Case Study: Testing Computational Archival Science frameworks using AI tools in analyzing the Spelman College Archives photograph collection

Kaila Fewster[1]

Educational applications: This case study illustrates how AI tools can be used in archives to support context preservation in digital image or other non-textual collections as well as open-linked data. It also highlights the importance of high-quality metadata in improving digital image collection accessibility, and the need for archivists and records managers to have a strong grasp of algorithmic thinking, and the basic data science skills when working with this technology. By providing insight into the value of computer vision in archives, this case also emphasizes the importance of critically evaluating AI tools in the archives from ethical perspectives to look for bias and discrimination built into the code.

Educational topics: AI for non-textual records (photographs), types of AI/ML for photographs, biases in AI/ML algorithms, AI for archival description and access[2].

About: This case study is part of a series of learning materials developed by InterPARES Trust AI[3] researchers and educators to train archival professionals and students to effectively leverage artificial intelligence in their archival work. The final draft was completed on October 21st, 2023. It has a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International BY-NC-SA 4.0 license, which requires that reusers give credit to the creator. It allows reusers to distribute, remix, adapt, and build upon the material in any medium or format, for noncommercial purposes only. If others modify or adapt the material, they must license the modified material under identical terms.[4]

This case study addresses the issues associated with preserving context in digital photographic collections through a Computational Archival Science (CAS) framework at the Advanced Information Collaboratory at the University of Maryland. Since photographs typically carry little to no textual information, this often translates into a lack of metadata and other contextualizing information. As a result, this lack of information complicates connecting the items to Linked Open Data, which allows archivists to create access points throughout their collections and better integrate and reflect community voices in archival descriptions. As such, this project explores whether computational methods can improve textual and photographic resource integration to bring more context surrounding photographic collections into the metadata by

[1] InterPARES Trust AI Graduate Academic Assistant, University of British Columbia.

[2] Educational applications map to a Body of Knowledge proposed by InterPARES researchers for AI/ML for the archival professionals. https://docs.google.com/document/d/1UsjkkkGeSJrgCDJGASCAy5q0uo_ZkQpzi_Ch8XUcqYw/edit?usp=sharing

[3] This case study is an outcome of InterPARES Trust AI, an international research partnership led by Drs. Luciana Duranti and Muhammad Abdul-Mageed, University of British Columbia, and funded by the Social Sciences and Humanities Research Council of Canada (SSHRC). https://interparestrustai.org/

[4] Case Study: Testing Computational Archival Science frameworks using AI tools in analyzing the Spelman College Archives photograph collection © 2024 by Fewster, Kaila is licensed under CC BY-NC-SA 4.0. To view a copy of this license, visit https://creativecommons.org/licenses/by-nc-sa/4.0/

developing a framework to enhance linking people, places and events depicted in historical collections.

In order to test the CAS framework, the project researchers chose the Spelman College Archives Photograph Collection as their dataset, containing over 1200 images dating from 1880 to 2007. The team then statistically analyzed the entire collection using the OpenRefine software, with the goal of consolidating its metadata and yielding aggregate information about the collection. The final sample of 40 photographs and their associated metadata depict different eras, styles and subjects and roughly represent the range of the collection-as-a-whole. Once chosen, the sample underwent a textual analysis of the existing metadata to develop item-level descriptions, which were then linked in a web-like structure through semantic processing. The sample also went through a process of image analysis, which helped to create baseline accuracy measures for the facial recognition tools used in this framework. Establishing these accuracy measures is necessary as computer-vision algorithms have traditionally struggled recognizing people of colour, who make up much of the population captured in the Spelman Photograph Collection. As a result, this ensured that the system could reliably recognize people and faces in images and subsequently use this information to link individuals in the social web. Finally, the data gleaned from the sample through the textual, semantic and image analyses was linked with the broader Open Linked Data available online.

Although the study was initially conceived as a proof of concept rather than a practical and applicable tool, this project illustrates that the methods used in the CAS framework are scalable to actual archival practice. In this sense, as the process of creating linked data is usually prohibitively resource-intensive, incorporating automation through machine learning in this context can have a significant practical impact. Furthermore, the creation of this framework makes it easier for archives and other cultural heritage organizations to adopt Linked Data approaches in their collections, hopefully leading to more automation opportunities in these spaces in the future.

Potential Discussion Questions:

1. What ethical considerations arise from the use of computational methods and facial recognition in archival work, and how should archivists and records managers address these concerns while working with these technologies?
2. In what ways do the biases in current facial recognition algorithms that, in particular, disproportionately affect people of colour impact the accurate representation of diverse communities in digital archival collections?
3. What are the benefits and limitations of using semantic processing to create a web-like structure for linking archival materials, how can this approach be refined or expanded?

References

Arnold, T. (2023). *Distant Viewing Toolkit for the Analysis of Visual Culture* [Python]. Distant Viewing Lab. https://github.com/distant-viewing/dvt (Original work published 2017)

Atlanta University Center. (n.d.). *Spelman College Photographs Collection*. Retrieved October 21, 2023, from https://radar.auctr.edu/islandora/object/sc.002%3A9999/

Proctor, J., & Marciano, R. (2021, December 15). *An AI-Assisted Framework for Rapid Conversion of Descriptive Photo Metadata into Linked Data*. IEEE International Conference on Big Data. https://doi.org/10.1109/BigData52589.2021.9671715