This layer provides a description of the methods used to collect the data for the project, the models used to perform text analysis of the life histories, and the underlying technologies used to create the digital project. All of the code and data are available for download under permissive open-source licenses. Links are provided within the text.

**Data Collection**

The collection of life histories used in this project are held by The Wilson Library at the University of North Carolina at Chapel Hill in the Southern Historical Collection.[[1]](#footnote-2) The texts of the life histories have been digitized and made available for public access through the library’s web interface. Life histories are provided as PDF images; there are no searchable, machine-readable versions of the texts available on the website.[[2]](#footnote-3) Each life history has a short caption that includes information about the title of the interview, its location, date, and interviewee's name. This information is not structured into specific fields. Additional metadata information about the life histories are included in the digitized images in the form of headers at the top of each interview and as summary cards included in the archive. We structured the headers into a database for this project and created plain text machine-readable versions of each life history for data analysis.

Creating metadata records of the interviews required manually parsing the unstructured text and reading the individual metadata headers for each interview. The process of manual parsing was conducted by the authors, students in a class taught by Rivard, and a paid research assistant. An article focused on the pedagogical practices and lessons learned from this process appeared in *Digital Humanities Quarterly*.[[3]](#footnote-4) The article discusses various ways of crediting contributors—particularly how to credit those students who made substantive contributions beyond that of the typical class requirements—through authorship credit and other visualizations.

After being converted into structure records, the metadata was organized into a collection of normalized tables.[[4]](#footnote-5) These tables contain information about writers, revisers, interviewees, interviews, and professions. The collection of tables follows the “tidy data” model, with a different table dedicated to each type of record.[[5]](#footnote-6) Normalized relational tables are particularly important in our dataset because many of the relationships linking tables to one another are complex one-to-many and many-to-many relationships. For example, most writers wrote more than one interview, and some interviews were co-written by two or more writers. The normalized data model guarantees that data about each entry is consistent and easy to update.

While most of the variables in the metadata tables are relatively straightforward to record and describe, a few fields require some discussion. Some interviews used pseudonyms for the interviewee names. In many cases, the pseudonyms and real names are mapped to one another in the header of the typed interview; the digital archive hosted by The Wilson Library lists interviews by real names, not pseudonyms. For these two reasons, we have listed both forms of names when given in our dataset and use real names when displaying records on the digital site. This makes our data consistent with the archival source and does not reveal private data that is not available elsewhere.

Recording race and gender information about writers and interviewees also require careful decision-making. As digital humanities scholars such as Catherine D’Ignazio and Lauren Klein have argued, "what gets counted counts," yet must be done with careful attention to binaries, hierarchies, and classifications.[[6]](#footnote-7) Because we were interested in understanding the racial logic that the SLHP produced, we used the categories in the archival records. The SLHP subscribed to a gender binary of “male” or “female”, which they often typed in the header of the life history.

The gender of the interviewees and writers were inconsistently documented though. Since we were interested in the gender representation among the interviewees and writers, we assigned a gender based on the binary logic of the SLHP. Given how writers and editors wrote and edited the stories, the gender can usually be inferred based on pronouns and other gendered language (i.e., “wife”) used in relation to an interviewee. The gender of the writers was determined by archival records through a close reading of correspondences for an individual's pronouns. Given the number of writers who are silent in the archive except for the life history they wrote, we also turned to census data that used the gender binary.

Along with gender, we were interested in how race was configured at the time. We took as our guide cautions about encoding racial logic through data that digital humanities scholars such as Jessica Marie Johnson and DH initiatives such as #transformdh and #dhpoco have elucidated.[[7]](#footnote-8) As constructed and unstable categories, racial categorization was a significant site of contention in the 1930s as groups debated the names and boundaries of race and ethnicity. For interviewees, we used the categories described in the interview metadata. The SLHP, we learned through the process of creating this data, used three primary categories: Negro, White, and Other. Because of the implications of the term “negro” in the past and today, we used “Black” as the category that appears on the site. In an effort to account for the complicated relationship between race and ethnicity, they at times included categories such as Greek or Swedish. These categories should be used carefully, however, because some labels are inconsistently applied. For example, interviews of Greek families sometimes describe the interviewees as "White" and other times as "Greek." We included an additional field to capture more granular ethnic categories that can be inferred from the text.[[8]](#footnote-9) Racial information for the writers was determined by archival records. Records for which we were unable to determine a racial category are labeled as “unclear”.

Our dataset includes a record for every interview that is listed in the archive’s finding aid. The digital files are organized into folders, which most frequently contain a single interview, but occasionally include over a dozen individual interviews. Some interviews are duplicates or near-duplicates of each other, and others consist of a single sentence indicating that the record has been deleted. For consistency and simplicity, our data splits out each individual interview and includes duplicate records. Removal of duplicates and other processing is done during the analysis of the data, making explicit how modeling decisions relate to the data contained in the archive.

Along with the metadata, we also produced machine-readable versions of the text of each interview. Off-the-shelf optical character recognition (OCR) made a reasonable first pass of some of the interviews but produced unusable text in others. An external paid service was used to manually clean up the OCR into usable text. Some typed records contain either typed or handwritten corrections made by the revisers and editors in the Southern Life History Project; these include fixing typos, rephrasing sentence structure, editorial comments about the content and quality of writing, and making substantive edits to the content of the interview. For consistency, and because the crossed-out text was often unreadable, the machine-readable files used only the corrected versions of the text without any handwritten editorial comments. In a limited number of cases, due to physical imperfections, fading, or issues with the digitization process, small portions of the texts were unreadable. These are marked with the phrase "[text not clear]." The final machine-readable text files include all of the header information contained in the typed pages but exclude page numbers and any written comments that are not corrections to the main text.

The metadata and machine-readable versions of the life histories are published under the open-source GNU Public License (GPL-2).[[9]](#footnote-10) These can be downloaded in bulk. All of the data are contained in plain text forms that can be read by most data analysis software. Metadata is provided as CSV files, and the texts are provided as text files (one file per interview). All of the material is encoded using UTF-8 and internal consistency of the records was checked with a set of unit tests.[[10]](#footnote-11) The points on the interactive map and embedded figures in this project were all created using this metadata.

**Text Analysis**

Layer 4, alongside other archival evidence, presents several computational models to help understand and organize the 1248 life histories in the collection. For these analyses, duplicated interviews and interviews whose text was removed from the archive were not included in the analysis. When two slightly different versions of an interview appeared in the collection, we selected the longest text. After filtering, the collection contained 1106 life histories. Finally, these were cleaned to remove headers and instances of “[text not clear]” within the machine-readable text.[[11]](#footnote-12)

The two text analysis models used in our analysis—topic modeling and document clustering—are both built to analyze term frequencies (TFs). TFs count how frequently particular words or wordforms occur within each document within a corpus. To compute these counts, we passed the text of each life history through a language processing (NLP) pipeline using the R package cleanNLP. [[12]](#footnote-13) The package applies pre-built models to split the text into individual words and punctuation marks (tokenization), construct standardized forms of the words (lemmatization), and tagged each word with a part of speech code. [[13]](#footnote-14) Using this information, we constructed counts of lemmas for all lemmas tagged as nouns, verbs, adjectives, and adverbs. [[14]](#footnote-15) Because of the heavy influence of place names in the lexicon of the texts, we also explicitly removed any place names (i.e., cities and census-designated places) contained in any of the geographic columns in the metadata.[[15]](#footnote-16)

The life histories use a significant amount of “eye dialect” in which words are intentionally misspelled to signal what is problematically called “non-standard” pronunciations. Examples include “git” for “get” and “wuz” for “was”.[[16]](#footnote-17) The use of dialect is an interesting and important feature that we investigate in Layer 4. It offers insight into how spelling was used as a racial signifier and inculcated in white supremacist ideologies. At the same time, it dominates the signal within the topic models and document clusters, making it hard to detect other linguistic features. We have made available versions of our models with dialect included and with dialect removed for these reasons. To identify and remove dialect, we started by comparing each of the words identified by the NLP pipeline with a “standard” spelling dictionary of American English and removed words that were not included.[[17]](#footnote-18) Then, we looked at the most over-represented words in interviews of Black interviews and manually constructed a list of additional dialect terms to remove. These consisted of relatively uncommon terms that are English words but have alternative meanings. For example, the word "den" was used heavily in the corpus as dialect for the word "then." Using this approach as a strategy to explore other linguistic features, the approach is intended to offer another way to explore the topics that does not reduce the stories of people of color to primarily racist applications of dialect.

Using the term frequencies, we computed two sets of topic models, one with dialect terms and one without them. We used latent Dirichlet allocation (LDA) as implemented by the R package topicmodels to construct the models.[[18]](#footnote-19) LDA is a common and well-known technique in digital humanities and digital history research that calculates a probability for word co-occurrence.[[19]](#footnote-20) After some experimentation, we included 16 topics to display. The visualization of these topics on the digital project includes the probability distributions over words and the probability distributions of documents that each model defines.

Finally, document clustering was also applied to each of the two sets of term frequencies. Document clustering is the process of grouping all of the items in a corpus of texts into discrete groups, called clusters, based on linguistic features. Clustering is not commonly used in DH projects, primarily because of the momentum around the use of topic models. For many projects, including ours, it is useful to find groups of documents that use similar language features. It is possible to find some groups of documents by looking at the results of a topic model, but this approach will miss documents that cross between multiple topics. Also, the results of LDA are quite sensitive to a number of parameters, most notably the number of topics used. Document clustering can avoid these issues and move directly to the task of grouping together similar documents.

We used spectral clustering to produce clusters of documents. Spectral clustering is a relatively well-known technique in statistical computing for grouping together textual documents.[[20]](#footnote-21) We used the implementation from the R package casl to apply this algorithm.[[21]](#footnote-22) Document clustering is a hierarchical clustering method. The algorithm starts by splitting all of the documents in a corpus into two groups such that the two groups differ as much as possible in their usage of words. Then, the same algorithm is applied to each of these groups separately to split the entire collection into four subgroups. Applying again yields eight subgroups, then sixteen, and so forth. We applied this algorithm five times to yield a set of 32 clusters. Due to the iterative method, the clusters are related to one another. Documents in cluster 1 and cluster 2, for example, were split only in the final round of the algorithm.

All of the code to produce the text analysis models are made available under the open-source GNU Public License (GPL-2).[[22]](#footnote-23) The code works directly off of the data described in the previous section. It also includes the code to create the JSON files that serve as the backend for the digital project.

**Digital Platform**

The project is a part of a growing community of scholars, publishers, and foundations working together to expand the forms of academic scholarship. This includes expanding what counts as evidence, ways of knowing, and communicating knowledge. Along with archival evidence, the creation, analysis, and communication of data sits at the core of this project. The digital platform offered an opportunity to make visible *what* kindsof data we created, *how* we analyzed the data, and *why* through visualizations and text. It also provided a space to communicate scholarship through visual ways of knowing – graphs, maps, and interactive visualizations. Additionally, the interactive visualizations encourage visitors to explore the archive alongside us, build off our scholarship, and pose their questions. As Cox and Tilton have written, developing open access and interactive digital public projects can expand our argumentative strategies and reorient the reader/viewer as not just a person to be persuaded but as a participant engaged in humanistic inquiry and communication.[[23]](#footnote-24)

The platform is designed to pair text and interactive visualizations to convey the project’s arguments and scope. The project is structured in layers. Like a chapter, each layer offers insights into the social, political, and cultural work of the SLHP. In the same way that audiences have learned how to read a text to interpret an argument, audiences also have tools to interpret visualizations; there is a system of symbols and signifiers that people have learned to "read" visualizations that they employ daily. This project uses visualizations such as interactive mapping as a form of argumentation and then puts them in conversation with textual argumentation. As a result, this project is not strictly a textual book on a digital platform as we harness layers and the interpretive power of interactive visualizations to convey a set of arguments through an interactive platform made possible by the affordances of digital technologies.

The structure of the project builds off the spatial and visual turn in DH. As scholars such as Tara McPherson, Jentery Sayers, and Lev Manovich have argued, DH remains a text-heavy field.[[24]](#footnote-25) The call to use the affordances of computational methods to use visual ways of knowing, such as graphs and interactive visualizations, is amplifying. Led by scholars such as Stanford’s Richard White and over two decades of critical cartography, scholars have also used visualizations such as maps to convey scholarly knowledge and arguments. The project brings together these turns through the digital platform.

The digital platform is written using a number of open-source technologies. The main functional elements of the site are written using the popular JavaScript framework ReactJS. Additional JavaScript packages were used within ReactJS to add specific functionality, such as the use of router to create meaningful URLs and React-Dropdown to create interactive menus. Modern web standards are used to provide responsive, cross-platform compatible code using documented CSS3 and HTML5. The source code is available on GitHub.[[25]](#footnote-26) Any original code produced for the project is under a GPL-2 license; some derivative components are released under an alternative open-source license as required by their respective authors.

The mapping component of the website uses the JavaScript library Leaflet within ReactJS. Each of the maps was manually georeferenced using QGIS and projected into the Albers conic projection.[[26]](#footnote-27) This projection preserves areas and more accurately represents distances between points when compared to a Mercator projection.[[27]](#footnote-28) Map tiles were created using the Geospatial Data Abstraction Library (GDAL).[[28]](#footnote-29) The map tiles are served locally by the project; this is slightly less efficient than a purpose-built GIS server but is more stable and reliable for long-term preservation and access to the digital project.

The visualization of the topic models is written as a custom JavaScript code. It was adapted from a similar visualization produced for the Signs@40 project, made available under an open-source license.[[29]](#footnote-30) All of the data for the visualizations are stored as JSON files to simplify deployment and increase the stability of our application. This design choice removes the need for a backend database, making it relatively easy to run the website from alternative sources. Keeping the backend of the website minimal also facilitates long-term access to the project by minimizing the ways that the site could become obsolete. Code to create the topic model files are included in the repository containing the code for producing the topic models.

1. Federal Writers' Project papers #3709, Southern Historical Collection, The Wilson Library, University of North Carolina at Chapel Hill. [↑](#footnote-ref-2)
2. For more, see https://catalog.lib.unc.edu/catalog/UNCb2431314 [↑](#footnote-ref-3)
3. Courtney Rivard, Taylor Arnold, Lauren Tilton. “Building Pedagogy into Project Development: Making Data Construction Visible in Digital Projects.” Digital Humanities Quarterly, 13.2 (2019). [↑](#footnote-ref-4)
4. We recognize that there are debates about the term "normalized". In this context, we use the word to indicate a specific approach to structuring, which is referred to as "database normalization". [↑](#footnote-ref-5)
5. Hadley Wickham, “Tidy data,” *The Journal of Statistical Software*. 59.10 (2014). [↑](#footnote-ref-6)
6. “What Gets Counted Counts.” (2020). In Data Feminism. Retrieved from https://data-feminism.mitpress.mit.edu/pub/h1w0nbqp [↑](#footnote-ref-7)
7. Jessica Marie Johnson; Markup Bodies: Black [Life] Studies and Slavery [Death] Studies at the Digital Crossroads. *Social Text* 1 December 2018; 36 (4 (137)): 57–79. doi: <https://doi.org/10.1215/01642472-7145658>. See also Earhart, Amy E., and Toniesha L. Taylor. “Pedagogies of Race: Digital Humanities in the Age of Ferguson.” Debates in the Digital Humanities, JSTOR, 2016, pp. 251–64; Mahoney, Jennifer, et al. “Data Fail: Teaching Data Literacy with African Diaspora Digital Humanities.” Journal of Interactive Technology & Pedagogy, Dec. 2020; Noble, Safiya Umoja. “Toward a Critical Black Digital Humanities.” Debates in the Digital Humanities, 2019, pp. 25–35. [↑](#footnote-ref-8)
8. These are not exhaustive, focusing most on flagging a few interviews of people of Greek and Cuban descent. [↑](#footnote-ref-9)
9. Arnold, Taylor, Alexander, Emeline Blevins, Rivard, Courtney, Tilton, Lauren, & Wexler, Laura. (2020). FWP Life History Project in the American South: Machine Readable Text and Metadata (Version 1.0). Zenodo. http://doi.org/10.5281/zenodo.3865765 [↑](#footnote-ref-10)
10. The tests ensure, for example, that all of the table keys link together and that the interviews appear to correctly match their metadata (i.e., contains the title of the interview in the text and has a reasonable number of words given the number of pages). [↑](#footnote-ref-11)
11. For more about debates on “cleaning” data, see Katie Rawson and Trevor Muñoz, “Against Cleaning” in *Debates in the Digital Humanities 2019* eds. Lauren Klein and Matt Gold. Minnesota: University of Minnesota, 2019.

    https://dhdebates.gc.cuny.edu/read/untitled-f2acf72c-a469-49d8-be35-67f9ac1e3a60/section/07154de9-4903-428e-9c61-7a92a6f22e51 [↑](#footnote-ref-12)
12. Taylor Arnold. A Tidy Data Model for Natural Language Processing Using cleanNLP The R Journal, 9.2 (2017): 248-26. [↑](#footnote-ref-13)
13. Honnibal, Matthew and Montani, Ines and Van Landeghem, Sofie and Boyd, Adriane. “spaCy: Industrial-strength Natural Language Processing in Python." 2020, Zenodo. 10.5281/zenodo.1212303. [↑](#footnote-ref-14)
14. Bies, Ann, Justin Mott, Colin Warner, and Seth Kulick. “English Web Treebank LDC2012T13”. *Linguistic Data Consortium*, 2012. https://doi.org/10.35111/m5b6-4m82 [↑](#footnote-ref-15)
15. These should have been removed by the part of speech tags, which have a separate category for proper nouns, but errors are common with this process, particularly when proper capitalization is not used (not uncommon in direct quotes in this corpus). [↑](#footnote-ref-16)
16. Morrison, Toni. *Playing in the Dark: Whiteness and the Literary Imagination*. Harvard University Press, 1992. [↑](#footnote-ref-17)
17. Németh, L., Trón, V., Halácsy, P., Kornai, A., Rung, A., & Szakadát, I. (2004). Leveraging the open source ispell codebase for minority language analysis. *Proceedings of SALTMIL*, 56-59. [↑](#footnote-ref-18)
18. Hornik, K., & Grün, B. (2011). topicmodels: An R package for fitting topic models. *Journal of statistical software*, *40*(13), 1-30. [↑](#footnote-ref-19)
19. For a good technical overview focused on humanities applications, see David Blei, “Topic Modeling and Digital Humanities,” *Journal of Digital Humanities*, 2.1 (2012). [↑](#footnote-ref-20)
20. William Donath and Alan Hoffman (1972). "Algorithms for partitioning of graphs and computer logic based on eigenvectors of connections matrices". *IBM Technical Disclosure Bulletin*. [↑](#footnote-ref-21)
21. Taylor Arnold, Michael Kane, Bryan Lewis. *A Computational Approach to Statistical Learning*. New York: Chapman & Hall/CRC Texts in Statistical Science (2019). [↑](#footnote-ref-22)
22. “FWP Life History Project in the American South: Text Analysis Code,” https://github.com/statsmaths/fwp-life-histories-analysis. Accessed 28 February 2021. [↑](#footnote-ref-23)
23. Jordana Cox & Lauren Tilton (2019) The digital public humanities: giving new arguments and new ways to argue, Review of Communication, 19:2, 127-146, DOI: 10.1080/15358593.2019.1598569 [↑](#footnote-ref-24)
24. Manovich, Lev. *Cultural* Analytics, Cambridge: MIT Press, 2021. McPherson, Tara. "Introduction: Media Studies and the Digital Humanities." *Cinema Journal* 48, no. 2 (2009): 119-23. Accessed March 1, 2021. http://www.jstor.org/stable/20484452; Jentery Sayers, ed. *The Routledge Companion to Media Studies and Digital Humanities*. Routledge, 2018. [↑](#footnote-ref-25)
25. https://github.com/statsmaths/voice-of-a-nation [↑](#footnote-ref-26)
26. Gray, J. (2008). “Getting started with quantum GIS.” *Linux Journal*. [↑](#footnote-ref-27)
27. Kimerling, J. A., Overton, S. W., & White, D. (1995). Statistical comparison of map projection distortions within irregular areas. Cartography and Geographic Information Systems, 22(3), 205-221. [↑](#footnote-ref-28)
28. Warmerdam, F. (2008). The geospatial data abstraction library. In *Open-source approaches in spatial data handling* (pp. 87-104). Springer, Berlin, Heidelberg. [↑](#footnote-ref-29)
29. Andrew Goldstone, Susana Galán, C. Laura Lovin, Andrew Mazzaschi, and Lindsey Whitmore. “An Interactive Topic Model of Signs”, https://github.com/signs40th/topic-model [↑](#footnote-ref-30)