The Making of *America's Public Bible*: Computational Text Analysis for Religious History

Lincoln A. Mullen
Department of History and Art History
George Mason University
lincoln@lincolnmullen.com

This is a preprint of a chapter to appear as “The Making of America’s Public Bible: Computational Text Analysis in Religious History,” in *Introduction to Digital Humanities: Research Methods for the Study of Religion*, edited by Christopher D. Cantwell and Kristian Petersen (DeGruyter, under review). This version incorporates revisions suggested by the editors but has been posted prior to peer review. This preprint has been posted with the permission of the editors. Please cite the forthcoming version.

The Making of *America's Public Bible*: Computational Text Analysis for Religious History

Lincoln A. Mullen
Department of History and Art History
George Mason University
lincoln@lincolnmullen.com

America's Public Bible is a website that charts biblical quotations in U.S. newspapers over the nineteenth and early twentieth centuries.¹ The prototype version uses the *Chronicling America* corpus of over 12 million newspaper pages as a source base. It finds and identifies quotations from the King James Version of the Bible. The prototype was created for the *Chronicling America* Data Challenge hosted by the National Endowment for the Humanities.² The initial version, which will remain available until the revised version is published, is available at the project website:

<<http://americaspublicbible.org>> (figure 1). A much expanded version—featuring an additional newspaper corpus, more versions of the Bible and other scriptures, and, most importantly, expanded interpretation in the form of visualizations and prose—will be published by Stanford University Press as a part of their Andrew W. Mellon Foundation funded digital publishing program.³

¹ Lincoln A. Mullen, *America's Public Bible* (Stanford University Press, forthcoming): <<http://americaspublicbible.org>>.

² "NEH Announces the Winners of the Chronicling America Data Challenge," National Endowment for the Humanities, 27 July 2016, <https://www.neh.gov/news/press-release/2016-07-25>.

³ Stanford University Press's digital publishing program: <https://www.sup.org/digital/>.

America's Public Bible

Biblical Quotations in U.S. Newspapers

For most of its issues in 1902, the *Ellensburg [Washington] Dawn* featured a quotation from Benjamin Franklin prominently on its front page. "A Bible and a newspaper in every house," the masthead proclaimed, "are the principal support of virtue, morality, and civil liberty."¹ Though the quotation from Franklin was doubtless spurious, the combination of newspapers and the Bible would have been familiar to readers. Throughout the nineteenth century and into the twentieth century, newspapers in the United States—even newspapers which were not published by a religious denomination or organization—had frequent recourse to the Bible. Newspapers printed sermons and Sunday school lessons, and ministers offered lessons through newspaper Bible clubs.² Newspapers featured jokes whose punchlines required familiarity with the Bible. They aired political commentary that cited the Bible on all sides of a given issue. They ran features on [Thomas Jefferson's edited Bible](#) and [Abraham Lincoln's use of the Scripture](#).³ On [Good Friday](#), [Easter](#), and [Christmas](#) they reprinted long portions of the Scripture.⁴ They [opined on revisions to the English Bible](#), and offered [word-by-word comparisons](#) of the changes in new translations.⁵ They made money from [advertisements](#) for Bibles of every kind, and some newspapers even [sold Bibles directly](#) as a way of raising revenue. But most of all, newspapers quoted the Bible.

America's Public Bible uncovers the presence of biblical quotations in the nearly 11 million newspaper pages in the Library of Congress's [Chronicling America](#) collection. Using the [techniques of machine learning](#) I have identified over 866,000 quotations of the Bible or verbal allusions to specific biblical verses on those newspaper pages. For now, the project has looked only for quotations to the [King James Version](#) (or Authorized Version) of the English Bible, by far the most commonly used Bible among American Protestants during

An advertisement for *The Illuminated Holy Bible*, which several newspapers sold directly. From the *Brattleboro Daily Reformer* (Brattleboro, Vt.), 15 Nov. 1916.

Figure 1. The front page of the prototype version of *America's Public Bible*, created for the *Chronicling America* Data Challenge.

America's Public Bible as a digital history or digital religious studies project could be discussed in a variety of ways. The project uses the techniques of machine learning and text analysis to find the quotations, and it uses data analysis and visualization to make sense of them. It is therefore a kind of computational humanities project, like Frederik Elwert's use of network analysis of ancient Egyptian and Indian religion or Marcus Bingenheimer's digital analysis of Buddhist texts, both described in this volume. Since such projects are rather unusual for humanities scholars (though decidedly less so), the specific computational techniques and methods that they use are understandably an object of curiosity: what did the data look like? what methods were used to analyze it? But it is also possible to talk about the methods behind these projects in a more humanistic way, to understand how computational methods drawn from other disciplines can be used within humanities disciplines. In other words, how does one use computational methods to make a historical interpretation? And then it is also possible to discuss such projects in terms of the contributions they make to a specific field of study. What does this project tell us about nineteenth-century American religious history?

In this essay, I invert the typical way of discussing such projects. Instead of

discussing methods, then modes of interpretations, then results, I begin with a case study of an interpretative result, describe a general pattern of interpretation that is useful to humanities scholars working with digital methods, show how America's Public Bible offers an interface that enables such interpretations, and finally describe the computational methods for finding quotations. Scholars in history and religious studies who are contemplating a digital project must acquaint themselves with the various computational methods available to them, to be sure. But the more pressing problem is finding a mode of argumentation which usefully applies those methods to make meaningful interpretations in one's given field.

Interpretation: The McKinley assassination and 2 Chronicles 7:14

William McKinley, the twenty-fifth president of the United States, was assassinated on September 6, 1901. *The Cameron County Press* reported on the sermons preached in the churches of Emporium to commemorate the slain president. The Methodist, Catholic, Episcopal, and Presbyterian churches all paid tribute to McKinley. The Rev. Robert McCaslin called his Presbyterian congregation to prayer and repentance, juxtaposing prosperity and repentance. "In our national prosperity we were forgetting God and we were becoming self-reliant," the minister claimed. He called his hearers to repent using 2 Chronicles 7:14: "If my people, which are called by my name, shall humble themselves, and pray, and seek my face, and turn from their wicked ways; then will I hear from heaven, and will forgive their sin, and will heal their land."

There could be no mistaking who McCaslin and his hearers thought the "people" in that biblical verse were. "The people" were not ancient Israelites, or even the Presbyterians sitting in the pews of Emporium, but the citizens of the United States. McCaslin drove home this point by encouraging his listeners to heed "the loud call of God to nations to rise and stamp out the curse of anarchy," regretting that it was "deeply humiliating that such a deed could take place in this christian land."⁴

This brief scene should be familiar to any historian of American religion. Christian nationalism has long been a subject of concern to religious historians, and this episode in Emporium is a classic example of the jeremiad.⁵ A public event in the wake of a national tragedy linked Christianity and the state, and the sacred scriptures were

⁴ *Cameron County Press* (Emporium, PA), 19 Sept. 1901, p. 1. All citations to newspapers in this chapter come from *Chronicling America: Historic American Newspapers*, Library of Congress, <http://chroniclingamerica.loc.gov/>.

⁵ Sacvan Bercovitch, *The American Jeremiad* (Madison: University of Wisconsin Press, 1978); Andrew R. Murphy, *Prodigal Nation: Moral Decline and Divine Punishment from New England to 9/11* (New York: Oxford University Press, 2009).

used to undergird the power of the state and condemn its enemies. Like a Puritan fast day, the day was set aside for prayer, preaching, and repentance. And the scriptural verse, 2 Chronicles 7:14, used would later become popular in the late twentieth-century rise of the religious right.

The question, though, is whether this use of the Bible was typical or unusual. When the citizens of Emporium opened their Bibles to 2 Chronicles, would the verse calling them to humility have naturally seemed to mean what their ministers said it meant that day? Or might it have held other meanings that they saw as more obvious?

In the case of 2 Chronicles 7:14, an answer is readily apparent. In the decade on either side of the McKinley assassination, the connection between that text and Christian nationalism was highly unusual. Instead the verse had two other uses which were far more common.

Far more frequently, the verse was used as a call to humility for Christians, occasionally in the context of a revival. The *Watchman and Southron* in Sumter, SC, ran an unsigned column on "Humility," assembling the verse from 2 Chronicles and other texts, encouraging its readers to "never mind the showy array and costly equipage of the worldly, but to be clothed with humility."⁶ In St. Louis, the pastor of the Presbyterian church preached on the text, reminding his congregants that, as the newspaper headline put it, "to give up habits of sin ... is harder than pulling teeth."⁷ In Sacramento, a Congregationalist pastor surprisingly used the text to recommend the practice of Lent, and less surprisingly to encourage preparation for the famous preacher D. L. Moody's upcoming revival in that city. The pastor did regard the Lenten observance as tied to the responsibility of Christians in the state, observing that "as a nation and people we need many things," among them "higher ideals of business integrity" and "higher standards of political morality," besides the "deeper consciousness of God."⁸

The verse was also popular in revivals. The *Pacific Commercial Advertiser* in Honolulu tried to gather "the Christian people of our city" in a "call to prayer" quoting that text as a means of "putting away of sin."⁹ A cooperative revival in Kentucky of "God's own people," meaning clearly the members of churches and not the citizens of the United States, was announced by the *Mt. Sterling Advocate*.¹⁰ The *Monroe City Democrat* in Missouri specifically addressed the verse to Christians, quoting it as the

⁶ *Watchman and Southron* (Sumter, SC), 19 Dec. 1882, p. 5.

⁷ *St. Louis Republic* (St. Louis, MO), 17 Feb. 1902, p. 6.

⁸ *Record-Union* (Sacramento, CA), 6 Mar. 1899, p. 4.

⁹ *Pacific Commercial Advertiser* (Honolulu, HI), 5 Apr. 1905, p. 2.

¹⁰ *Mt. Sterling Advocate* (Mt. Sterling, KY), 22 March 1905, p. 7.

solution “when a church wants a real revival.”¹¹

The second common use of 2 Chronicles 7:14 was as a prayer for rain in response to drought. After all, verse 13 framed the call to repentance as being in a time when “there is no rain.” A Baptist pastor in Clinton, Missouri, promised that “God will answer prayer with rain.”¹² In Utah the *Deseret Evening News* quoted the verse while worrying about a drought’s effect on the corn crop.¹³ In 1901, the governor of Nebraska proclaimed a day of prayer “for relief from destructive winds and drouth.” Ten years later the *Omaha Daily Bee* ran a commemoration, claiming that by 1pm on the day of prayer it began to rain and that in a few days “the whole state was wet down.” The verse from 2 Chronicles was used to bolster that meteorological claim.¹⁴

The point of undertaking this brief history of 2 Chronicles 7:14 is not to suggest that Christian nationalism was not a significant factor in American life, but to put it in a larger context. That verse did not provoke automatic and easy assumptions of the role of God’s favor on the United States or his promised defense against anarchists and other enemies. For farmers in Utah or Missouri, the danger against which God might defend them was drought. Their link to ancient Israel might not be an assumed identity as the people of God but their common agricultural occupation. For a churchgoer the enemy was pride and hypocrisy, if also sometimes false dealing in business and politics. A pattern of quotation that became a favorite of Jerry Falwell and the Moral Majority in the 1980s had antecedents at the turn of the twentieth century, but those antecedents were unusual rather than typical.

Pattern of argumentation: the typical and the exceptional

This vignette of a civil religious event in the McKinley assassination combined with the more complicated history of a biblical text highlights a fundamental tension in how scholars approach the study of religion and history. The problem lies in knowing whether some phenomenon that we are interested in studying is typical or exceptional. Arguments in the humanities tend to be structured around common patterns. One common pattern explicates some unusual text or event. An equally important strand of research aims at explicating the typical, the everyday, and the ordinary.

But how can we know whether the object of our study is one or the other? Scholars often assert but seldom prove their claims about whether something is

¹¹ *Monroe City Democrat* (Monroe, MO), 22 Mar. 1900, p. 6.

¹² *St. Louis Republic* (St. Louis, MO), 15 July 1901, p. 7.

¹³ *Deseret Evening News* (Salt Lake City, UT), 15 July 1901, p. 4.

¹⁴ *Omaha Daily Bee* (Omaha, NE), 23 July 1911, p. 3.

exceptional or typical. Such claims typically rest on the basis of the scholar's expertise. This is not without justification. After all, a career spent immersed in the sources does give scholars some ability to distinguish between the ordinary and the extraordinary.

Our intuitive understanding, though, is inadequate in the face of the scarcity and abundance of our sources—a problem faced by all humanistic disciplines, but perhaps especially by historians. Historical sources are scarce in that we always have the problem of an incomplete set of sources, and must therefore be attentive to the silences of the archives.¹⁵ The art of being a historian is taking sources that were created for one purpose and reading against their grain to answer the questions which interest us.

Yet historians also face the problem of the abundance of sources. However partial they may be, the archival and the printed record is vast. For all the labors of librarians and archivists, our sources are inadequately cataloged and indexed. For even the most narrowly targeted scholarly question, the sources available outstrip the historian's time and ability to read. There are no two ways about it: the way we go about finding our sources are inextricably tied up with chance and happenstance.

This problem of scarcity and abundance is rendered all the more acute by the rise of digitized sources. Digitized sources, as everyone knows, hardly represent the whole of the human record available in libraries and archives, nor do the collections in libraries and archives represent everything that was created in the past. The easy availability of digitized sources leads historians to use those sources rather than other sources which are available in the archive but remain undigitized.¹⁶ Digital sources thus exacerbate the problem of scarcity, in the sense that they are an incomplete and partial record of the past which absorbs scholarly attention.¹⁷ But digitized sources also exacerbate the problem of abundance. Scholars can access a much larger source base than ever before thanks to collections of primary source materials such as newspapers, photographs, government documents, as well as large book collections including the HathiTrust and Google Books.

We should not have a rosy view of this digitization, not least because of two problems that limit their usability. One is the enclosure of our cultural heritage by the large, for-profit publishers often doing the digitizing. The importance of large scale, publicly funded projects like the *Chronicling America* collection of over 13.4 million

¹⁵ Roy Rosenzweig, "Scarcity or Abundance? Preserving the Past in a Digital Era," *American Historical Review* 108, 3 (2003): 735–762, <https://doi.org/10.1086/ahr/108.3.735>.

¹⁶ For a useful example, see Ian Milligan, "Illusionary Order: Online Databases, Optical Character Recognition, and Canadian History, 1997–2010," *Canadian Historical Review* 94, no. 4 (2013): 540–569.

¹⁷ Michael O'Malley, "Evidence and Scarcity," blog post, 2 October 2010, <http://theaporetic.com/?p=176>; and Sean Takats, "Evidence and Abundance," blog post, 18 October 2010, <http://quintessenceofham.org/2010/10/18/evidence-and-abundance/>.

newspaper pages cannot be understated. These projects are free both in terms of cost and free in terms of the way that scholars can use them for any purpose, including computational research. (They are both *gratis* and *libre*, in the parlance of open source software.) But the norm for digital collections is the subscription database from a for-profit company. Though paid for at great cost by university libraries, such databases are not usually available for text mining.¹⁸ The second limitation of digitized sources is the way that increasing the scale of a source base tends to decrease the diversity of sources used. A thousand, a million, or ten million newspaper pages are all still just newspaper pages. Historians and most humanities scholars tend to rely on combining many different kinds of sources in order to make useful interpretations, but using large-scale text corpora, for instance, paradoxically narrows the kinds of source scholars use.¹⁹

The abundance of digitized sources has already transformed historical research.²⁰ As Lara Putnam has pointed out in an article on “The Transnational and the Text Searchable,” searching digitized collections is now a basic scholarly practice. Putnam argues that the ability to search for sources without the constraints of the national archive has allowed new angles of vision on transnational history, because “transnational approaches among historians did not become commonplace until technology radically reduced the cost of discovering information about people, places, and processes outside the borders of one’s prior knowledge.” Yet as Putnam points out, “Digital search makes possible radically more decontextualized research” and makes it possible to find examples of what we are looking for without a sense of its significance. To deal with this problem of context, Putnam observes that “computational tools can discipline our term-searching if we ask them to. By measuring proximity and comparing frequencies, topic modeling [or other text analysis methods, we might add] can balance easy hits with evidence of other topics more prevalent in those sources.”²¹

In explaining what my digital project is and how it functions, I hope to show two things. First, the project creates serendipitous findings through computational history

¹⁸ Thomas Padilla, “Text and Data Mining: Seeking Traction,” LIS Scholarship Archive, 7 March 2018, <https://osf.io/preprints/lissa/qxs9j>.

¹⁹ For a discussion of this problem, see the Arguing with Digital History working group, “Digital History and Argument,” white paper, Roy Rosenzweig Center for History and New Media (13 November 2017): <https://rrchnm.org/argument-white-paper/>.

²⁰ Jennifer Rutner and Roger Schonfeld, “Supporting the Changing Research Practices of Historians” Ithaka S+R, 11 August 11, 2015, <https://doi.org/10.18665/sr.22532>, p. 9, 14-15.

²¹ Lara Putnam, “The Transnational and the Text-Searchable: Digitized Sources and the Shadows They Cast,” *American Historical Review* 121, no. 2 (2016): 377–402, <https://doi.org/10.1093/ahr/121.2.377>. Quotations at p. 383, 392. See also Tim Hitchcock, “Confronting the Digital: or How Academic History Writing Lost the Plot,” *Cultural and Social History* 10, no. 1 (2013): 9–23; Ted Underwood, “Theorizing Research Practices We Forgot to Theorize Twenty Years Ago,” *Representations* 127, no. 1 (2014): 64–72.

by surfacing sources that would otherwise go unnoticed. And second, the project disciplines those searches by setting the results in a much broader chronological context in which the typical and the exceptional can be identified. This disciplined serendipity constitutes a method of approaching the past, and in this case for approaching a past relevant to religious studies.²²

Interface: disciplined serendipity through interactive visualization

The prototype version of *America's Public Bible* has as its centerpiece a visualization that lets users interactively explore a time series visualization of the trend in quotations for over one thousand of the most quoted verses in the Bible (figure 2).

Figure 2. An example of the interactive verse browser, showing time series of the rate of quotations for the five most quoted verses from Proverbs. Users can enter verse references on their own, or they can choose from pre-selected collections of verses, such as the Lord's Prayer or verses on marriage and divorce.

Most important, users can disaggregate the time series and find each quotation in

²² This idea of disciplined serendipity comes out of conversations with my colleague, Mike O'Malley, who shared an unpublished essay on the topic.

the context of the newspaper page at *Chronicling America*. A table below the visualization shows a row for each instance of a quotation. Users can follow links to that specific newspaper page in *Chronicling America* (figure 3).

Text of selected verses

2 Chronicles 7:14 If my people, which are called by my name, shall humble themselves, and pray, and seek my face, and turn from their wicked ways; then will I hear from heaven, and will forgive their sin, and will heal their land.

Pages with selected verses

Show entries Search:

Newspaper	State	Date	Reference	ChronAm Link
Press And Daily Dakotaian	South Dakota	1887-11-01	2 Chronicles 7:14	See at ChronAm
Record-Union	California	1899-03-06	2 Chronicles 7:14	See at ChronAm
Monroe City Democrat	Missouri	1900-03-22	2 Chronicles 7:14	See at ChronAm
Deseret Evening News	Utah	1901-07-15	2 Chronicles 7:14	See at ChronAm
St. Louis Republic	Missouri	1901-07-15	2 Chronicles 7:14	See at ChronAm
Bossier Banner	Louisiana	1901-07-18	2 Chronicles 7:14	See at ChronAm
Cameron County Press	Pennsylvania	1901-09-19	2 Chronicles 7:14	See at ChronAm
St. Louis Republic	Missouri	1901-10-21	2 Chronicles 7:14	See at ChronAm
St. Louis Republic	Missouri	1902-02-17	2 Chronicles 7:14	See at ChronAm
Spokane Press	Washington	1904-02-25	2 Chronicles 7:14	See at ChronAm

Showing 11 to 20 of 46 entries Previous 1 **2** 3 4 5 Next

Figure 3. After identifying an interesting verse and see the trend in quotations, users can find every quotation which was used to compute that trend line. The links to the right take users to that specific newspaper page in *Chronicling America*.

On arriving at the newspaper page in *Chronicling America*, the key words in the quotation are highlighted on the page. This allows the user to readily identify the quotation on the page, which would otherwise be quite difficult.

The ability to move between the trend line of the verses quotation and its location in the actual primary source is the way that the site enables disciplined serendipity. The serendipity lies in how the site surfaces hundreds of thousands of instances of quotations which the user can readily browse. This may be a subjective judgement, to be sure, but this has been the most fun project that I have ever worked on because I am constantly surprised by the quotation finder (and not just that it works at

all!). For example, how could I have known to look for the time when a Democratic newspaper thought Samuel Tilden had been elected in the disputed presidential race of 1876, and plastered the banner “The Lord called Samuel” across the paper (figure 4). Biblical jokes are another frequent category that I did not expect.²³ I take this as a sign that the method truly is serendipitous.

Figure 4. In the disputed 1876 election for the president of the United States between Rutherford B. Hayes and Samuel Tilden, several Democratic newspapers announced Tilden’s supposed victory using the verse 1 Samuel 3:8 (“The Lord called Samuel”). Stark County Democrat (Canton, OH), 7 December 1876.

The discipline in this approach lies in how the site contextualizes those quotations. Each text is set within two different contexts. The first context is the trend in the verses quotation. This context allows the user to see both how often the verse was quoted in an absolute sense, and to compare that trend to other verses to get a sense of its relative popularity.

The second context is the place of the text on the newspaper page itself (figure 5). This context allows the scholar to understand how the Bible verse was used. The Bible was a common yet contested text, and the fact that a verse was quoted does not show the meaning of that quotation. To add to the example of how 2 Chronicles 7:14 was used in the wake of the McKinley assassination, take the trend for John 15:13 (“Greater love hath no man than this, that a man lay down his life for his friends”). This verse exploded in popularity around World War I, and looking at how the verse was used in specific newspapers confirms that it was popular because of obituaries. Investigating

²³ “Our own inability to get the joke is an indication of the distance that separates us from the workers of preindustrial Europe. ... When you realize that you are not getting something—a joke, a proverb, a ceremony—that is particularly meaningful to the natives, you can see where to grasp a foreign system of meaning in order to unravel it.” Robert Darnton, *The Great Cat Massacre: And Other Episodes in French Cultural History* (Basic Books, 1984), 77–78.

Cameron County press., September 19, 1901, Image 1
About [Cameron County press. \(Emporium, Cameron County, Pa.\) 1866-1922](#)

First, it will broaden the source base and improve the reliability of the method used to find the quotations. This will include extending the source base to Gale's corpus

of *19th Century U.S. Newspapers*.²⁴ This corpus is available only by license from Gale, which in my case was made possible because my university library purchased the text mining rights. Restrictions on how the corpus can be used, in particular a requirement that only the briefest of snippets be reproduced from the text of the newspapers, means that displaying the full text of the quotation and allowing back-and-forth interaction between contexts is simply not possible using the Gale dataset. The Gale corpus is also an order of magnitude smaller than *Chronicling America*. But it has cleaner OCR than *Chronicling America*, and it is segmented into articles rather than pages. It is therefore useful for validating the trends identified in *Chronicling America*. Article- rather than document-level segmentation may allow me to use computational analysis to understand the topics of articles in which the verses were quoted.

Second, the new version will extend the finding of quotations to other versions of the Bible. The prototype version used the King James version since it was a standard text for English-speaking Protestants in the United States. The revised version will use other versions of the Bible as well. It is an extremely difficult problem both identify quotations *and* distinguish between versions of the Bible which have only small (if significant) verbal differences. Nevertheless this is an important addition to the project, since the rise of different translations of the Bible was a crucial moment for American Protestants, and since Catholics and Jews (and non-English speakers, for that matter) have all used versions other than the KJV.²⁵

Third, the site will take the exploratory analysis of the prototype and add interpretative work on the history of the Bible. In digital history there is a widespread sense that digital work has failed to make historical interpretations.²⁶ This digital monograph will feature a series of chapters that interpret various aspects of the history of the Bible, using prose and visualizations. This interpretative process will necessarily involve more data analysis of the quotations, such as clustering the time series of quotation trends in order to find verses that had similar trajectories over time. But at its root will be much close reading of the context of the verses. The project will thus rely on the serendipity of the quotations that are turned up but also on the discipline of its empirical approach.

Method: text analysis and machine learning for quotation identification

²⁴ Gale, *19th Century U.S. Newspapers*, <https://www.gale.com/c/19th-century-us-newspapers>

²⁵ Peter J. Thuesen, *In Discordance with the Scriptures: American Protestant Battles Over Translating the Bible* (Oxford University Press, 1999).

²⁶ Cameron Blevins, "Digital History's Perpetual Future Tense," in *Debates in the Digital Humanities 2016*, ed. Matt Gold and Lauren Klein (University of Minnesota Press, 2016), <http://dhdebates.gc.cuny.edu/debates/text/77>; "Digital History and Argument" white paper.

Having described the public facing component of the project, and considered the disciplined serendipity it provides, I want to go behind the scenes of the project to show how it was created. In particular, I want to show how sources, methods, and questions came together to shape the project. The availability of large digital corpora and the knowledge of digital text analysis methods are not sufficient to create a meaningful intellectual project. Nor does having a worthwhile intellectual question in one's discipline imply that that one can apply the right methods to the right sources to answer that question in some meaningful way. Sources, methods, questions—these three must come together to make a worthwhile interpretative project. To the extent that *America's Public Bible* is as useful as a model—let the reader be the judge—I want to show how that project combined those three components.

The initial problem was formulating a worthwhile historical question that was amenable to some kind of computational research. When the National Endowment for the Humanities opened a data challenge for the best project using *Chronicling America*, I had an incentive to compete but not a question worth asking. The sources were chosen, but they were not obviously “religious” and so there was no clear connection to my primary research area. I was familiar with an array of methods that might be applied to the dataset. (Topic modelling! the newly initiated to DH is likely to suggest.) But without a question to answer there was no purchase for the methods.

My question came after hearing a paper that Mark Noll gave at the 2016 annual meeting of the American Society of Church History. The paper compared how the Bible was used in sermons after the deaths of Presidents Washington, Lincoln, and Garfield.²⁷ Noll had counted scriptural references in those sermons, then analyzed the way the quotations were used. That paper, along with the work he had done in his books *In the Beginning Was the Word* and *The Civil War as a Theological Crisis*, established that the Bible could be studied as a text that was been used in the public sphere.²⁸ Other historians have of course studied the Bible from a number of angles, not least in studying its reception by Christians and its use as a material object.²⁹ The intellectual

²⁷ Mark A. Noll, “Presidential Death and the Bible: 1799, 1865, 1881” (American Society of Church History, Atlanta, 2016). The latter two presidents died from an assassin's bullet, so consider the opening of this essay on the McKinley assassination an homage to the conference paper which gave rise to this project.

²⁸ Mark A. Noll, *In the Beginning Was the Word: The Bible in American Public Life, 1492–1783* (Oxford University Press, 2016); Mark A. Noll, *The Civil War as a Theological Crisis* (University of North Carolina Press, 2006), ch. 3.

²⁹ James P. Byrd, *Sacred Scripture, Sacred War: The Bible and the American Revolution* (Oxford University Press, 2013); Valerie C. Cooper, *Word, Like Fire: Maria Stewart, the Bible, and the Rights of African Americans* (University of Virginia Press, 2011); Philip Goff, Arthur E Farnsley II, and Peter J. Thuesen,

frame that Noll was using, however, was one that both justified the source base and pointed to the methods that should be used. Why was a large corpus of newspapers worth investigating for a project in the history of the Bible in America? Because it could provide evidence of how the Bible was used in the public sphere. What kind of method would be worth using? Finding quotations was a task that could be turned over to the computer.

America's Public Bible is an example of a methodological approach that I would label computational history. The project uses the techniques of machine learning to find examples of biblical quotations, then uses data analysis combined with close reading to understand the meaning of those quotations. Finding the quotations computationally required a series of steps.³⁰

The first step was to download the plain text versions of the roughly 12 million newspaper pages available at the time. While for many online resources this would be an enormous chore or impossible due to licensing restrictions, the data for *Chronicling America* is easily downloaded thanks to the Library of Congress's bulk data downloads. *Chronicling America* also offers JSON and RDF application programming interfaces (APIs) for downloading the machine-readable metadata for each newspaper, such as its title, dates of publication, and location.³¹ I parsed this API into a database of newspaper pages and publication information. From there I also computed other basic metadata such as the word count for each page.

Once the data was in hand (or, more precisely, on a RAID array) it was time to apply a series of digital methods. The second step was to identify potential quotations on the page. Detecting quotations can be thought of as a problem of detecting text reuse.³² While this is a well-known area of research, for my particular problem it was

eds., *The Bible in American Life* (Oxford University Press, 2017); Paul C. Gutjahr, *An American Bible: A History of the Good Book in the United States, 1777–1880* (Stanford University Press, 1999); Nathan O. Hatch and Mark A. Noll, eds., *The Bible in America: Essays in Cultural History* (Oxford University Press, 1982); Colleen McDannell, *Material Christianity: Religion and Popular Culture in America* (Yale University Press, 1995); Seth Perry, "Scripture, Time, and Authority among Early Disciples of Christ," *Church History* 85, no. 4 (December 2016): 762–83, <https://doi.org/10.1017/S0009640716000780>; Jonathan D. Sarna and Nahum M. Sarna, "Jewish Bible Scholarship and Translations in the United States," in *The Bible and Bibles in America*, ed. Ernest S. Frerichs (Scholars Press, 1988), 83–116; Stephen J. Stein, "America's Bibles: Canon, Commentary, and Community," *Church History* 64, no. 2 (June 1, 1995): 169–84, <https://doi.org/10.2307/3167903>; Peter J. Wosh, *Spreading the Word: The Bible Business in Nineteenth-Century America* (Cornell University Press, 1994); John Fea, *The Bible Cause: A History of the American Bible Society* (Oxford University Press, 2016); Paul Gutjahr, ed., *The Oxford Handbook of the Bible in America* (Oxford University Press, 2017).

³⁰ Users who are interested in how the project was created are welcome to investigate the code for themselves: <https://github.com/public-bible>. The prototype project has a methodological appendix: <http://americaspublishbible.org/methods.html>.

³¹ *Chronicling America's* bulk data and API: <https://chroniclingamerica.loc.gov/about/api/>

³² For a useful overview of the problem, see David A. Smith, Ryan Cordell, and Abby Mullen,

necessary to invent a new kind of method.

The idea behind my method is straightforward. A common process in text analysis is to tokenize the text, that is, to turn it into single words or phrases (called n-grams).³³ The corpus can then be represented in a matrix, often called a document-term matrix, where the rows are documents (in this case a newspaper page), the columns are tokens, and the counts in the cells are the number of times that token appears in that document. I created a document-term matrix for both newspapers and for the Bible. As long as one limits the tokens in the matrix to only those words and phrases which appear in the Bible, it is possible to multiply one matrix by the transpose of the other matrix. The result is a document-document matrix with Bible verses down the rows, newspaper pages across the columns, and the count of the number of biblical words and phrases in the cells. If all that talk of tokenizing and matrix multiplication sounds confusing, just remember that at the root it is all just counting: the new matrix has the counts of how many words and phrases from each Bible verse appeared on each newspaper page. That is a common sense understanding of what a quotation looks like.

In practice it is not always possible to distinguish between actual quotations and false positives using only the counts of tokens. So I created other features that could indicate a quotation. I weighted the matrices by their TF/IDF score.³⁴ This technique gives more weight to phrases which are unusual and characterize a document. For example, the four-word phrase “went into the city” appears several times in the Bible and many times in newspapers, but the mere fact that phrase appears on a page is unlikely to indicate that it is a quotation from the Bible. But the four-word phrase “through a glass darkly” cannot be anything other than a biblical allusion to 1 Corinthians 13:13. Another useful feature is the percentage of the Bible verse that is quoted. If every word or phrase from a verse appears on a page, then it is more likely to be a genuine quotation than if a single phrase appears. Then too, a bunch of random phrases from the Bible scattered around the newspaper page are unlikely to be a quotation, but if those phrases are concentrated in a single location, then they are far likelier to be a quotation. The spread of phrases around the page can be computed with a statistical test of randomness called a runs test.

Applying this algorithm to the newspapers resulted in millions of possible

“Computational Methods for Uncovering Reprinted Texts in Antebellum Newspapers,” *American Literary History* 27, no. 3 (2015): 1–15, <https://doi.org/10.1093/alh/ajv029>.

³³ For an overview of the text analysis process generally, see Kasper Welbers, Wouter Van Atteveldt, and Kenneth Benoit, “Text Analysis in R,” *Communication Methods and Measures* 11, no. 4 (October 2, 2017): 245–65, <https://doi.org/10.1080/19312458.2017.1387238>.

³⁴ Jure Leskovec, Anand Rajaraman, and Jeff Ullman, *Mining of Massive Datasets*, 2nd ed. (Cambridge University Press, 2014), section 1.3.1, <http://www.mmids.org/>.

quotations from the Bible. But many of these are false positives, as could be verified by looking up the newspaper page and seeing if the biblical text actually appeared. What was necessary was a way of weeding through all of these potential matches and identifying the correct quotations and discarding the incorrect ones. The computational tool for the job is called machine learning, or more precisely, supervised classification.

The aim of supervised classification is to take inputs (potential quotations) which have certain features (the count of tokens, how unusual those tokens were, etc.) and assign them a label (quotation or not-a-quotation). Typically when writing a program one takes input data and creates a series of rules about what to do with it. To make up some rules, for instance, one might instruct the computer that matches with more than five tokens are genuine quotations and the rest are noise. However such rules are extremely difficult to write. Ten meaningless tokens might not be a quotation, but one unusual token could be a quotation.

Machine learning takes a different approach. Instead of writing rules and getting answers, you give the computer a set of answers and get back the rules, called a model. In my case, I went through a number of potential matches and labeled them as quotations or not (see figure 6). The idea behind supervised classification is that by showing a machine-learning model what a bunch of real quotations and a bunch of false quotations look like, the model can learn to tell the difference between them.³⁵ The accuracy of this model can be estimated to test its precision (what proportion of its results are true matches) and recall (the proportion of all matches that it has found). After creating the training data and evaluating the model, I ran the model on the millions of possible quotations using my university's high-performance computing cluster.

³⁵ This description of machine learning is indebted to Francois Chollet and J.J. Allaire, *Deep Learning with R* (Manning, 2018), section 1.1.2, especially figure 1.2.

	A	B	D	G	H	I
1	match	reference	url	token_count	proportion	tfidf
2	FALSE	Luke 10:31 (KJV)	http://chroniclingamerica.loc.gov/lccn/sn83040198/1897-12-23/ed-1/seq-3/#	1	0.02631578947	0.1992378928
3	TRUE	Romans 3:23 (KJV)	http://chroniclingamerica.loc.gov/lccn/sn85034235/1883-01-25/ed-1/seq-1/#	5	0.2941176471	2.763034851
4	FALSE	Judges 19:1 (KJV)	http://chroniclingamerica.loc.gov/lccn/sn88076523/1911-08-18/ed-1/seq-2/#	2	0.01612903226	0.2649320585
5	FALSE	Proverbs 10:28 (KJV)	http://chroniclingamerica.loc.gov/lccn/sn84026853/1886-04-28/ed-1/seq-1/#	3	0.125	1.170029201
6	FALSE	Galatians 3:15 (KJV)	http://chroniclingamerica.loc.gov/lccn/sn83025121/1897-03-02/ed-1/seq-2/#	1	0.02222222222	0.1671054807
7	FALSE	1 Maccabees 13:7 (KJV)	http://chroniclingamerica.loc.gov/lccn/sn82014760/1862-03-19/ed-1/seq-1/#	2	0.05882352941	0.9353564175
8	TRUE	John 7:17 (KJV)	http://chroniclingamerica.loc.gov/lccn/sn84026688/1903-02-05/ed-1/seq-1/#	9	0.2195121951	2.129326177
9	TRUE	Ecclesiastes 1:2 (KJV)	http://chroniclingamerica.loc.gov/lccn/sn86083274/1889-12-13/ed-1/seq-2/#	1	0.05882352941	0.5777842192
10	TRUE	Matthew 6:13 (KJV)	http://chroniclingamerica.loc.gov/lccn/sn86072143/1876-12-01/ed-1/seq-1/#	30	0.6511627907	6.657510873
11	FALSE	Acts 10:28 (KJV)	http://chroniclingamerica.loc.gov/lccn/sn82015387/1898-12-17/ed-1/seq-4/#	1	0.01234567901	0.1162576126
12	FALSE	Isaiah 28:15 (KJV)	http://chroniclingamerica.loc.gov/lccn/sn84020104/1853-12-16/ed-1/seq-4/#	1	0.002409638554	0.1183413461
13	FALSE	Ecclesiasticus 17:5 (KJV)	http://chroniclingamerica.loc.gov/lccn/sn86072143/1872-03-15/ed-1/seq-3/#	1	0.006172839506	0.1615503599
14	FALSE	Tobit 7:15 (KJV)	http://chroniclingamerica.loc.gov/lccn/sn82016419/1877-07-27/ed-1/seq-1/#	1	0.3333333333	3.274110575
15	FALSE	1 Kings 6:30 (KJV)	http://chroniclingamerica.loc.gov/lccn/sn97071028/1888-06-08/ed-1/seq-1/#	1	0.05263157895	0.4804833971
16	FALSE	Baruch 2:26 (KJV)	http://chroniclingamerica.loc.gov/lccn/sn88078726/1854-06-15/ed-1/seq-1/#	1	0.008928571429	0.1681583325
17	FALSE	Luke 1:63 (KJV)	http://chroniclingamerica.loc.gov/lccn/sn83016942/1838-07-21/ed-1/seq-3/#	3	0.05172413793	1.016103282
18	FALSE	1 Samuel 17:3 (KJV)	http://chroniclingamerica.loc.gov/lccn/sn83016943/1910-06-24/ed-1/seq-5/#	3	0.06382978723	0.4916472619
19	FALSE	Ecclesiasticus 19:15 (KJV)	http://chroniclingamerica.loc.gov/lccn/sn85025620/1888-03-01/ed-1/seq-2/#	1	0.04347826087	0.4270579011
20	FALSE	Galatians 4:9 (KJV)	http://chroniclingamerica.loc.gov/lccn/sn84024716/1908-06-03/ed-1/seq-3/#	1	0.01754385965	0.1723216092
21	FALSE	1 Maccabees 2:21 (KJV)	http://chroniclingamerica.loc.gov/lccn/sn83016810/1905-08-05/ed-1/seq-2/#	3	0.2	1.825836909

Figure 6. Some sample data used to train the classification model. The combination of the “reference” and “url” columns indicate a single potential quotation. I then filled out the “match” column to indicate whether that row was a genuine quotation (“TRUE”) or just noise (“FALSE”).

The result was a table, where each row was an instance of a Bible verse quoted on a specific newspaper page, along with the probability that it was a quotation, according to the model.³⁶ This table was then able to be used for data analysis. The most obvious approach, and the one which is at the center of the prototype site, was to chart time series of the popularity of the verses. Another approach is to look for verses which often appear next to one another, omitting verses which appear in the same passage. And ultimately the research involves lots of reading of the context of the quotations to see how the Bible was actually used.

To return to the initial contention of this chapter, this method of using machine learning to identify biblical quotations is both like and unlike the kinds of keyword searching and, more broadly, the lack of context and evidence of typicality that I was critiquing.

This method is unlike that kind of keyword searching in that it provides a rigorous contextualization of the quotations I’ve found. Unlike a keyword search, which in a corpus of 60+ billion words is guaranteed to turn up at least something, this shows me the relative frequency of occurrence, so that I can see the context of a quotation in terms of the change in quotations over time and in the context of the Bible as a whole.

³⁶ Technically any quotation with a probability above 50% was predicted to be a quotation by the model. But I chose to use only predicted quotations above 90% for the initial studies, since I judged the problem of false matches to be worse than the problem of missing genuine quotations. This cutoff may change for the final version of the project.

But this method is still a kind of search. In particular, the way that machine learning works requires you to feed it a bunch of examples, and the model then finds other things that look like the examples. It is impossible for my model to find things other than biblical quotations. Fundamentally, the model cannot surprise because it only ever returns biblical quotations. It is thus subject to the limitations of the question that I have framed.

Broader applications

The method of quotation detection that I have described is only one example of a kind of computational text analysis applicable to the humanities. Other forms of identifying the reuse of text can show the migration of ideas.³⁷ Another category of textual analysis relies on understanding discursive structures. Topic modeling shows how categories of topics change over time at the document or corpus level.³⁸ Word-embedded models can show the structure of the relationships between words, allowing a kind of multidimensional map of concepts which can also be compared over time.³⁹ In short, there are a range of text analytical methods which can be applied to questions in religious studies.

What does it take to learn these kinds of methods? At a minimum it requires learning some kind of programming language suitable for data analysis such as R or Python. There are plenty of digital humanities approaches where one can work without ever seeing, let alone writing, a single line of code. But any serious text analysis or data analysis requires the ability to write scripts in a programming language. There are now abundant resources targeted at the humanist who wishes to learn these skills.⁴⁰ And in fact learning to program is the least difficult part of what it takes to do computational

³⁷ Ryan Cordell, "Reprinting, Circulation, and the Network Author in Antebellum Newspapers," *American Literary History* 27, no. 3 (September 1, 2015): 417–45, <https://doi.org/10.1093/alh/ajv028>; Kellen Funk and Lincoln A. Mullen, "The Spine of American Law: Digital Text Analysis and U.S. Legal Practice," *American Historical Review* 123, no. 1 (February 1, 2018): 132–64, <https://doi.org/10.1093/ahr/123.1.132>.

³⁸ Andrew Goldstone, Susana Galán, C. Laura Lovin, Andrew Mazzaschi, and Lindsey Whitmore, *An Interactive Topic Model of Signs*, part of *Signs at 40*. <http://signsat40.signsjournal.org/topic-model>.

³⁹ Ben Schmidt, "Word Embeddings for the Digital Humanities," blog post, 25 October 2015, <http://bookworm.benschmidt.org/posts/2015-10-25-Word-Embeddings.html>; Ben Schmidt, "Rejecting the Gender Binary: a Vector-Space Operation," 30 October 2015, <http://bookworm.benschmidt.org/posts/2015-10-30-rejecting-the-gender-binary.html>

⁴⁰ E.g., Taylor Arnold and Lauren Tilton, *Humanities Data in R* (Springer, 2015), <http://link.springer.com/10.1007/978-3-319-20702-5>; Matthew L. Jockers, *Text Analysis with R for Students of Literature* (Springer, 2014), <http://link.springer.com/10.1007/978-3-319-03164-4>.

work in the humanities. More important is gaining a conceptual familiarity with methods such as data analysis or machine learning. And most important of all is the problem how to frame worthwhile disciplinary questions in the humanities and religious studies. Integrating questions, sources, and methods, and that is a problem which is not well solved.⁴¹

And what does this project cost? Both the prototype version and the eventual digital monograph have been created with no budget, apart from a couple hundred dollars I spent out of pocket on web hosting for the first year or so of the project. The lack of direct costs, however, obscures the very real costs behind the project. The *Chronicling America* dataset was provided for free by the NEH and the Library of Congress, while the George Mason University Libraries entered into an agreement with Gale to allow text mining of their newspaper collection. I have access to the high-performance computing cluster at George Mason, along with web servers at the Roy Rosenzweig Center for History and New Media to host the site. And it is not to be taken for granted that my department affords me the time to undertake research and has the intellectual flexibility to regard such experimental digital projects as scholarship.

But in this chapter I have set these details about writing code to do text analysis and within the large conceptual problems in computational research. I have described the process of creating a large-scale text analysis project. This project is a kind of a search, and therefore subject to the limitation that it depends on the question asked. But on the other hand the project is generative of new knowledge in that it offers a disciplined serendipity in searching, rigorously contextualizing its results but continually turning up surprises. Whatever methods religious studies scholars employ, to the extent that they try to cope with large data sets and text analysis, they may find the problem of distinguishing between the typical and the exceptional a useful starting question, and the kind of disciplined serendipity that I have described a useful method.

⁴¹ For that reason my own contribution to the literature on learning computational methods, which is still very much at the beginning stages, focuses primarily on the patterns of argumentation in computational history. Lincoln A. Mullen, *Computational Historical Thinking: With Applications in R* (2018–): <http://dh-r.lincolnmullen.com>.