Machine Learning as an Archival Science:
Narratives behind Artificial Intelligence, Cultural Data, and Archival Remediation

Giulia Taurino 1,2
David A. Smith 1
1 Khoury College of Computer Sciences, Northeastern University / US
2 Institute for Experiential AI, Northeastern University / US

In contemporary Western anthropology and sociology, the concept of culture has evolved from being perceived as an abstract subject—culture as ideas, concepts, knowledge (Kroeber and Kluckhohn 1952)—to a concrete and observable matter (White 1959). Culture, therefore, is often presented as the history of its apparatuses—whether in the form of inherited traditions, learned behaviors, social habits (Bourdieu 1977), or as the supra-individual organizational structures and institutional systems that make up societies (Steward 1955), or else in the more or less homogeneous networks of traces left by human activity (Lorusso 2015). These traces—*monumenta*, memorials, signs and other "things-and-events-dependent-upon-symboling" (White 1959: 230)—lead eventually to the creation of archives of various forms (Briet 2006 [1951]). Over time, the study of archival records in particular has gained importance as valuable insight for reconstructing histories, building collective memories, witnessing the present, and even predicting the future of human cultures. Indeed, in the digital age, the very act of collecting and documenting culture through signs and symbols has blended with the notion of culture itself: web archives, online repositories, and media platforms all focus on processes of reorganization of data, and data *is* culture.

In this paper we tackle the use of AI for cultural preservation, processing, management, and retrieval in archival contexts, while also looking at how emerging sets of AI applications in the GLAM sector might shift dominant approaches and perspectives in the culture of AI at large. We notably discuss two main frameworks that have emerged in AI as a tool for curatorial practices: one that tends towards data disaggregation and one that promotes reaggregation of content. On the one hand, earlier attempts of "archival rehabilitation"—that is to say, the revival of cultural records that are made available in more durable digital formats—have been defined by a tendency to disaggregate the archive into fragments of digitized material. Examples of disaggregation practices in GLAM collections can be found in many library and museum initiatives that repost digitized images on online platforms as single records, devoid of the archival context of origin, or else in computational art projects that reuse individual images for creative purposes using AI-based generative methods. On the other hand, attempts to partially reassemble cultural records in digital scholarship started with a renewed interest for the spatial turn in humanities, since researchers have been able to process geolocated data using open-access computational tools. Overall, initiatives for reactivating archival collections have emerged in close connection with new technologies, with the intent of turning cultural artifacts into machine-readable, searchable, retrievable data.

Throughout the years, many archives have published online digitized images from their repositories, an ongoing process that sparks further practices of archival remediation inside and outside the institution of origin. More recently, with the introduction of open-access tools able to process large datasets, the archiving of cultural artifacts took a different approach: the interest among digital humanities scholars in clustering objects across collections resulted in an abundance of research projects tinkering with topic modeling, textual classification, object detection, and other ways to find similarities among records by creating assemblages and re-aggregating the archive leveraging on machine learning techniques. In order to show how computer vision can be used to work with both textual and visual data from archival collections, we take as a case study a project developed at Northeastern University. Building upon the digitization of millions of published and unpublished images from the Boston Globe photo collection, this project utilizes computer vision to recover handwritten information present in the physical archive (data archeology), highlight patterns across images via automated sorting (ML-augmented classification), and finally reassemble fragmented records into coherent narrative paths (narrative-based information retrieval). Rather than disaggregating images into separate units, we use machine learning to create a cataloging system able to recover and enrich the metadata available, thus rendering each record in the photo-archive as a fluid text (Bryant 2002) that can be manipulated and reorganized into ever new aggregates.

This project takes a critical stance on existing discourses in the field of AI and culture that privilege the use of machine learning uniquely for content filtering and predictive purposes. The work on the Boston Globe archive expands the conversation on practices of categorization, by showing that computer vision (e.g. OCR, image recognition) applications in the study of cultural histories and archival material have more implications than just the indexing of isolated objects. Much like archival and curatorial practices, machine learning activates a transformative process, it helps historicize, remediate and recontextualize records as part of the archive as a whole in a generative and dynamic way. While internet culture have been stressing on the creation of search engines and recommendations system for retrieving content, public records are still confined between two extremes: either in the form of decontextualized objects at risk of misinterpretation, or else in the form of infinite queues that might result in information bias. Our project proposes to tackle this fault in algorithmic design, by returning to an exploratory mode where multiple narratives and perspectives can coexist in the same archival space through processes of algorithmic-human co-curation (Hendricks 2017; Dekker and Tedone 2019). For instance, training machine learning models on the Globe collection will allow us to document past curatorial practices of the newspaper's photo librarians, retrace the editorial selection from photo-assignments and propose archival paths supported by several historical documents. It also promotes dialogue between archives, by relying on metadata and image recognition, instead of unique identifiers as seen in ontology-based curatorial practices (Menghini *et al.* 2019).

A closer look at archival practices in recent years reveals a history of constant cultural remediation. More and more, archival work relies on transformative operations that turn analog objects into digital records, in need to be reorganized and recontextualized. The archive as we understand it today goes beyond the practical activity of storage, preservation, and categorization of single items in a given physical collection. It acquires new conceptual challenges and responsibilities in redefining cultural memories and imaginations. By deploying computer vision as a form of archival remixing, with this project we make a case for machine learning as an archival science. In the wake of previous work that discusses archival strategies for socio-cultural datasets (Jo and Gebru 2019) and approaches big data as archives themselves (Thylstrup et al. 2021), we call for an archival turn in AI. By engaging machine learning with archival collections and cultural records, we present a novel framework for understanding the narratives behind artificial intelligence, cultural data and archival remediation. In a moment when digitization is offering the possibility of turning cultural data into a media format that we can manipulate and remediate by means of computational tools, artificial intelligence and archival studies can cooperate in fruitful ways. By intersecting the field of machine learning with that of archival studies we ultimately hope to contribute in laying the groundwork for an innovative, experimental domain that still needs to be explored: AI in culture, or maybe better, AI *for* culture.

References:

Bourdieu, P. (1977). *Outline of a Theory of Practice* (R. Nice, Trans.; 1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511812507

Briet, S. (2006). *What is documentation? English translation of the classic French text*. Scarecrow Press.

Bryant, J. (2002). *The Fluid Text: A Theory of Revision and Editing for Book and Screen*. University of Michigan Press. https://doi.org/10.3998/mpub.12024

Dekker, A., & Tedone, G. (2019). Networked Co-Curation: An Exploration of the Socio-Technical Specificities of Online Curation. *Arts*, *8*(3), 86. https://doi.org/10.3390/arts8030086

Hendricks, Manique. (2017) “The Algorithm as Curator: In Search of a Non-Narrated Collection Presentation” Stedelijk Studies Journal 5 . DOI: 10.54533/StedStud.vol005.art11

Jo, E. S., & Gebru, T. (2020). Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 306–316. https://doi.org/10.1145/3351095.3372829

Lorusso, A. M. (2015). *Cultural Semiotics*. Palgrave Macmillan US. https://doi.org/10.1057/9781137546999

Meghini, C., Bartalesi, V., Metilli, D., & Benedetti, F. (2019). Introducing narratives in Europeana: A case study. *International Journal of Applied Mathematics and Computer Science*, *29*(1), 7–16. https://doi.org/10.2478/amcs-2019-0001

Kroeber, A.L. and Kluckhohn, C. (1952) Culture: A Critical Review of Concepts and Definitions. Peabody Museum, Cambridge, MA.

Steward, J. H. (1955). Theory of Culture Change: The Methodology of Multilinear Evolution (244 p). Urbana, IL: University of Illinois Press.

Thylstrup, N. B. (Ed.). (2021). *Uncertain archives: Critical keywords for big data*. The MIT Press.

White, L. A. (1959). The Concept of Culture*. *American Anthropologist*, *61*(2), 227–251. https://doi.org/10.1525/aa.1959.61.2.02a00040

White, L. A. (1959). The Concept of Culture. *American Anthropologist*, *61*(2), 227–251. http://www.jstor.org/stable/665095