Elevating "Everyday" Voices and People in Archives through the Application of Graph Database Technology

Mark Conrad
Advanced Information Collaboratory (AIC)
University of Maryland
Keyser, USA
conradsireland@gmail.com

Lyneise Williams
Department of Art and Art History
University of North Carolia at Chapel Hill
Chapel Hill, USA
lyneise@gmail.com

***Abstract*—In a simple experiment using a graph database we demonstrate that it is possible to increase the number of access points to individual items in archival collections. We do this by leveraging existing machine readable and searchable data and metadata to identify and display relationships between persons, places, dates, events, etc. across items and collections. We discuss some of the financial, ethical and representational implications of decisions made in applying technology to archival holdings. Many decisions are made without considering the ethical and representational implications. Our experiment has illustrated some of these ethical and representational implications.**

Keywords—Archives, Representation, Access, Ethics

I. The Use Case

People who are new to searching for archival (primary source) material often have expectations for that search experience based on their experience of looking for non-archival information on the internet. They want to enter a few words into a search engine related to their research topic and have immediate access to a digital surrogate for each item that is relevant to their topic. Many of them are unaware of, or choose to ignore, the fact that the search engine algorithm makes choices about what information to display. The query results are neither comprehensive nor representative.[1]

These people, we will call them consumers, are less likely to understand the limitations of the finding aids and catalogs used to provide access to archival material.[2] An archive may have thousands, millions, or billions of records in its collections. Many of them, however, are not in a machine readable or searchable form. They are in analog (non-digital) forms that cannot be retrieved or displayed over the internet. As a result, the number of access points available to the online consumer may be limited to high level summary descriptions of whole collections rather than full text indexing of individual items. Needless to say, most consumers are not satisfied with this level of service.

II. Addressing the Challenges

This scenario plays out on a regular basis. In this paper we look at some tools and practices that might make it easier for consumers to find more – not all – of the information they are looking for. Some of these tools and practices are already in use. We suggest ways these practices could be modified and more effective ways to use some of the existing tools. We also discuss some of the ethical issues around the representations that are generated from these practices and tools. We especially focus on one type of tool – graph databases – to examine their utility in increasing the number of access points across multiple collections.

The use of graph databases with archival collections is not new. The European Holocaust Research Infrastructure (EHRI) began using graph databases almost a decade ago to make connections between collections, and research, that are scattered across the continent.[3]

Likewise, "**SNAC (Social Networks and Archival Context) is a free, online resource** that helps users discover biographical and historical information about persons, families, and organizations that created or are documented in historical resources (primary source documents) and their connections to one another. Users can locate archival collections and related resources held at cultural heritage institutions around the world."[4]

These two examples, EHRI and SNAC, may help a consumer find collections that may contain information about their research topic, but they won't find individual items – documents, maps, films, photographs, etc - that contain that information and make that information instantly available over the internet.[5] That's because the information that they are looking for is often "trapped" in analog forms.

Making some of that trapped information available to anyone, anytime, anyplace is a resource intensive operation. For documents with text, the usual process is to scan the documents, then maybe employ optical character recognition (OCR) or intelligent character recognition (ICR) to get some machine-readable text out of the digital image. Depending on the quality of the output the decision may be made to use data cleaning tools to correct the mistakes in the OCR'd text. If the OCR/ICR results are really bad the decision might be made to have the text in the images manually transcribed or maybe

978-1-7281-6251-5/20/$31.00 ©2020 IEEE

simply to manually add tags to the digital images to provide some additional access points to the item.

At the end of all this work, if there is a "good" machine searchable text, the decision may be made to apply natural language processing (NLP), machine learning (ML), or artificial intelligence (AI) tools to analyze these formerly "trapped" texts to make them easier to find. Similar processes with similar decision points can be employed with digitized audio (voice recognition, NLP, ML), digitized photos (facial and object recognition, AI), digitized videos (voice, facial and object recognition, AI).

Each of the decision points in the previous two paragraphs and many others have financial, representational and ethical implications. The financial implications revolve around the resource intensive nature of moving the message from analog items to a digital form that can be searched for, reproduced and served to someone over the internet. Most archives have limited budgets, limited personnel, and limited expertise in the technologies needed. Diverting resources to make records available over the internet often means that other important work of the archives (e.g., preservation of holdings, arranging and describing new collections, in-person services, etc) may be deferred or reduced. Many archives undertaking even modest digitization projects seek outside funding to cover some or all of the costs.

The costs of making information available over the internet are substantially less for materials that are already in machine readable, searchable, and manipulable form (i.e., born digital information). There are still costs associated with making these materials available, but the expenses associated with converting the information from an analog to a digital form and cleaning up the results do not apply.

Each of the decision points above have representational and ethical implications as well. The conversion of analog materials to machine readable and searchable forms is fraught with more challenges and considerations than what is usually discussed and addressed. Guidelines for this process frequently focus on technical specifications for the conversion technology. This kind of information is a necessary component. However, it is not the only component that needs to be considered. Technical specifications are geared towards technological considerations of a material's technological aspects. That equation fails to address what and who is being represented in the material. This technological-focused model is a one-size-fits-all strategy that is potentially and actually detrimental because conversion technology impacts the variations in what/who is being represented in the analog sample.

Indeed, conversion technology does not evenly convert analog materials featuring images of people. For example, some conversion technologies, like microfilm, and technology combinations, like the digitizing of microfilm, such as what has already been used on most pre-2006 newspapers, have racial implications[6]. Microfilm is made to record extreme dark and extreme light tones---those commonly associated with text-based documents. The skin tones of most people of color fall between those two tones and into the dark tones. Photographic representations of people of color in pre-2006 newspaper archives range from distorted, minimized, to erased because of microfilm's inherent qualities. Digitizing microfilmed newspapers, which is how most newspapers have been converted to machine readable and searchable forms for digital archives, flattens out already distorted images. People of color, marginalized in U.S. society, are erased from the historical record. Who is represented is inextricably tied to decisions about transformation technology and the aesthetics of representation, or how they are represented. Specialists, like art historians, versed in a deep understanding of representation and conversion technologies are much-needed, invaluable contributors to this decision-making process.

There is much value to be gained by repurposing information that is already in machine readable form. The archival community has engaged in extended discussions about the implications of who is represented in archives.[7][1] The who referenced by many authors refers to people and populations who have been marginalized historically and currently within the archive, as well as stewarding the archive. Equally pertinent, as in this case study, is the identification and elevation of "everyday" voices and people. Archives like SNAC focus on well-known figures, or "famous" people and places. Many "everyday" people already exist in archives centered on well-known figures. We are considering and exploring strategies to identify them and amplify their presence.

III. THE EXPERIMENT

Some archives have born digital records in their holdings already. Most archives have at least some metadata (e.g., descriptions, indexes, finding aids) in digital form. There are many tools available today that can be used to manipulate and reuse that digital information to create new access points for consumers to use to find items that are directly relevant to their research.

We recently conducted an experiment using a spreadsheet, a word processing document, XML exports from an archival catalog, a text editor, and a graph database to increase access points and identify relationships between items from multiple series in the holdings of an archival repository.[2] The spreadsheet contained an index to some of the photographs in

[1] For example, see Mary A. Caldera and Kathryn M. Neal, eds.,*Through the Archival Looking Glass: A Reader on Diversity and Inclusion* (Chicago: Society of American Archivists, 2014); Shannon O'Neil, Elvia Arroyo-Ramirez, Jasmine Jones, and Holly Smith, eds., Radical Empathy in Archival Practice." *Journal of Critical Library and Information Studies* 3 (2020); Panel, "Intersectionality in Identity-Focused Archives," Society of American Archivists Annual Conference, July 28, 2017; Panel, "Immigration Archival Collections: Difference, Transnationality, and Relevance," Society of American Archivists Annual Conference, July 28, 2017.

[2] We purposely do not identify the repository or the collections because some of the records are still unprocessed and we do not have the permission of the repository to share that information.

the repository's holdings. It contained columns for the collection, series, sub-series, box number, photo id, date taken, caption, and photo credit. After manipulation using the text editor (BBEdit)[3] we created a character separated value (csv) file with columns for date taken, caption, person(s) pictured, location name, city, state, country.

The word processing document was an index to correspondence from multiple series in a collection. Using the text editor, we were able to create a csv file that contained columns for the collection, series, box, folder, sender, recipient, date and subject of the correspondence.

At this point we noticed that there were a number of similar names in both csv files. For example, there would be a Dr. Smith, Ms. Smith, Mrs. William Smith, Mrs. Annie Smith in the csv files[4]. We wanted to see if we could determine if any of these names belonged to the same person. We took a look at the XML files that had been exported from the repository's catalog. There was a file for different levels of the archival hierarchy – collection, series, folder, and item. Using the text editor we were able to isolate the authoritative list of names found in the archival collections. While this authoritative list allowed us to consolidate a few names there are more where we will have to check the actual documents to see if the names are for the same person.

At this point we loaded the two csv files into a Neo4j Desktop database.[5] Neo4j Desktop is a graph database. It stores entities (persons, places, things, etc) as nodes and relationships between the entities as edges. Using a query language (Cypher) used by Neo4j, you can explore relationships that would not be obvious if you were using tables in a relational database or csv files. One of the key advantages of a graph database is that it can display query results as a graph. Fig. 1. shows a simple example of such a graph using only a few rows from csv files holding thousands of rows.

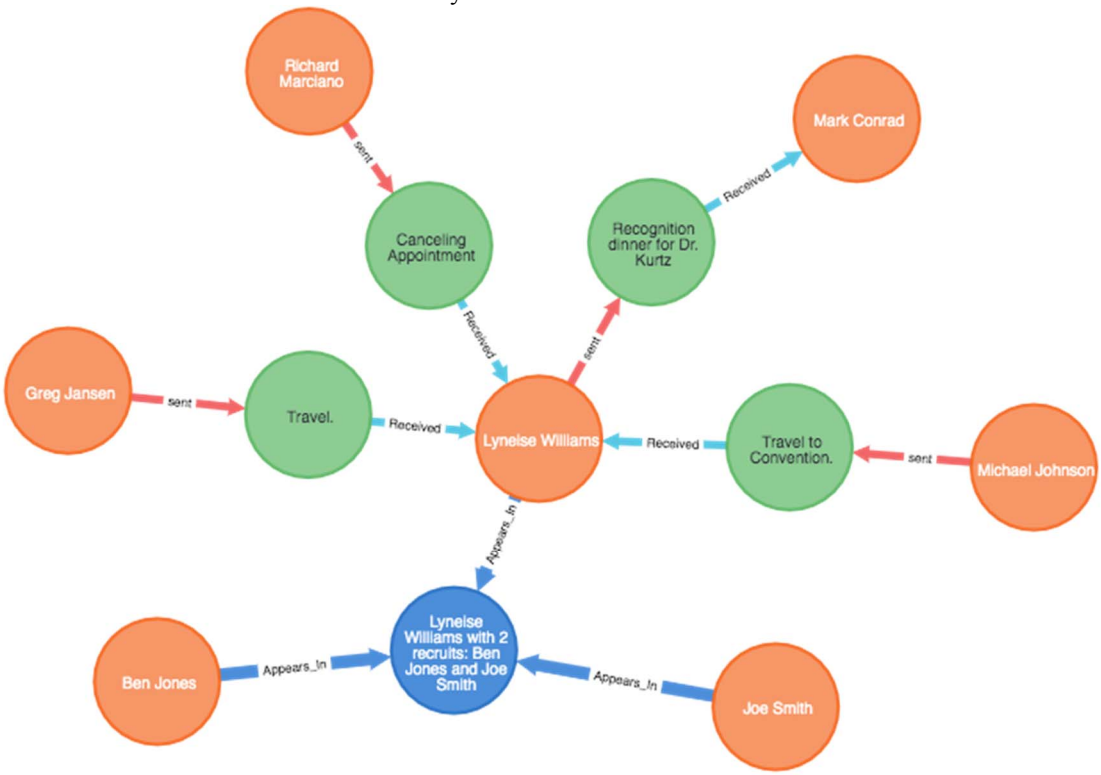

Fig. 1: Graph Connecting People with Archival Items

In Fig. 1. the circles represent the nodes – in this case persons, letters and photographs. The colors of the circles are significant. Orange represents persons, green represents letters, and blue represents photographs. The labels in each circle provide additional information about each node - in this case, personal names, subjects of the letters, and the captions of the photographs. The arrows represent the relationships between the nodes. The arrows are labeled to identify the relationship and the direction of the relationship is indicated by the arrowhead.

[3] https://www.barebones.com/products/bbedit/index.html Mention of a specific tool should not be taken as an endorsement of that tool. These are simply the tools we had readily available for this experiment.

[4] These names are fictitious and are used for illustrative purposes only.

[5] https://neo4j.com/

Using the information available in Fig. 1. it is easy to understand what is being displayed about the relationships between the different entities. For example, Greg Jansen sent a letter about travel and Lyneise Williams received that letter. We can also see that Lyneise Williams appeared in a photograph, we can see the caption for that photograph and we can see that two other persons have also been identified as appearing in the picture.

We can learn a great deal by looking at this simple graph. Each green and blue circle represents an archival item. For each item we see the names of at least some of the persons associated with that item. By examining each person node we can see items that they are associated with throughout the repository's holdings. In addition, clicking on one of the blue or green nodes will display additional information about that item (e.g., location of the item, date, etc).

Despite the use of already processed machine readable and searchable data, there are difficulties associated with the strategies we propose. Scaling up such markups by manually examining, identifying potential access points and tagging them in individual items requires additional expenses and labor. Specialists attuned to searches for identifying specific individuals and cultural attributes are necessary to the accuracy of the conversion process.

IV. Conclusion and Future Steps

We have demonstrated in a simple experiment that it is possible to increase the number of access points to individual items in archival collections by leveraging existing machine readable and searchable data and metadata. It is also possible to identify and display relationships between persons, places, dates, events, etc across archival items and collections. This experiment was conducted in a matter of hours using readily available data and a few simple tools.

We have discussed some of the financial, ethical and representational implications of decisions made in applying technology to archival holdings. Many decisions are made without considering the ethical and representational implications. Our experiment has illustrated some of these ethical and representational implications.

The csv file we assembled from a pre-existing index to photographs, contained lists of photographs. Finding aids may only list the name of the well-known person who the collection features. Viewing the photographs demonstrates that many may depict more individuals besides the well-known person. Identifying and tagging that person renders them machine readable and searchable.

The graph database used to identify and demonstrate relationships between the well-known person and others in the collection can also be implemented to center everyday people. Connecting those individuals to well-known people, places, and things, as well as across collections provides greater access to varying viewpoints, authorities and interpretations.

Beyond additional people, the items may depict particular objects like sculptures, books, furniture, and even architecture that can serve to illuminate the cultural values and preferences of those imaged. Tagging and identifying these elements offers not only more access points but increased possibilities for deepening our understanding of the contexts constructed by the individuals and from which the individuals engage.

One of the next steps we would like to take is to find ways to create access points for the "non-famous" persons and objects in the collections. For machine readable and searchable collections, we will consider using Named Entity Recognition (NER) tools (e.g., ANNIE: a Nearly-New Information Extraction System[8] and GATE: general architecture for text engineering[9]) for extracting more names, dates, events, and places than are found in the indexes we have used so far.[10] For digital photographic images we may apply some of the tools from the Brown Dog project[11] (e.g., OpenCV[12] or Discriminatively Trained Deformable Part Models[13]) to try to detect more access points in the images. We will also look for more ways to harvest additional access points from existing sources of machine-readable data and metadata held by the repository using our current methods. Finally, we will consider the aesthetic aspects of machine-ready data with an eye towards appropriate depictions of individuals from an ethical perspective.

Acknowledgment

The authors thank Richard Marciano and Greg Jansen for their comments, pointers, and support.

References

[1] Algorithms of Oppression: How Search Engines Reinforce Racism *by Safiya Umoja Noble. NYU Press, 2018*

[2] "The role played by those persons, or client systems, who interact with OAIS services to find preserved information of interest and to access that information in detail." ISO 14721:2012 – Reference Model for an Open Archival Information System, Section 1.7.4.

[3] T. Blanke, Bryant, M. , and Speck, R. , "**Developing the collection graph**", *Library Hi Tech*, vol. 33, pp. 610-623, 2015.

[4] SNAC (Social Networks and Archival Context) https://portal.snaccooperative.org/about

[5] See, Teddy Randby and Richard Marciano, Digital Curation and Machine Learning Experimentation in Archives, IEEE Big Data 2020, for another example of using technology to provide better item level finding aids.

[6] Lyneise Williams, "What Computational Archival Science Can Learn from Art History and Material Culture Studies," *2019 IEEE International Conference on Big Data (Big Data)*, Los Angeles, CA, USA, 2019, pp. 3153-3155, doi: 10.1109/BigData47090.2019.9006527.

[7] For information about ethical representations in archives see VERA Collaborative (Visual Electronic Representations in the Archive) https://veracollaborative.com

[8] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02). Philadelphia, July 2002.

[9] H. Cunningham, V. Tablan, A. Roberts, K. Bontcheva (2013) Getting More Out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics. PLoS Comput Biol 9(2): e1002854. doi:10.1371/journal.pcbi.1002854

[10] See earler work in this area, William Underwood and Sheila Isbell, Semantic Annotation of Presidential E-Records, Technical Report

ITTL/CSITD 08-01, May 2008 https://www.archives.gov/files/applied-research/papers/semantic-annotation.pdf

[11] S. Padhy, G. Jansen, J. Alameda, E. Black, L. Diesendruck, M. Dietze, P. Kumar, R. Kooper, J. Lee, R. Liu, R. Marciano, L. Marini, D. Mattson, B. Minsker, C. Navarro, M. Slavenas, W. Sullivan, J. Votava, K. McHenry, **"Brown Dog: Leveraging Everything Towards Autocuration"**, IEEE Big Data, 2015

[12] You can find information on the OpenCV project here: Kari Pulli, Anatoly Baksheev, Kirill Kornyakov, and Victor Eruhimov. 2012. Realtime Computer Vision with OpenCV: Mobile computer-vision technology will soon become as ubiquitous as touch interfaces. Queue 10, 4 (April 2012), 40–56. DOI:https://doi.org/10.1145/2181796.2206309 and information about the Brown Dog scripts that work with OpenCV here: https://browndog.ncsa.illinois.edu/toolscatalog/tools/54f87fe02200005701332b88

[13] You can find more information about Discriminatively Trained Deformable Part Models here: P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan, "**Object Detection with Discriminatively Trained Part Based Models",** IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 32, No. 9, Sep. 2010 and more information about the associated Brown Dog script here: https://browndog.ncsa.illinois.edu/toolscatalog/tools/54f92ba92200006902332b9b