GoodlyLabs

From ‘Static’ to ‘Research-Ready’:
Preparing Congressional Archives for Computational Text Analysis

a Statement of Work
from GoodlyLabs to The Social Science Research Council
April 14, 2017

Background & Narrative

Libraries, archives, and other memory organizations are nearly as interested in ensuring their documents are well-used as they are in collecting and preserving them for posterity. But the researchers who access and use digital archives often find them organized only for traditional, forensic queries –ideal for researchers interested in reading just one or a few documents at a time. But digital humanists and data scientists seeking to analyze thousands of archived documents using new techniques find that they must first invest (too often) dozens of hours arduously re-organizing and cleaning the data to get them to a ‘research-ready’ state.

Since digital culture researchers are typically found outside of memory organizations, they are often unable to influence the policies and procedures affecting the presentation and accessibility of archival data. Instead, they spend mind-numbing hours pointing, clicking, saving, and repeating. And after considerable toil, they usually do not share the fruits of their labor (the better organized, 'research-ready' data) with their colleagues or the memory organization from whence it came. As a result, independently working scholars duplicate painful struggle and memory organizations learn little about how they could make their archives more useful for (especially digital) scholarship.

The SSRC can support both communities by supporting the development and dissemination of practices that transform 'static' archives into 'research-ready' databases serving broader research communities. The GoodlyLabs and Berkeley's Computational Text Analysis Working Group (CTAWG) are engaged in one such effort right now, converting the Government Publishing Office’s static archive of Congressional Records and Committee Reports into a research-ready database of Congressional activity (since 1994).

Our team’s computational approaches to natural language allow us to instruct a computer to "read" thousands of documents per hour, extracting socially meaningful textual units, like speech acts (embedded in a conversational sequence of such acts), and link those units to all the data we have about the speaker, where they are from, the constituents they represent, their gender, party affiliation, and so forth. Such well-defined, socially relevant units of text can be analyzed by topic (using known and our custom techniques), and across time as well. But the hundreds of possible research inquiries into an archive like the Congressional Record (CR) always start with the same arduous process of gathering the data from the archive, chunking it into meaningful text units, and linking them to external databases housing data about the people and entities appearing in the documents.

The present Statement of Work from GoodlyLabs to The Social Science Research Council (The SSRC) describes work over the term May 15th to November 15th, 2017, to result in a pilot database of the Congressional Record, and documentation (including programming scripts and tutorials) that may serve as useful guides for memory organizations, digital culture researchers, and anyone seeking to transform archival data from static to research-ready.

Phases of the Work and *Deliverables*

Phase I: Acquiring and Cleaning Data

Our work identifying, acquiring, and cleaning *Congressional Records* (CR) data has already begun. Since data quality, readiness, and temporal breadth vary across the data sources we have identified, we will be working with two datasets in parallel.

- **GPO Data:** We have already acquired the U.S. Government Publishing Office's (GPO) archive of the CR from their website, and are able to begin Phase I work writing documentation and tutorials explaining our process.
    - Details: GPO-data were only accessible as hundreds of clickable links spread across several web pages, a display of data fundamentally incompatible with an era when computers "read" thousands of documents at a time.
    - Limiting Factor: The GPO only maintains CR data since 1994. So, while this dataset is appropriate for a 'research ready' pilot database (and the creation of associated documents and code), we will also begin immediately pursuing data from the Hathi Trust.

- **Hathi Data:** Librarians on our team, Cody Hennesy and Jesse Silva, have already opened conversations with Hathi Trust about collecting, OCR-ing, and cleaning their CR data, which (except for some short periods) includes all of Congressional history. Getting Hathi data to 'research-ready' will require considerable effort on our part. So, during the term of this Statement of Work, we will be focusing the bulk of our efforts on completing a GPO-based pilot database (from the GPO data described immediately above).
    - Details: We are excited about the Hathi Data. It will be able to generate many more learning opportunities for Digital Culture stakeholders interested in preparing research-ready databases. And ultimately, it could result in a database of great historical scope and significance for scholars. So, we will begin the process of acquiring and cleaning Hathi data during the term of this Statement of Work (even though it is not listed as a deliverable).

***Deliverables**: Workbook including code and curriculum demonstrating and explaining how others can acquire and clean archival data. Progress report/email.*

Phase II: Structuring, Chunking, and Tagging the Text

Once we have completed Phase I on the GPO data (including documentation and workbook tutorials), we will begin using automated approaches to identify and label speech acts in the record. This process, often called 'chunking,' requires specialized knowledge about how to architect and code a streaming, regular expressions-based text identification/extraction program. It also requires a practical knowledge of the archival documents under investigation, allowing researchers to write effective rules specifying where the computer can find socially relevant text units like speech acts. We will demystify this process with an educational (Jupyter or Rmarkdown) workbook showing researchers and memory organizations how to:

- Identify the overall structure of their texts
- Identify social units in their documents of interest
- Develop an XML schema (tag set) for the structure and the units of interest to researchers
- Write expressions to identify similar features and units across the corpus
- Design a computationally efficient, streaming chunking program
- Test and validate the chunking program

> ***Deliverables**: Workbook including code and curriculum demonstrating and explaining how others can structure, chunk, and tag their archival documents. Progress report/email.*

Phase III: Packaging the Data for Researchers and Archives

Once GPO CR data have been structured and tagged, we will work with UC Berkeley Research Computing (BRC) to create a data storage infrastructure that will allow researchers to query our CR data in Python and R environments. This will require, as a first step, the building of an XML Database. Then, since we want to empower users to query and use our database, we will set up a (Docker) research environment to house the database, stable versions of R and Python text analysis libraries, and some instructional Jupyter and R-Markdown notebooks showing researchers how to get started. (We expect these workbooks will be useful across a range of pedagogical contexts, including D-Lab's popular Python and R workshops.) With these materials, any researcher who has completed a basic 'Introduction to Text Analysis' course should be able to begin producing novel research findings heretofore buried in 'open' but static and difficult-to-search archives. Finally, we will write a covering report for the whole pilot project addressing an audience of memory organizations and other members of the Digital Culture community.

- XML Database (BaseX)
- Docker on Jetstream
- Python and R environments
- Example Jupyter and R Studio notebooks

***Deliverables**: GPO database in an accessible Docker environment. Workbook including code and curriculum demonstrating and explaining how others can query the database and use it for basic text analysis tasks. Covering report for a Digital Culture (SSRC and memory organizations) audience.*

Phase X: Beyond the Scope of this Statement of Work

The primary objectives of the efforts specified by this Statement of Work are to produce a GPO pilot database useable by researchers, and a set of educational programming workbooks demonstrating how others can move archives "from static to research-ready." But we will note, here, that we also hope for the project to grow after (and even while) this work is completed. As explained above, our team will soon begin cleaning Hathi Trust CR data even though that work is not listed as a deliverable. We will also wish to link our efforts with other researchers studying the CR or Congressional activity. And with our CTAWG partners, we have already applied for a small amount (~$7K) of UC Berkeley campus funding to support the expansion of this pilot project over the 2017-2018 Academic year. We will be thankful for The SSRC's support in thinking through how the project might grow and who we might partner with in the future.

Timeline

The term of this Statement of Work shall begin on May 15th, 2017 and expire on November 15th, 2017. We (GoodlyLabs) are confident that we can deliver the prototype database (and all other deliverables listed above) within this 6-month term.

Deliverables: *The SSRC can expect deliverables for each of the three phases of work on or about the 15th of August, September, and October, respectively.*

Team

Nick Adams, Ph.D., Founder & Conductor, GoodlyLabs

Additional Team Members
Johannes Fritz, D-Lab's Computational Text Analysis Working Group, UC Berkeley
Aaron Culich, Sr. Research Computing Architect, Berkeley Research Computing, UC Berkeley
Jesse Silva, Federal Documents and Political Science Librarian, UC Berkeley
Cody Hennesy, E-Learning and Information Studies Librarian, UC Berkeley
Rebecca Fan, D-Lab's Computational Text Analysis Working Group, UC Berkeley
Niek Veldhuis, Professor of Assyriology, UC Berkeley
Laura K. Nelson, Assistant Professor of Sociology, Northeastern University
Ben Gebre-Medhin, PhD Candidate in Sociology, UC Berkeley
Scott Paul McGinnis, PhD Candidate in History, UC Berkeley
Various members, D-Lab's Computational Text Analysis Working Group, UC Berkeley

GoodlyLabs