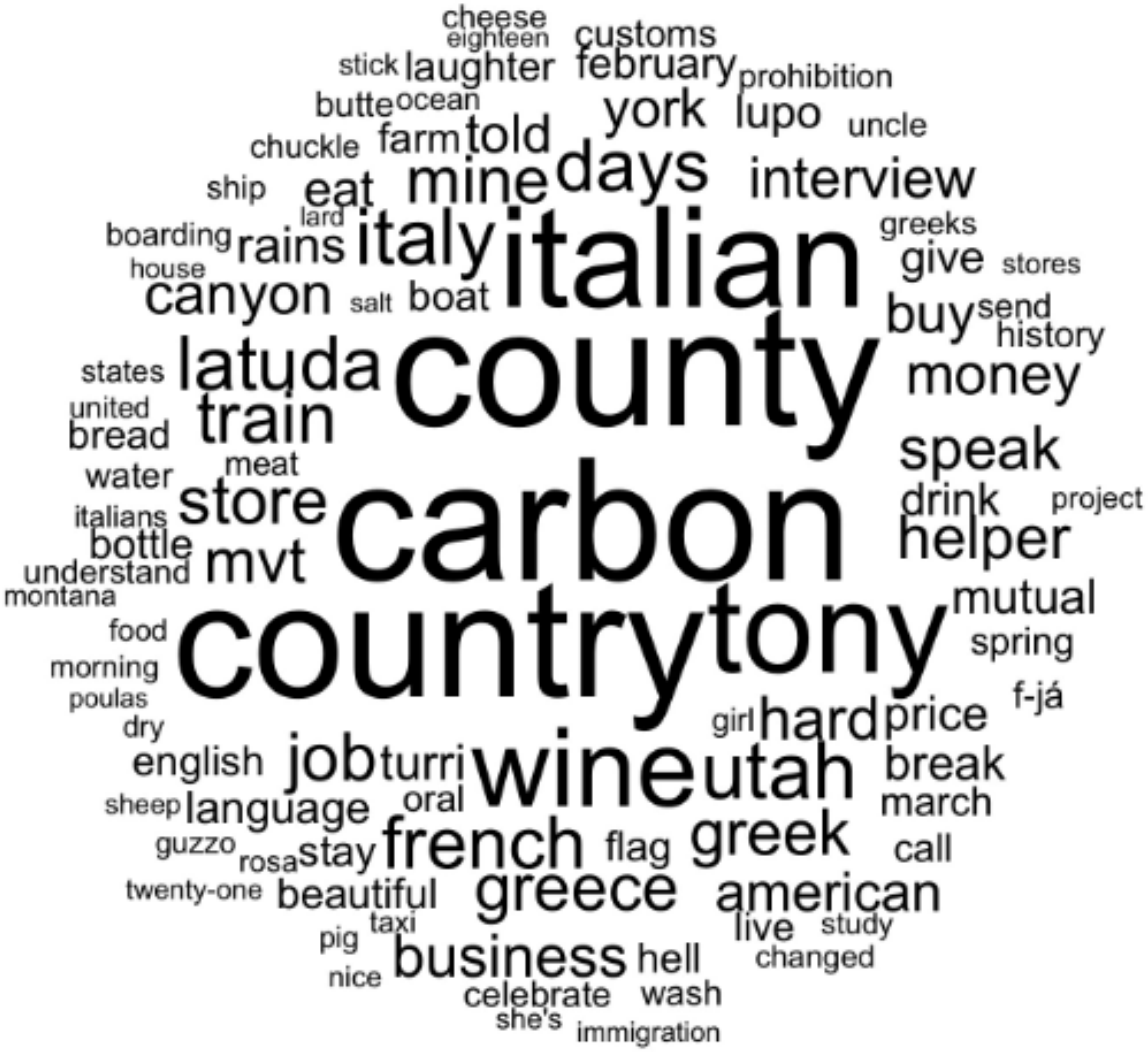


Figure 1. Topic model from text mining the mining-related oral histories found in the University of Utah's Digital Library.

From the mining-related oral history corpus, twenty-six topic models were created. Once generated, each topic model points to five interviews that are most related to the words in a particular model. In figure 1, the words carbon, county, country, and Italian are the largest, because the interviews are about Carbon County, Utah. Considering this geographical area of Utah was the most ethnically diverse in the late 1800s due to the coal mining industry recruiting labor from abroad, including Italy, these words are not surprising. As indicated by their prominence in the topic model, the set of words co-occur most often in the interview set. We approached the process of topic modeling the oral histories as an exploration, but after

HTTPS://DOI.ORG/10.6017/ITAL.V38I4.11101

reviewing the results, we discovered that many of the words which surfaced through this process pointed to deficiencies in the original descriptive metadata, highlighting new possibilities for access points and metadata remediation. Honing in on the midsize words tended to uncover unique material that is not covered in descriptive metadata, as these words are often mentioned more than a handful of times and across multiple interviews. The largest words in the model are typically thematic to the interview and included in the descriptive metadata. For example, when investigating the inclusion of "wine" in the topic model found in figure 1, conversations about the winemaking process amongst the Italian mining community in Carbon County, Utah were revealed. From an interview with Mary Nicolovo Juliana conducted in 1973 from the Carbon County Oral History Project, Nicolovo discusses how her father, a miner, made wine at home.[12]

As the topic models are based on co-occurrences in the corpus, there was another interview with Emile Louise Cances, from the Carbon County Oral History Project conducted in 1973. Cances, from a French immigrant mining family, discusses the vineyards her family had in France.[13] With both of these oral histories, there was no reference to wine in the descriptive metadata. A researcher may miss this content because it isn't included as an access point in metadata. Thus, topic modeling allowed for the discoverability of potentially valuable topics that may be buried in hundreds of pages of content.

From this collections as data project, text mining the mining oral history texts to produce topic models, we are considering employing topic modeling when creating new descriptive metadata for similar collections. Setting a precedent, the text files for this project are hosted on the growing Marriott Library Collections as Data Github repository. After we developed this corpus, we discovered that a graduate student in the History department had developed a similar project, demonstrating the research value of oral histories combined with computational analysis.[14]

HAROLD STANLEY SANDERS MATCHBOOKS COLLECTION

When assessing potential descriptive metadata for the Harold Stanley Sanders Matchbooks Collection, an assortment of matchbooks that reflect many bygone establishments predominately from Salt Lake City that include restaurants, bars, hotels, and other businesses, non-Dublin Core metadata was essential for computational purposes. With the digital project workflow now extending beyond publishing the collection in the Digital Library, to publishing the collection data to the Marriott Library Collections as Data GitHub repository, assessing metadata needs has evolved. As matchbooks function as small advertisements, they often incorporate a mix of graphic design, advertising slogans, and addresses of the establishment. The descriptive metadata was created first with the most relevant fields for computational analysis, including business name, type of business, transcription of text, notable graphics, colors of matchbooks, and street addresses. For collection mapping capabilities, street addresses were then geocoded using a Google Sheets add-on called Geocode Cells, which uses Google's Geocoding API (see figure 2).

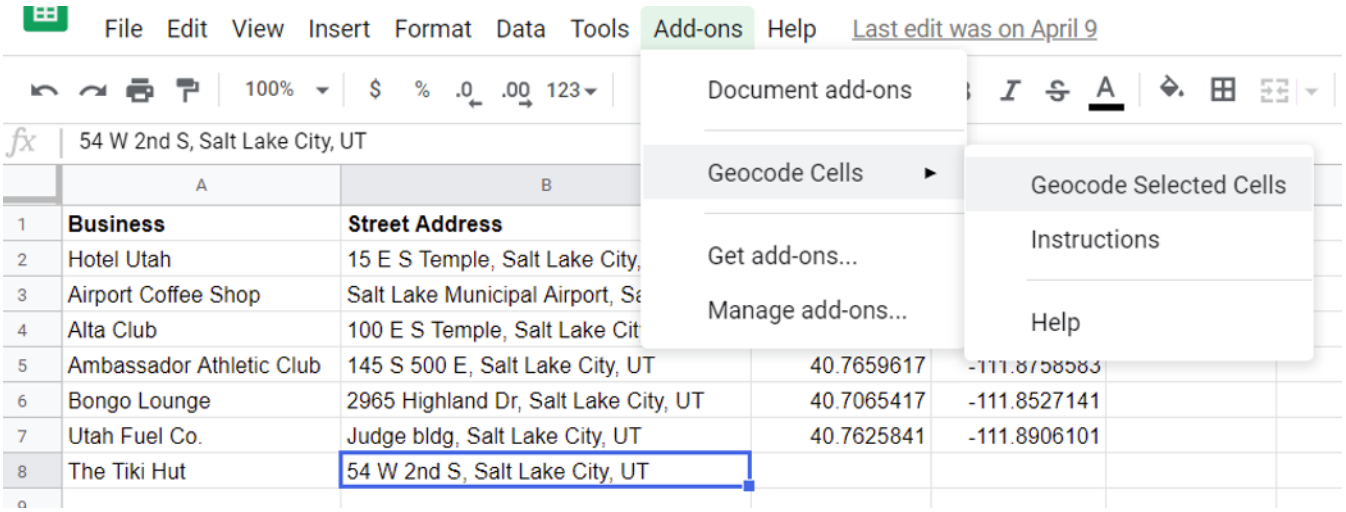


Figure 2. A screenshot of Google Sheets add-on, Geocode Cells. (https://chrome.google.com/webstore/detail/geocode-cells/pkocmaboheckpkcbnnlghnfccjjikmfc).

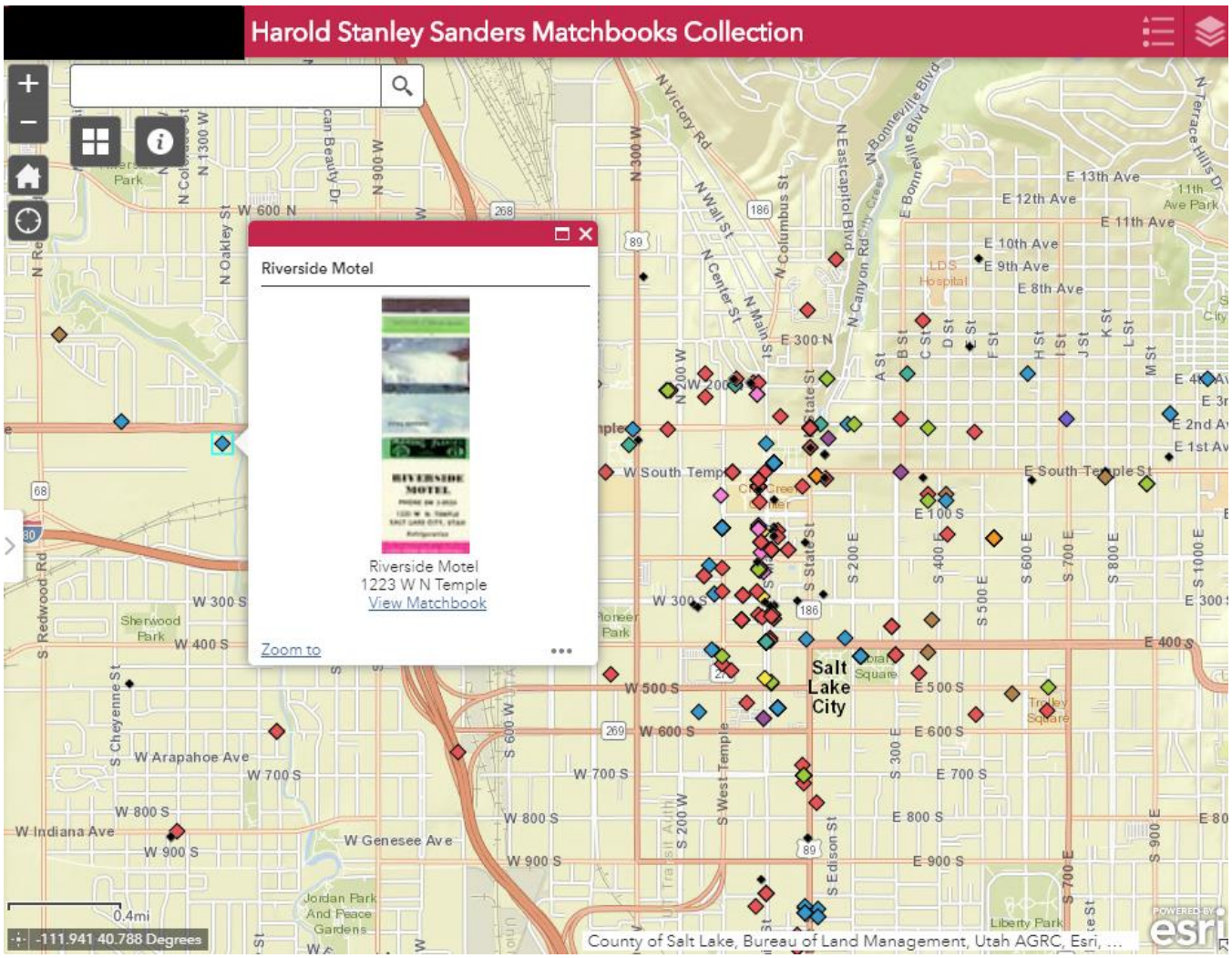


Figure 3. A screenshot of Harold Stanley Sanders Matchbook Collection Map, made with ArcGIS Online.

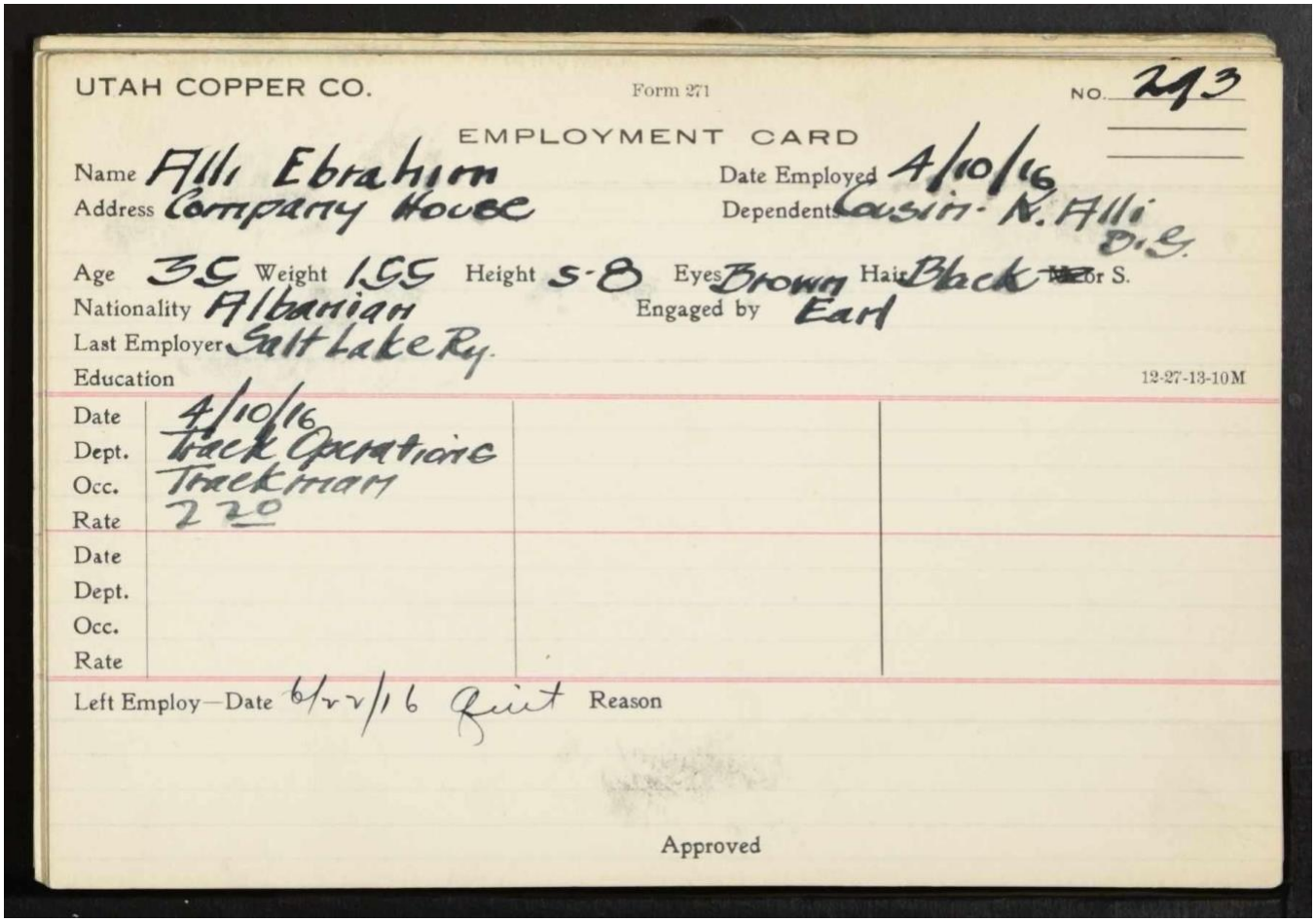
This proved efficient for this collection, as other geocoding services required zip codes for street addresses which were not present on the matchbooks. With the latitude and longitude addition to the metadata, the collection was then mapped using ArcGIS Online (see figure 3).[15]

The extensive metadata, including geographic-coordinate data, is available on the library's GitHub repository for public use. After the more computationally ready metadata was created, it was then massaged to fit library best practices and Dublin Core (DC) standards. This included deriving Library of Congress Subject Headings for DC subjects from business type and concatenating notable matchbook graphics and slogans for the DC description. While providing the extensive metadata is beneficial for computational experimentation, it adds time and labor to the lifespan of the project.

KENNECOTT COPPER MINER RECORDS

One aspect of our collections as data work at the University of Utah moving forward is the need for long-term planning for resources that contain interesting information that could eventually be used for computational exploration, even if we currently don't have the capacity to make the envisioned dataset available at the current time. The Marriott Library holds a variety of personnel records from the Kennecott Copper Corporation, Utah Copper Division. These handwritten index cards contain a variety of interesting demographic data about the workers who were employed by the company from 1900-19 such as name, employee ID, date employed, address, dependents, age, weight, height, eyes, hair, gender, nationality, engaged by, last employer, education, occupation, department, pay rate, date leaving employment, and reason for leaving. Not all the cards are filled out with the complete level of detail as listed in the fields above, however, usually name, date employed, ethnicity, and notes about pay rates for each employee are included.

Developing a scanning and digitization procedure for creating digital surrogates of almost 40,000 employment records was fairly easy due to an existing partnership and reciprocal agreement with FamilySearch, however developing a structure for making the digitized records available and providing full transcription is a long-term project. Librarians used this project as an opportunity to think strategically about the limits of Dublin Core when developing a collections as data project from the start. The digital library repository at the University of Utah provides the ability to export collection level metadata as .tsv files. With this in mind, the collection metadata template was created with the aim of eventually being able to provide researchers with the granular information on the records. This required introducing a number of new, non-standard field labels to our repository. Since we are not able to anticipate exactly how a researcher might interact with this collection in the future, our main priority was developing a metadata template that would accommodate full transcription for every data point on the card. Twenty new fields in the template reflect the demographic data on the card, and ten are existing fields that map to our standard practices with Dublin Core fields. Because we do not currently have the staffing in place to transcribe 40,000 records, we are implementing a phased approach of transcribing four basic fields, with fuller transcription to follow if we are able to secure additional funding.



UTAH COPPER CO. Form 271 NO. 293

EMPLOYMENT CARD

Name Alli Ebrahim Date Employed 4/10/16

Address Company House Dependents Cousin - N. Alli D.G.

Age 35 Weight 155 Height 5-8 Eyes Brown Hair Black ~~Wh~~ or S.

Nationality Albanian Engaged by Earl

Last Employer Salt Lake Ry.

Education 12-27-13-10M

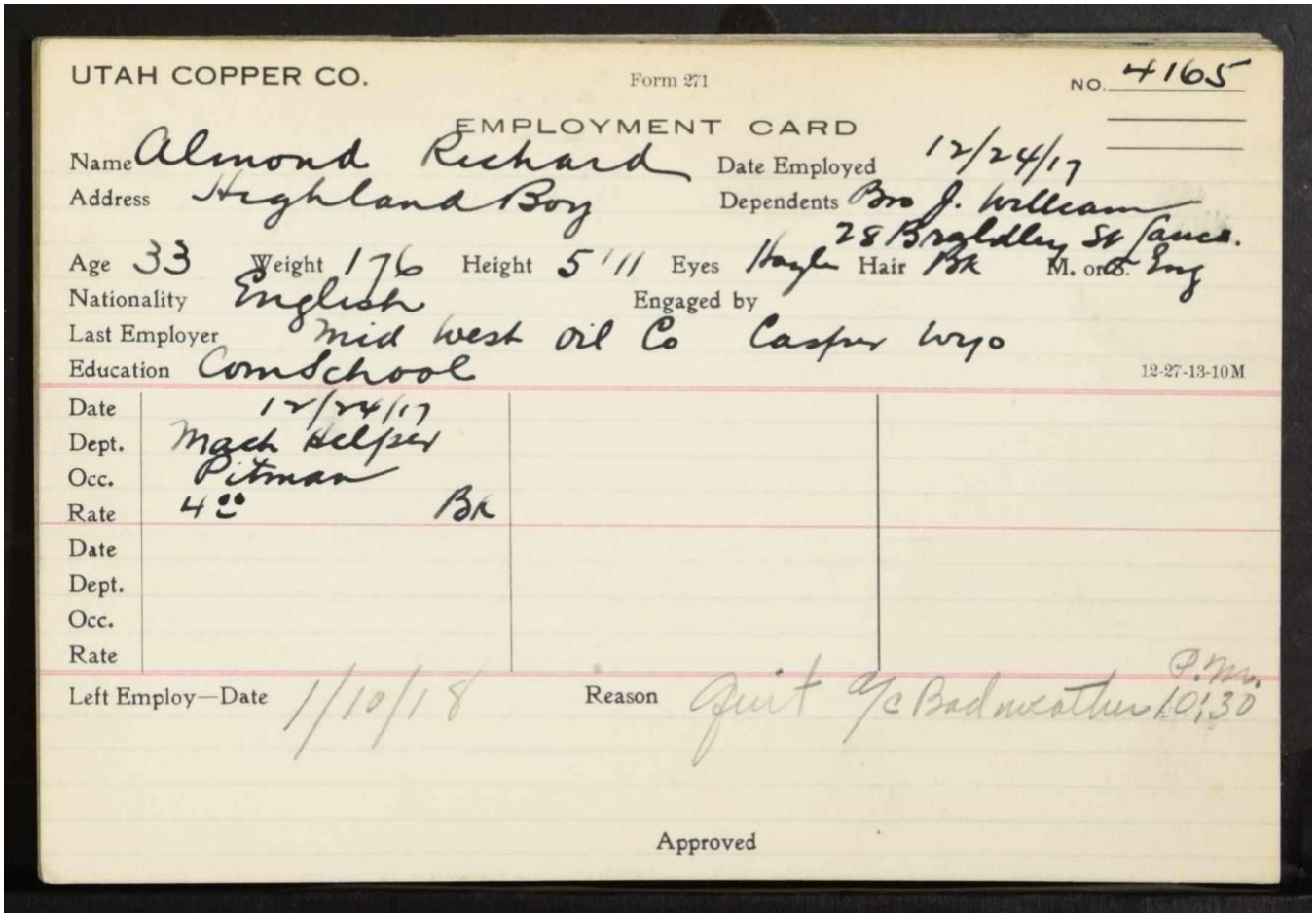
Date	4/10/16		
Dept.	Track Operations		
Occ.	Trackman		
Rate	2 20		
Date			
Dept.			
Occ.			
Rate			

Left Employ—Date 6/22/16 Quit Reason

Approved

Figure 4. Employment card for Alli Ebrahim, 1916.

HTTPS://DOI.ORG/10.6017/ITAL.V38I4.11101


UTAH COPPER CO. Form 271 NO. 4165

EMPLOYMENT CARD

Name Almond Richard Date Employed 12/24/17

Address Highland Boy Dependents Bro J. William 28 Brighley St Lawes.

Age 33 Weight 176 Height 5'11 Eyes Hazel Hair Bk M. or S. Eng

Nationality English Engaged by

Last Employer Mid West Oil Co Casper Wyo

Education Comschool

12-27-13-10M

Date	12/24/17		
Dept.	Mach Helper		
Occ.	Pitman		
Rate	4.00 Bk		
Date			
Dept.			
Occ.			
Rate			

Left Employ—Date 1/10/18 Reason Quit Yc Badweather P.M. 10:30

Approved

Figure 5. Employment card for Richard Almond, 1917.

WOMAN'S EXPONENT

A stated goal for Digital Matters is to be a digital humanities space that is unique to Utah and addresses issues of local significance such as public lands, water rights, air quality, indigenous peoples, and Mormon history.[16] When considering what digital scholarship projects to pursue in 2019, Digital Matters faculty became aware of the upcoming 150th anniversary of women in Utah being the first to vote in the nation. Working with a local nonprofit, Better Days 2020, and colleagues at Brigham Young University (BYU), Digital Matters faculty and staff decided to embark on a multimodal analysis of the 6,800-page run of the *Woman's Exponent*, a Utah women's newspaper published between 1872-1914 primarily under the leadership of Latter-day Saint Relief Society President Emmeline B. Wells. In its time, the *Woman's Exponent* was a passionate voice for women's suffrage, education, and plural marriage, and chronicled the interest and daily lives of Latter-day Saint women.

Initially, we hoped to access the data through the Brigham Young University Harold B. Lee Library, which digitized the *Exponent* back in 2000. We quickly learned that OCR from nearly twenty years ago would not suffice for digital humanities research and considered different paths for rescanning the *Exponent*. After accessing the original microfilm from BYU, we leveraged existing structures for digitization. Through an agreement that the Marriott Library has in place with a vendor for completing large-scale digitization of newspapers on microfilm for inclusion in the Utah Digital Newspapers program, we were able to add the *Woman's Exponent* to the existing project without securing a new contract for digitization. The vendor digitized the microfilm, created an index of each title, issue, date, and page, and extracted the full text

through an OCR process. They then delivered 330 GB of data to us, including high-quality TIFF and JP2000 images, a PDF file for each page, and METS-ALTO XML files containing the metadata and OCR text.

Acquiring data for the *Woman's Exponent* project illuminated the challenges that digital humanists face when looking for clean data. Our original assumption was that if something had already been scanned and put online, the data must exist somewhere. We soon learned, when working with legacy digital scans, that the OCR might be insufficient or the original high-quality scans might be lost over the course of multiple system migrations. As librarians with existing structures in place for digitization, we had the content rescanned and delivered within a month. Our digital humanities partners from outside of the library did not know this option was available and assumed our research team would have to scan 6,800 pages of newspaper content before we were able to start analyzing the data. This incongruity highlighted cultural differences between digital humanists with their learned self-reliance and librarians who are more comfortable and conversant looking to outside resources. Indeed, our digital humanities colleagues seemed to believe that "doing it yourself" was part and parcel of digital humanities work.

The *Woman's Exponent* project is still in early phases, but now that we have secured the data, we are considering what digital humanities methods we can bring to bear on the corpus. With the 2020 150th anniversary of women's suffrage in Utah, we have considered a topic modeling project looking at themes around universal voting, slavery, and polygamy and tracking how the discussion around those topics evolved over the 42-year run of the paper. Another potential project is building a social network graph of the women and men chronicled throughout the run of the paper. Developing curriculum around women in Utah history is of particular interest to the group as women are underrepresented in the current K-12 Utah history curriculum. Keeping in line with our commitment to collections as data, we have released the *Woman's Exponent* as a .tsv file with OCR full-text data, which can be analyzed by researchers studying Utah, Mormon studies, the American West, or various other topics. Collaborators have also developed a digital exhibit on the Woman's Exponent which includes essays about a variety of topics as well as sections showcasing its potential for digital scholarship.[17]

OBITUARY DATA

The Utah Digital Newspapers (UDN) program began in 2002 with the goal of making historical newspaper content from the State of Utah freely available to the public for research purposes. Between 2002 and 2019, there have been over 4 million newspaper pages digitized for UDN. Due to search limitations of the software system used for UDN at the time, the data model for newspapers was made more granular, and included segmentation for articles, obituaries, advertisements, birth notices, etc. This article segmentation project ended in 2016 when it was determined that the high cost of segmentation was not sustainable with mass digitization of newspapers and users were still able to find the content they are looking for on a full newspaper page.

Before the article segmentation project concluded, UDN had accrued over 20 million articles, including 318,044 articles that were tagged as obituaries or death notices. In 2013, the Marriott Library partnered with FamilySearch to index the genealogical information that can be gleaned from these obituaries. The FamilySearch Indexing (FSI) program crowdsourced the indexing of this data to thousands of volunteers worldwide. Certain pieces of data, such as place names, were mapped to an existing controlled vocabulary and dates were entered in a standardized format to ensure that certain pieces of the data are machine actionable.[18]

After the obituaries were indexed by FSI in 2014, a copy of the data was given to the Marriott Library to use in UDN. The indexed data included fields such as name of deceased, date of death, place of death, date of birth, birthplace, and relative names with relationships. Since this massive amount of data didn't easily fit within the UDN metadata schema, it was stored for several years without the Marriott Library doing anything with the data.

Now that we are thinking about our digital collections as data, we are exploring ways that researchers could use this vast amount of data. The data was delivered to the library in large spreadsheets that are not easily usable in any spreadsheet software. We are exploring ingesting the data into a revised newspaper metadata schema within our digital asset management system or converting the data into a MySQL database so it is possible to search and find relationships between pieces of data.

Working with a large dataset such as this can be challenging. The data from only two newspapers, including 1,038 obituaries, is a 25 MB file. The full database is over 10 GB of data. Since this is a large amount of data, we are working through issues related to how we can distribute this data in a usable way in order for researchers to make use of the data. We are also looking at the possibility of having FSI index additional obituary data from UDN, which will make the database continually expand.

CONCLUSION

As the digital library community recognizes the need for computational-ready collections, the University of Utah Digital Library has embraced this evolution with a strategic investment. Implementing the collections as data GitHub repository for computational users is a first step towards providing access to collections beyond the traditional digital library environment. While there may be improved ways to access this digital library data in the future, the GitHub repository filled an immediate need.

Developing standardized metadata for computational use can often require more time from metadata librarians who are already busy with the regular work of describing new assets for the digital library. Developing additional workflows for metadata enhancement and bulk download can delay the process in making new collections available. In most cases, collections need to be evaluated individually to determine what type of resources can be invested in making them available for computational use. For a project needing additional transcription, like the Kennecott Mining Records, crowdsourcing might seem like potential avenue to pursue. However, the digital library collection managers have misgivings about the training and quality assurance involved in developing a new large-scale transcription project. Combined with the desire to ensure that the people who are working on the project have adequate training and compensation for their labor, we are making the strategic decision to transcribe for some of the initial access points to the collection now, and attempt full transcription at a later date pending additional funding. For the UDN obituary data, leveraging an existing transcription program at no cost with minimal supervision needed by librarians worked well in being able to surface additional genealogical data that can be released for researchers.

The collections as data challenge mirrors a perennial digital library conundrum—how much time and effort should librarians invest for unknown future users with unknown future needs? Much like digitization and metadata creation, creating collections as data requires a level of educated guesswork as to what collections digital humanists will want to access, what metadata fields they will be interested in manipulating, and in what formats they will need their data. Considering the limited resources of librarians, should we convert our digital collections into data in anticipation of use or convert our collections on demand? This “just in case” vs. “just in time” question is worthy of debate and will naturally be dependent on the resources and priorities of individual institutions.

With an increasing number of researchers experimenting with digital humanities methods, collections as data will be a standard consideration when working with new digitization projects at the University of Utah. Visualization possibilities outside of the digital-library environment will be regularly assessed. Descriptive metadata practices beyond Dublin Core will be developed when beneficial to the computational and experimental use of the data by the public. Integrating techniques like topic modeling into descriptive metadata workflows provides additional insight about the digital objects being described. While adding collections as data to existing digitization workflows will require an additional investment of time, developing these projects has also created new opportunities for collaboration both within the library and

in developing expanded partnerships at the University of Utah and other institutions in the Mountain West. By leveraging our existing partnerships, we were able to create collections as data pilots organically by taking advantage of our current workflows and digitization procedures. While we have been successful in releasing smaller-scale collections as data projects, we still need to consider integration issues with our larger digital library program and experiment more with enabling access to large datasets. With librarians engaged in producing curated datasets that evolve from unique special collection materials, they can extend the research value of the digital library and the collections that are unique to each institution. As we look towards the future, we see this work continuing and expanding as librarians engage more with digital humanities teaching and support.

ACKNOWLEDGEMENTS

The authors would like to acknowledge Dr. Elizabeth Callaway, former Digital Matters postdoctoral fellow and current Assistant Professor in the Department of English at the University of Utah, for developing the topic modeling workflow used in the collections as data project, Text Mining Mining Texts. Callaway's expertise was invaluable in creating the scripts to enable distance reading of the text corpus, documenting this process, and training library staff.

REFERENCES

[1] Thomas G. Padilla, "Collections as Data: Implications for Enclosure," *College & Research Libraries News*; Chicago 79, no. 6 (June 2018): 296, https://crln.acrl.org/index.php/crlnews/article/view/17003/18751.

[2] Thomas Padilla et al., "The Santa Barbara Statement on Collections as Data (V1)," n.d., https://collectionsasdata.github.io/statementv1/.

[3] Christine L. Borgman, "Data Scholarship in the Humanities," in *Big Data, Little Data, No Data: Scholarship in the Networked World* (Cambridge, MA: The MIT Press, 2015), 161–201.

[4] Miriam Posner, "Humanities Data: A Necessary Contradiction," *Miriam Posner's Blog* (blog), June 25, 2015, http://miriamposner.com/blog/humanities-data-a-necessary-contradiction/.

[5] Thomas Padilla et al., "The Santa Barbara Statement on Collections as Data (V1)," n.d., https://collectionsasdata.github.io/statementv1/.

[6] Thomas Padilla, "Always Already Computational," *Always Already Computational: Collections as Data*, 2018, https://collectionsasdata.github.io/; *Thomas Padilla, "Part to Whole," Collections as Data: Part to Whole*, 2019, https://collectionsasdata.github.io/part2whole/.

[7] "Marriott Library Collections as Data GitHub Repository," April 16, 2019, https://github.com/marriott-library/collections-as-data.

[8] "Century of Black Mormons," accessed April 25, 2019, http://centuryofblackmormons.org.

[9] Anna Neatrour et al., "A Clean Sweep: The Tools and Processes of a Successful Metadata Migration," *Journal of Web Librarianship* 11, no. 3-4 (October 2, 2017): 194-208, 111, https://doi.org/10.1080/19322909.2017.1360167.

[10] Anna L. Neatrour, Elizabeth Callaway, and Rebekah Cummings, "Kindles, Card Catalogs, and the Future of Libraries: A Collaborative Digital Humanities Project," *Digital Library Perspectives* 34, no. 3 (July 2018): 162–87, https://doi.org/10.1108/DLP-02-2018-0004.

[11] David M. Blei et al., "Latent Dirichlet Allocation," *Journal of Machine Learning Research* 3, no. 4/5 (May 15, 2003): 993–1022, http://search.ebscohost.com/login.aspx?direct=true&db=asn&AN=12323372&site=ehost-live.

[12] "Mary Nicolovo Juliana, Carbon County, Utah, Carbon County Oral History Project, No. 47, March 30 1973," Carbon County Oral Histories, accessed April 29, 2019, https://collections.lib.utah.edu/details?id=783960.

[13] "Mrs. Emile Louise Cances, Salt Lake City, Utah, Carbon County Oral History Project, No. CC-25, February 24, 1973," Carbon County Oral Histories, accessed April 29, 2019, https://collections.lib.utah.edu/details?id=783899.

[14] Nate Housley, "A Distance Reading of Immigration in Carbon County," *Utah Division of State History Blog*, 2019, https://history.utah.gov/a-distance-reading-of-immigration-in-carbon-county/.

[15] "Harold Stanley Sanders Matchbooks Collection," accessed May 8, 2019, https://collections.lib.utah.edu/search?facet_setname_s=uum_hssm; "Harold Stanley Sanders Matchbooks Collection Map," accessed May 8, 2019, https://mlibgisservices.maps.arcgis.com/apps/webappviewer/index.html?id=d16a5bc93b864fc0b9530af8e48c6c6f.

[16] Rebekah Cummings, David Roh, and Elizabeth Callaway, "Organic and Locally Sourced: Growing a Digital Humanities Lab with an Eye Towards Sustainability," *Digital Humanities Quarterly*, 2019.

[17] "Woman's Exponent Data," https://github.com/marriott-library/collections-as-data/tree/master/womansexponent; "Woman's Exponent Digital Exhibit," https://exhibits.lib.utah.edu/s/womanexponent/.

[18] John Herbert et al., "Getting the Crowd into Obituaries: How a Unique Partnership Combined the World's Largest Obituary with the Utah's Largest Historic Newspaper Database," in Salt Lake City, UT: International Federation of Library Associations and Institutions, 2014, https://www.ifla.org/files/assets/newspapers/SLC/2014_ifla_slc_herbert_mynti_alexander_witkowski_-_getting_the_crowd_into_obituaries.pdf.