
Baylor University Sample Humanities Project DMP

A Data Management Plan created using DMPTool

Creator: Christina Chan

Affiliation: Baylor University

Template: National Endowment for the Humanities (NEH)

Project abstract:

This project will examine the letters of little-known abolitionist Abe O. L'Sean. The entire correspondence is divided between a collection of 218 letters housed at the Abe O. L'Sean library in Augusta, GA and a smaller collection of 112 letters at Baylor. The letters will be digitized at the Reilly Digitization Center. The 86 typewritten letters will be initially transcribed using OCR, and a team of seven undergraduates will transcribe the remainder. The PI and two graduate students will review and correct all the transcriptions. The letters will be textually analyzed for references to Abe O. L'Sean personal life (although he lived his whole life in Augusta, his three wives were from France, Australia, and Hong Kong), his business (he was a bottle cap importer/exporter), his religious activities (he was an ordained pastor with the Church of the Flock) and his abolitionist activities as well as references to current events. A sentiment analysis will also be conducted on the letters. The letters will also be analyzed to connect the locations of the sender and recipient over time.

Last modified: 06-06-2019

Copyright information:

The above plan creator(s) have agreed that others may use as much of the text of this plan as they would like in their own plans, and customize it as necessary. You do not need to credit the creator(s) as the source of the language used, but using any of the plan's text does not imply that the creator(s) endorse, or have any relationship to, your project or proposal

Baylor University Sample Humanities Project DMP

Roles and responsibilities

The Riley Digitization Center at Baylor University will be responsible for the digitization of the letters. The Abe O. L'Sean Library and the Baylor Libraries will be responsible for the digital copies and transcriptions of the letters that they respectively own.

The PI will oversee the transcription portion of the project. If she is unable to complete the transcription of the project, the head of the Riley Digitization Center will continue this portion of the project.

The PI of the project will be responsible for the textual analysis and the resulting data.

Expected data

The digitized letters and their transcriptions will be stored on contentDM servers at both libraries. It is anticipated that there will be between 200-300 images and 133 transcripts totaling approximately 5GB of data. The file naming convention will include the date, sender, and recipient. ContentDM will allow the images of letters and their transcripts to be searchable and available to the public. The data will be added incrementally to the servers as the letters are digitized and transcribed. Once the collection is complete, it will be made available to the public.

During the project, all analyzed data (eg. word frequency data, location data, etc.) will be stored in Box. Each type of analysis will generate at the most 5 files resulting in 20-30 files totaling less than 100MB. After the project has been completed, the analyzed data will be deposited in BEARdata, Baylor University's data repository along with README files which explain the steps of the analysis (eg. keywords used, how to upload the transcript data into the analytical/data visualization tools) and describe the format of the data files. BEARdata allows datasets to be searchable and citable.

Data visualizations will be available through the Abe O. L'Sean webpage/blog. Screenshots of various data visualizations will also be deposited into BEARdata. We anticipate generating approximately 50 screenshots totaling approximately 500M. The file naming convention will include the type of data visualization and the parameters of the visualization shown in the screenshot.

Period of data retention

The letter images and transcriptions will be available on completion of the digitization and transcription process and be stored in contentDM indefinitely. If the libraries switch to a different digital content platform, the data will be migrated to the new system.

Analyzed data described above will be deposited into BEARdata within 1 year of the completion of the project and will be held indefinitely.

Data visualization on the Abe O. L'Sean Library webpages will be available within 3 months of the completion of the project for a minimum of 5 years after the completion of the project.

Data formats and dissemination

The digitized letters will be stored as high quality .tif files. The transcripts will be stored as .docx, .txt, and .pdf files. The image and transcript content is owned by the respective libraries and available to the public under a CC-BY license.

The README files will be stored as .txt files; analyzed data will be stored as .xlsx and .csv files; and screenshots of various data visualizations as .tif and .jpg files. The README files will be available openly to the public. The data files will be available under a CC-BY license with permission of the project director. BEARdata allows the public to automatically contact the PI for access permission. Screenshots will be available openly through a CC-BY license.

The family of Abe O. L'Sean gave permission for all materials to be available public when they donated the materials for the library. No private materials are part of the collection.

Data storage and preservation of access

Both digital content platform (contentDM) used at both libraries and BEARdata (an instance of Harvard Dataverse maintained by the Texas Digital Library) have comprehensive preservation plans and are backed up using Chronopolis.

The website/blog hosted by Abe O. L'Sean Library will be archived yearly using Archive-It.