Journal of Research in Health Sciences
ISSN: 2228-7809 | https://jrhsjournal.com

Computational Approaches to Archival Data Extraction: An Archivally-Informed Framework

Dr. Thabo Ndlovu[1]*, Dr. Ayesha Khan[2]
[1]Department of Public Health, University of Cape Town, Cape Town, South Africa
[2]Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

Abstract

Computational approaches to archives have gained traction in recent years. Archivists are using computational technologies to appraise, arrange and describe, and create access to digital records. While computational tools and methods have proven useful in specific cases, they have only recently been adopted by archivists, and the complexities of their implementation in an archival context is only beginning to be understood. This paper discusses a pilot project to implement computational approaches to digitized archival records: The Cybernetics Thought Collective: A History of Science and Technology Portal Project, was a collaborative multi-institutional initiative funded by the National Endowment for the Humanities (US). As an example of using computational methods on archives, this paper seeks to contribute to discussions about how computational archival projects can employ models that are informed by archival practices and that generate trustworthy data.

Keywords – Archives; data; computational archival science; machine learning; natural language processing; cybernetics.

1 Introduction

For archivists, computational approaches to archives present potentially complementary and efficient ways to appraise and curate digital materials. Despite recent projects to implement and use computational tools and methods, questions remain about where computational archival work fits within archival theory and practice. For example, technologies like machine learning and natural language processing have potential for enhancing access to archival materials, but questions remain about the best way to use machine-extracted metadata that aligns with descriptive standards. Likewise, providing access to outputs of computational pipelines in ways that align with the spirit and ethics of communities of practice like Collections as Data (Padilla et al., 2018) is important for meeting new and emerging research needs. Above all, best practices for generating data through archival projects need to be developed and widely adopted.

This paper discusses a pilot project to implement computational approaches to digitized archival records. The Cybernetics Thought Collective: A History of Science and Technology Portal Project was a collaborative multi-institutional initiative funded by the National Endowment for the Humanities (US). Over a two-year period (2017-2019), participating institutions (the University of Illinois at Urbana-Champaign, American Philosophical Society, British Library, and MIT) used machine learning and natural language processing technologies to extract and classify data from 61,067 pages of digitized correspondence and journals from four founding members of the field of cybernetics. This data was used to create metadata for the digitized materials as well reveal latent connections between the cyberneticians' papers. The project presented a number of challenges and questions about how to document the many decision-points in computational workflow and provide transparency about the results of the machine-learning process. It also raised questions about how to best develop a computationally-driven pipeline that is tailored specifically for deployment in an archival context.

2 Background

Cybernetics was a revolutionary and transdisciplinary scientific movement in the twentieth century. Its engineers, computer scientists, anthropologists, and philosophers asked new questions, conducted thought experiments with cybernetic machines, and posited new ideas that pushed the boundaries of how we understand the relationships between nature, culture, and machines. As a transdisciplinary space, cybernetics developed distinct concepts to explore behavior and feedback. Cyberneticians likewise developed a specific vocabulary for articulating cybernetic concepts. The Cybernetics Thought Collective (CTC) project sought to use this vocabulary as the basis for inputs for a computational analysis pipeline to not only generate metadata from the records themselves, but to also

reveal connections between the records around these cybernetic concepts.

2.1 A Brief History of Cybernetics

Cybernetics found its beginnings in ten conferences hosted by the Josiah Macy, Jr. Foundation between 1946 and 1953. These meetings included scholars from fields in the natural, physical, social, and information sciences. Known as the Macy Conferences on "Circular Causal and Feedback Mechanisms in Biological and Social Systems," the meetings eventually adopted "cybernetics" as the unifying theme following the publication of Norbert Wiener's *Cybernetics: Or the Control and Communication in the Animal and the Machine* (1948).

Research in cybernetics initially sought to develop automated defense mechanisms during the Second World War, but cybernetics evolved for different uses and meanings over time. For example, given its utopian and counterculture leanings during the 1960s and 70s, it is no surprise that Stewart Brand's *Whole Earth Catalog* drew from cybernetic concepts amid new social currents (Turner, 2006). As the "new science," cybernetics spread broadly across the United States, Europe, South America, and the Soviet Union, where it was applied in different social and political contexts (Kline, 2005).

Given its breadth and ambitious aims, cybernetics has also been defined as a "universal science" (Bowker, 1993). But at its heart, cybernetics aims to study communication and control as manifested through behavior in organisms, machines, and social systems (Wiener, 1948; Ashby, 1956). Its provided participants from the Macy Conferences with a shared vocabulary to discuss similar phenomena across disciplinary divides. In addition to serving as a forum for established scholars such as Margaret Mead, Claude Shannon, and Greg Walter, the Macy Conferences also created a space for newer voices, such as Austrian émigré and physicist Heinz von Foerster (1911-2002). Von Foerster later established the Biological Computer Laboratory (BCL) at the University of Illinois at Urbana-Champaign, which became an important site for engagement around "second-order cybernetics" (Umpleby, 2005).

2.2 The Cybernetics Thought Collective Project

The Macy Conferences created a network between its participants who developed long-term relationships and exchanged correspondence. In a scientific context, Ludwik Fleck describes a phenomenon as a "thought collective" which consisted of scientists who form a network interrogate and exchange ideas (Fleck, 1979). While this exchange can take several forms, one way in which ideas were shared and debated was through the exchange of letters. Because cybernetics' international extent, however, the archives of cyberneticians are geographically dispersed. As means to unite these dispersed archives, the University of Illinois Archives, the American Philosophical Society, British Library, and MIT Distinctive Collections received a grant a from the National Endowment for the Humanities for a pilot project to digitize the papers of cyberneticians and make them accessible online. Each repository holds the papers of a founding member of the cybernetics movement—Heinz von Foerster (a physicist), Warren S. McCulloch (a neurophysiologist), W. Ross Ashby (a psychiatrist), and Norbert Wiener (a mathematician), respectively—who collectively formed a part of the larger cybernetics correspondence network.

Beyond digitizing and making the materials available online, the CTC project sought to use computational tools and methods on them to enhance access. Cybernetics' distinct vocabulary developed in association with its key concepts could serve as ready inputs for a computational analysis pipeline. Utilizing a computational approach to cyberneticians' archives ultimately served two main purposes: to generate data that could be used as metadata and to form the basis of a classification model that would illustrate connections between correspondence and related records across the archives forming this corpus. While machine-extracting data for such purposes is new territory for archivists, computational approaches are important to explore given new and emerging research needs to engage with computable archives within digital scholarship frameworks (Harris et al., 2020).

3 Approach

The four repositories digitized a select amount of correspondence and related materials for the purposes of the pilot project. A PDF file was created for each folder-level aggregation or multi-page item to align with archival descriptive practices across the four repositories (i.e., that do not describe materials at the item- or page-level). While the majority of the content is typewritten, some materials are handwritten. Typed materials were processed by optical character recognition (OCR) programs to make them machine-readable, while handwritten materials were manually transcribed as allowed within the project's limited timeframe.

The digitized materials also need to be normalized and transformed into plaintext files. This normalization entailed remediation of inconsistencies and errors

resulting from OCR as well as translation of some of the texts into English. Cybernetics' international extent meant that a number of letters were written in German, Spanish, French, and Italian. One of the programmers hired for the project used a combination of PDFMiner, Python tools, dictionaries, and Googletrans to remediate and extract the OCRed texts into plaintext and make them ready for entity extraction, natural language processing (NLP), and machine learning (ML) tools.

Before computational tools could be used on the texts, the project team needed to create inputs. As aforementioned, cybernetics generated a distinct vocabulary for its concepts. The CTC team sought a source for this vocabulary, but one that would reflect the dialogues around cybernetics at the time that von Foerster, McCulloch, Ashby, and Wiener were active. The team selected *Cybernetics of Cybernetics: Or, the Control of Control and the Communication of Communication* (von Foerster, ed., 1974), which contains discussions of the cybernetic ideas of the four cyberneticians as well as their contemporaries. To generate a list of cybernetic terms that could be used as inputs, *Cybernetics of Cybernetics* was uploaded to Voyant Tools, which produced a list of frequently occurring terms. This list was narrowed to the approximately 200 most frequently occurring terms and shared with the CTC project's advisory board (who comprised experts on the history of cybernetics) before finalizing.

The CTC project's programmer experimented with different Python libraries for entity extraction and NLP. A plaintext file containing extracted entities for each digital object (representing the digitized folder-level aggregations or multi-page items) was created. These entities were used as metadata for the digitized materials as well as the basis for a supervised ML model. Using a combination of Naïve Bayes and Weka, the materials were classified into four categories: Mathematics/Logic; Computers/Machines; Psychology/Neuroscience; and Personal Correspondence. This process is illustrated in Figure 1, but it is important to note that this was not a linear process; many steps were revisited as new tools were investigated to enhance the results and troubleshoot issues.

Percentages of certainty for the classification were generated and included along with the machine-extracted metadata in the University of Illinois Digital Library, where the digitized materials and project outputs have been made accessible (https://digital.library.illinois.edu/collections/38ec6eb0-18c3-0135-242c-0050569601ca-1). In addition to providing access through a digital collections portal, the machine-extracted data were also used to test visualization software as an access mechanism, which are available in the project's prototype portal (https://archives.library.illinois.edu/thought-collective/).

The methodology and process is described in more detail in the CTC project white paper (Anderson et al., 2019). In addition, a readme note delineating the process and the metadata fields (and which metadata was "machine-generated" or "human-generated) as well as the original inputs and outputs, have been made available along with the digitized materials (https://digital.library.illinois.edu/items/3c80ad40-8c95-0138-729a-02d0d7bfd6e4-b).

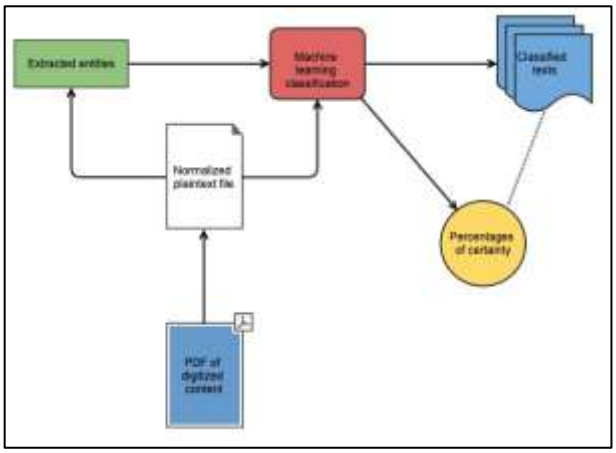

Figure 1: Illustration of the process employed to extract entities and classify the archival materials.

4 Machine-Extracting Data for Archives

The CTC project raised a number of questions about where computational archival work fits within emerging trends in data science in archives, and more broadly within archival theory and practice. For example, when utilized as an aid to description, machine learning blurs the line between "data" and "metadata"; named entities are extracted as *data*, and can be reused by researchers for their own computational analyses, but they can also be used as *metadata* once they form part of a metadata profile for a digital object. Computational approaches to archives also raise questions about how data generated from such projects should be documented, described, and made accessible.

When archivists appraise, arrange and describe, and make preservation decisions about the materials they steward, the decisions they make should ensure the reliability, authenticity, and ultimately the trustworthiness of that evidence. Ascertaining records' trustworthiness can be in many ways more complicated in the digital age (MacNeil, 2000). Because computational approaches are so opaque and akin to a "black box" (Seles, 2020), determining the trustworthiness of the outputs can be difficult for archivists

and researchers alike. On the one hand, this makes it vital for archivists to document the process employed and to make a description of that process accessible and understandable to researchers (Seles, 2020); on the other hand, it is equally vital for archivists to be closely involved in the implementation and use of computational algorithms (Hutchinson, 2020). While archivists cannot be expected to learn how to develop code and create NLP and ML tools themselves, it is important to for practitioners to understand the nature of the algorithms and to take time and care to generate the inputs. But this can take significant amount of time and preparation (Rolan et al., 2019).

Presenting extracted data as a product from a machine may inadvertently give the impression that such data are objectively-created outputs devoid of human subjectivity or influence. This can lead to assumptions that the data-as-metadata are somehow more authentic or accurate as descriptions of materials. Likewise, it can also lead to assumptions about the data being neutral products devoid of bias (Mordell, 2019). Documenting the process can illustrate the amount of human intervention in that process and fade the opaqueness surrounding computational tools. For example, the CTC project described the process that generated the data, but in retrospect key decision-points should have been better documented and described, such as which materials were selected for inclusion in the project and why, and which software was ultimately selected for inclusion in the computational analysis pipeline.

Documenting the process can also better facilitate reuse of the data by researchers wishing to analyze them with their own tools, or those wishing to replicate the process to ascertain the trustworthiness of the data. But the description of the process must also contain information about decision-points so that researchers can understand how those decisions might have affected the output. We can learn an important lesson here from science itself; documenting the process has long been a cornerstone of science to facilitate reproducibility and reuse. It is also important for those who use computational methods (Stodden et al., 2016). Such documentation of the process can help us better understand and develop models that are informed by archival principles and create access to trustworthy archival evidence and data.

5 Conclusion

Computational tools and methods potentially enable deeper (and perhaps different) engagements with archives. However, they must be employed (and developed) in ways that enable archivists to adapt them for use on archival materials. And to use computational tools responsibly, archivists must describe the processes that generated the data to facilitate reuse and reproducibility, and provide access to trustworthy output. It has long been known that scientific data and facts are socially constructed (Latour and Woolgar, 1979), and that different contextual factors shape and influence the results. Archival data are no different.

Acknowledgements

“The Cybernetics Thought Collective: A History of Science and Technology Portal Project” (NEH PW-253912-17), was funded by the National Endowment for the Humanities’ Humanities Collections and Reference Resources program (US).

References

Anderson, B.G., Prom, C.J., Hutchinson, JA, Chandrashekhar, A., Michael, B., Udhani, S., Sammons, M., Dolski, A., Hamilton, K., Kaushik, S., and Shrivastava, M. (2019). The Cybernetics Thought Collective: A History of Science and Technology Portal Project White Paper. [online] Available at: https://www.ideals.illinois.edu/handle/2142/106050.

Ashby, W. R. (1956). An Introduction to Cybernetics. London: Chapman and Hall.

Bowker, G. (1993). How to Be Universal: Some Cybernetic Strategies, 1943-1970. Social Studies of Science 23(1), 107-127.

Fleck, L. (1979). Genesis and Development of a Scientific Fact. Edited by Thaddeus J. Trenn and Robert K. Merton. Translated by Fred Bradley. Repr. 11. Aufl. Sociology of Science. Chicago: Univ. of Chicago Press.

Harris, G., Potter, A., & Zwaard, K. (2020). Digital Scholarship at the Library of Congress, Library of Congress, March 17, 2020. [online] Available at: https://labs.loc.gov/static/labs/work/reports/DHWorkingGroupPaper-v1.0.pdf.

Hutchinson, T. (2020). Natural Language Processing and Machine Learning as Practical Toolsets for Archival Processing. Records Management Journal. [online] Available at: https://doi.org/10.1108/RMJ-09-2019-0055.

Kline, R. (2015). The Cybernetics Moment: Or Why We Call our Age the Information Age. Cambridge: The MIT Press.

MacNeil, H. (2000). Trusting Records: Legal, Historical, and Diplomatic Perspectives. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Mordell, D. (2019). Critical Questions for Archives as (Big) Data. Archivaria 87, 140-161. [online] Available at: https://archivaria.ca/index.php/archivaria/article/view/13673.

Padilla, T., Allen, L., Frost, H., Potvin, S., Russey Roke, R., & Varner, S. (2019). Santa Barbara Statement on Collections as Data --- Always Already Computational: Collections as Data

(Version 2). Zenodo. [online] Available at: http://doi.org/10.5281.zenodo.3066209.

Seles, A. (2020). "Artificial Intelligence and Archives," Emerging Technologies, Big Data and Archives Webinar Series, Council on Library and Information Resources. [online] Available at: https://www.youtube.com/watch?v=noxwKS-cPh0&feature=youtu.be.

Stodden, V., McNutt, M., Bailey, D.H., Deelman, E., Gil, Y., Hanson, B., Heroux, M.A., Ioannidis, J.P.A., & Taufer, M. (2016). Enhancing Reproducibility for Computational Models. Science, 354, 1240-1241. [online] Available at: https://www.doi.org/10.1126/science.aah6168.

Turner, F. (2006). From Counterculture to Cyberculture: Stewart Brand, the Whole Earth Network, and the Rise of Digital Utopianism. Chicago: University of Chicago Press.

Umpleby, S. (2005). A History of the Cybernetics Movement in the United States. Journal of the Washington Academy of Sciences, 91(2), 54-66.

Von Foerster, H (ed). (1974). Cybernetics of Cybernetics, or the Control of Control and the Communication of Communication. Urbana, IL: Biological Computer Laboratory.

Wiener, N. (1948). Cybernetics: Or, Control and Communication in the Animal and the Machine. New York: J. Wiley.