The Tie that Binds: Developing a Scholarly Edition of Seventh-day Adventist Literary Systems

Jeri E. Wieringa[1]

[1] *The University of Alabama, Tuscaloosa, Alabama, United States*

Abstract

Computational text analysis with historical documents raises several challenges, from the scope of the digitized sources to the quality of the text derived through OCR. This paper builds on the work of Katherine Bode to argue for the utility of developing "scholarly editions" of literary systems as a mechanism for providing historical and technical context to the text used for computational analysis, with examples from my research on the print materials produced by the Seventh-day Adventist denomination between 1849 and 1920.

Keywords

Computational text-analysis, historical OCR, scholarly editions, Seventh-day Adventists

1. Introduction

In this paper I discuss the framing and motivation for my current project, *The Tie that Binds: A Scholarly Edition of Seventh-day Adventist Literary Systems.* Large-scale computational text analysis presents the opportunity to surface and examine patterns in historical texts, and in the case of religious studies, trace the cultural development through written materials. However, the known limitations of historical OCR and the unclear boundaries of available digital collections raise challenges for determining whether the patterns one finds in textual data have sufficient basis in the historical record. Informed by Katherine Bode's work with the Trove database in *A World of Fiction: Digital Collections and the Future of Literary History*, I am creating a scholarly edition of print materials produced by the Seventh-day Adventist denomination that identifies the documents available within the edition, the correspondence between the digital corpus and the known print materials, the quality of the available data, the known linkages between documents within the edition, as well as the generated metadata and extracted text [1]. The goal of this project is to produce a stable, well-described, and reusable resource that grounds my computational analysis of the culture of early Seventh-day Adventism. Additionally, the project makes possible the dissemination of a dataset that can be used by others interested in print culture during the nineteenth-century in a way that foregrounds the limitations of doing computational analysis with historical sources.

2. Print in the History of Seventh-day Adventism

While most nineteenth-century denominations utilized print to connect their members, as a medium for evangelism, and as a forum for articulating theological differences, periodical literature played a particularly important role in the development of early Seventh-day Adventism [2].[2] It was to periodicals that Ellen and James White first turned to unite the Adventists in the years after Miller's failed prediction of the second coming on October 21, 1844 [3]. Ellen White defended the use of

HistoInformatics 2021 -- 6th International Workshop on Computational History, September 30, online.
EMAIL: jewieringa@ua.edu
ORCID: 0000-0002-9364-2808

HistoInformatics 2021 -- 6th International Workshop on Computational History, September 30, 2021, online.

[2] Historian Catherine Brekus notes that while "earlier evangelicals had realized the potential of the press, the Millerites published tracts, memoirs, and newspapers on a scale never before imagined."

periodicals in 1850, reporting that she was shown in a vision that God was seeking to bring his people together around the truth of the seventh-day message, an effort that required "that the truth be published in a paper, as preached" [4].[3] The numerous publications of the denomination functioned as the force that brought into being the movement's "imagined community" [5].

While the content of the periodicals provides a window onto the denomination's developing theology and culture, the denomination has also invested in sharing their teachings as widely as possible through the large-scale digitization of their historical materials. This commitment includes the official editing and publication work of the Ellen G. White Estate, which has focused primarily on disseminating White's writings, lay efforts to digitize the denomination's periodicals, and current work to aggregate the digital resource of the denomination in the Adventist Digital Library [6]. As a result of these efforts, a large collection of the historical materials of the denomination are openly available in digital formats, making computational analysis of this material feasible.

3. Challenges of Historical OCR

For scholars working with digital sources, particularly for computational analysis such as text mining and natural language processing, there is much to be gained by attending to the quality of the data and, where possible, improving that quality. The strength of computational algorithms is that they perform consistently and logically upon whatever data they are given. The weakness is that, unlike a human reader who will generally infer that the "IN S TRUCT OR" they encounter in a periodical entitled "*Youth's Instructor*" most likely should be read as "INSTRUCTOR," the computer, unless trained to do so, will make no such inference. The data given to it will be processed literally, in this case as four distinct words. As a result, the quality of the output depends directly on the quality of the initial data. If the data is riddled with errors, then the models created will reflect those errors, and frequently exacerbate them. As the adage goes, "garbage in, garbage out."

Because of the significant effect transcription errors can have on the ability to search, classify, and analyze texts, scholars in computer science and information retrieval, as well as across the digital humanities, have developed several strategies for identifying and correcting regularly occurring errors. In addition to developing a series of corrections to received textual data, current work in information retrieval is focused on using probability and machine learning to estimate the most likely correct substitution for errors using general language patterns [7] [8] [9]. The work of computationally correcting the errors generated during character recognition is an important step in developing a source base that is reliable for tasks from information retrieval to computational analysis and is particularly important when working with the peculiarities of historical documents.

The challenges of working with text generated through OCR compound as one moves backward through time, as the documents become further removed from the materials used to develop and train recognition software and the standardization of everything from font and layout to spelling decreases. Differences in typography, in layout conventions, and lack of standardized spelling all contribute to the errors that OCR engines are likely to make when processing historic documents. In addition, physical blemishes, such as tears, stains, and discoloration in the case of the material objects, low image quality, and previous scanning errors introduce additional challenges. It is this textual layer, particularly when it is created using OCR software, its use in digital corpora, and its limitations and their potential remediation that both enables and constrains the use of computational methods with historic documents.

Within historical scholarship, the limitations of the current digital materials are often first encountered within the context of search and retrieval, as these are modes of interaction that define most scholars' interactions with the digital and digitized ecosystem. While the promise of the online database is that millions of texts are merely a keyword search away, how well that promise is being fulfilled is not easy to ascertain. Studies of the material included in the Eighteenth Century Collections Online (ECCO) collection and within the Burney Collection, which is a Gale product containing 17th and 18th century English newspapers, raise significant concerns about the representativeness of the documents included in these prominent subscription databases and the quality of the available data [10]

[3] In subsequent tellings of this history, a more direct link is told between Ellen's vision and the start of *The Present Truth*. See Floyd Greenleaf and Jerry Moon, "Builder," in *Ellen Harmon White: American Prophet*, ed. Terrie Dopp Aamodt, Gary Land, and Ronald L. Numbers (New York: Oxford University Press, 2014), pp. 126-7 & 140 n.2 and n.6.

[11]. For historians working from digital materials, regardless of method, tackling questions of data transparency and accuracy is of increasing importance as part of documenting the process of historical research.

These data challenges are not limited to large subscription sources, as documented by the *Mapping Texts* project from the University of North Texas and the Bill Lane Center for the American West at Stanford University. Originally conceived with the goal of developing tools and strategies for identifying meaningful patterns within the millions of newspaper pages of the *Chronicling America* project, the teams quickly realized that their ability to ask research questions of the data was dependent upon understanding which questions "*could* be answered with the available digital datasets …" [12]. To address this problem, the *Mapping Texts* team created an interface that highlights the coverage of the newspaper collection, particularly in terms of geography and dates, as well as the quality of the data. The interface shows an overall low "good word" rate across the Texas newspapers in *Chronicling America*, with most of the best performing areas hovering around 80% while worse performing areas hover around 60% of the identified words being recognized. This data encourages readers to temper their expectations about what can be known from the analyzed data, a useful counterbalance to the seeming expansiveness and authority of digital materials.

The situation encountered by the *Mapping Texts* team is not unique to the *Chronicling America* data. A quick browse through the "text view" of content in HathiTrust and Google Books reveals that it does not take very many OCR errors to make the text illegible. This is also true for other data sources, such as the online archives of the Seventh-day Adventist church. In the case of library databases, it is harder to ascertain the quality of the data because the venders tightly control their content. However, unless the vendors have devoted massive resources to the correction of their textual data, it is safe to assume that their data suffers similar error rates. To date, the data derived from scans of historical documents is known to be error-ridden, with few suggestions on how historians might better manage the situation of scarcity of reliable documents, despite an abundance of digitized materials.

4. Toward a "Scholarly Edition" of Digitized Texts

Current developments in computational textual analysis represent words in relation to their surrounding words and their grammatical functions. This form of representation is powerful, resulting in significant improvements in computer translation and enabling researchers to explore the relationships between words in new ways [13] [14]. The cutting edge of computational text analysis is to be found in analysis that considers language as networked, contextual, and relational.

However, without careful attention to the boundaries of a digital collection and the quality of the textual data, it is unclear whether the textual data from digitization efforts can support this sort of analysis. Careful analysis of the data quality from scanned historical sources is still rare among digital humanities projects, with the notable exception being studies into OCR quality for the purpose of improving information retrieval for large scale digitized collections. This absence is both curious and not unexpected. On the one hand, the quality of the data is critically linked to the potential success of experiments using machine learning algorithms with historical data, a connection that it seems should weigh heavily on researchers' minds. However, the work involved in diagnosing and addressing errors in character recognition is not trivial and has few rewards within the current academic structure. As a result, researchers interested in the intersection of computational text analysis and historical sources have tended to pursue one of two tracks: experiment with high quality data sets, such as in the cases of Martha Ballards Diary or the Mining the Richmond Dispatch, or to pursue modes of analysis considered more resilient to data errors, such as the text reuse work of Viral Texts and America's Public Bible [15].

Such work has opened the conversation about the use of computational techniques in historical research and provides the necessary first steps in the effort to bring these two methodological approaches together. However, for computational methods for history to continue to mature as a form of analysis, the limitations of the available data need to become a research question to be addressed head-on. With a clear understanding of the data and the types of analysis it reliably supports, the academic community can begin to improve the available data and develop algorithms designed to support the complex, interpretive forms of analysis of the humanities.

In her recent work, *A World of Fiction*, Katherine Bode proposed the format of the scholarly edition of a literary system as a response to the challenges facing digital historians looking to work with large corpora of textual data of unknown range and data quality. While large collections of digitized text have made possible several groundbreaking studies, there is yet a robust and agreed upon method for managing textual data, from identifying the relationship between the digitized texts and the literary systems in which they participated, identifying and addressing shortcomings in the digitization, and redistributing a dataset that is well documented and appropriate for assessment and additional research. Bode describes her proposed scholarly editions of literary systems as providing a "critical apparatus" that "elaborates the complex relationships between the historical context explored, the disciplinary infrastructure employed in investigating that context, the decisions and selections implicated in creating and remediating the collection or collections, and the transformations wrought by the editor's extraction, construction, and analysis of that data." For Bode, this results in a "curated dataset" that provides the critical apparatus necessary for developing reliable computational models [1, p. 53].

The publications of the Seventh-day Adventist church provide a rich example of a data source that is well suited for treatment as a digital scholarly edition in the tradition of Bode. There already exists collections of digitized materials, made available through the denominational archives and digital repository. While much smaller than the collections gathered in large repositories such as HathiTrust, Trove, and the Library of Congress, this collection, all produced by and related to Seventh-day Adventism, is itself too large for an individual researcher to easily ascertain the connection between the digital materials and the historical materials they represent and contexts from which they derive. Looking at only four regions within the United States between 1849 and 1920, I have to date identified over 13,000 periodical issues from 28 periodical titles published and digitized by the denomination, a figure that captures approximately 63% of the documented denominational titles produced in those areas during that period [16].[4] Additionally, the denomination produced an as yet unknown number of tracts, books, charts, hymnals, cookbooks, and additional miscellaneous print materials that together with their periodicals formed a crucial piece of infrastructure on which the religious movement formed.

A scholarly edition of this literary system would provide a mechanism for documenting the historical context and conditions of production for various documents, the data (both textual and contextual) extracted from them and its quality, and connections between the documents, from shared authors and editors to re-publication and revision patterns within and across titles over time. This contextual information would provide the basis for developing robust computational models of the development of the denomination, but also for studying the role of print technologies in its formation. As a curated, annotated, and stable collection of Seventh-day Adventist print literature, the scholarly edition will provide a mechanism for data publication and reuse by scholars interested in Seventh-day Adventism. The scholarly edition also becomes a resource for larger projects on cultural formation in the nineteenth century or other investigations into nineteenth-century print culture, religious or otherwise. And finally, it provides a clear scholarly output for the intellectual work of data curation and preparation for computational scholarship that to date has no generally recognized format and so is often insufficiently invested in and underexamined.

5. Conclusion

The format of the scholarly edition provides a structure for creating a scholarly object that does the work of identifying the boundaries, both in scope and quality, of the data to be used in computational research. It provides a familiar output for the scholarly work of identifying, aggregating, and assessing sources, while also making the resulting collection useful to other researchers because the parameters of the data are known and described. Encouraging the development of this genre of scholarly production provides an alternative to research based on patterns found in documents selected through search in large databases or with large collections of text where the quality and scope of the data in relation to the whole is unclear. Scholarly editions of literary systems provide a format for publication and reuse of the data, together with the necessary contextual framing, of particular collections of digitized

[4] For my discussion of sources, see http://dissertation.jeriwieringa.com/essays/chapter-2/#developing-a-corpus-for-a-gospel-of-health-and-salvation.

historical materials. In developing scholarly editions of textual worlds, we make possible a deeper interweaving of context with computational text analysis in historical research.

6. Acknowledgements

This research is an outgrowth of my dissertation, defended in 2019 at George Mason University. Early versions received funding from Summer Research Fellowships and a Completion Grant through the University. My committee members have been most supportive of my research, and I owe a particular thanks to Michael O'Malley, Sharon Leon, and Fred Gibbs. Thanks to my students in REL 370 at the University of Alabama who worked with me through Bode's book. Thanks also to the anonymous reviewers for this piece whose encouraging and insightful comments greatly improved this essay. And, as always, thanks to Celeste Sharpe and Erin Bush for their generous feedback on these ideas as they have developed.

7. References

[1] K. Bode, A World of Fiction: Digital Collections and the Future of Literary History., Ann Arbor: University of Michigan Press, 2018.

[2] C. A. Brekus, Strangers and Pilgrims: Female Preaching in America, 1740-1845 (Gender and American Culture), Chapel Hill : The University of North Carolina Press, 1998.

[3] J. White, "Dear Brethren and Sisters —," *The Present Truth, Vol 1, No. 1,* p. 6, 1849.

[4] E. G. H. White, "Dear Brethren and Sisters —," *The Present Truth 1, no. 11,* pp. 86-87, 1850.

[5] B. Anderson, Imagined Communities: Reflections on the Origin and Spread of Nationalism, Revised Edition, New York: Verso, 2006.

[6] E. Koester, *Inquiry Regarding Data from the Adventist Digital Library.,* 2017.

[7] J. a. K. F. Evershed, "Correcting Noisy Ocr: Context Beats Confusion," *DATeCH '14: Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage,* pp. 45-51, 2014.

[8] T. a. S. H. Lasko, "Approximate string matching algorithms for limited-vocabulary OCR output correction," *SPIE Proceedings Vol. 4307,* 2000.

[9] T. Underwood, "The Challenges of Digital Work on Early-19c Collections.," 2011. [Online]. Available: https://tedunderwood.com/2011/10/07/the-challenges-of-digital-work-on-early-19c-collections/.

[10] P. Spedding, "'The New Machine': Discovering the Limits of ECCO," *Eighteenth-Century Studies 44, no. 4,* p. 437–53, 2011.

[11] T. Hitchcock, "Confronting the Digital: Or How Academic History Writing Lost the Plot," *Cultural and Social History 10, no. 1,* p. 9–23, 2015.

[12] A. J. e. Torget, "Mapping Texts: Combining Text-Mining and Geo-Visualization to Unlock the Research Potential of Historical Newspapers," 2011. [Online]. Available: http://mappingtexts.org/whitepaper/MappingTexts_WhitePaper.pdf.

[13] B. M. Schmidt, "Interactive Visual Bibliography: Describing Corpora," 2018. [Online]. Available: http://creatingdata.us/techne/bibliographies/.

[14] Y. e. Wu, "Google's Neural Machine Translation System: Bridging the Gap Between Human and Machine Translation," 2016. [Online]. Available: https://arxiv.org/abs/1609.08144.

[15] D. A. a. R. C. Smith, "A Research Agenda for Historical and Multilingual Optical Character Recognition"," 2018. [Online]. Available: https://ocr.northeastern.edu/report/..

[16] J. E. Wieringa, "Constructing Computational Models from Historical Texts: A Consideration of Methods," 2019. [Online]. Available: http://dissertation.jeriwieringa.com/essays/chapter-2/.