Plan Overview

A Data Management Plan created using dmptool

Creator: Austin Nelsen

Affiliation: University of Florida

Funder: National Endowment for the Humanities (NEH)

Template: Digital Curation Centre

Project abstract:

This project seeks to leverage the powerful tools of data management in preparation for a future work of textual analysis of 17th century colonial administrative documents held by the Arquivo Histórico Ultramarino in an effort to better understand Palmares, a maroon community founded in colonial Brazil by runaway slaves. Since these formerly-enslaved peoples left behind no documents of their own, and reading colonial documents to understand anti-colonial societies such as Palmares presents its own challenges, innovative strategies are needed in order to overcome a lack of traditional historical sources. Additionally, this project hopes to serve as an example of new methods and outcomes made possible through the intersection of traditional archival research and the tools of the digital humanities. In accomplishing both of these aims, this project offers a unique opportunity to demonstrate how emerging technology may help historians find new answers to old questions.

Last modified: 04-30-2021

Image Data Extraction and Organization for Textual Analysis of Historical Sources

Data Collection

What data will you collect or create?

This project will consist of two main datasets. The first set consists of 288 .jpg files representing photos taken by the author of archival documents and accompanying official documentation. The second dataset will be the corpus of text extracted from said archival documents, in .txt format. These datasets will serve as the basis of a future formal analysis of the datasets, taking the form of a journal article. The file formats selected are well-regarded for long term access and preservation, have a demonstrated history of longevity in spite of the ubiquitous volatility of digital file formats, and are likely to remain in use well into the foreseeable future.

How will the data be collected or created?

Photos of archival documents have already been taken and stored, retaining original filenames and organization conventions created when the photos were taken. Although such an approach contradicts currently-accepted best practices of data management, historians are well-positioned to appreciate how information that may seem useless to contemporaries can often transform into unexpected -and critically important- sources of information at a later date. Fortunately, modern photo management software (such as Tropy) leverages metadata in order to take advantage of all of the benefits of systematic data organization and file naming conentions, without overwriting or otherwise sacrificing any data that may one day prove to be important.

Once the photos have been organized and stored in Tropy, relevant metadata will be extracted from the documents and added to the document's record in Tropy. Wherever possible, this process will be undertaken by automated OCR via RStudio code. The metadata targeted for extraction from images include: date and location of document creation, authors, subjects, the document's official archive designation, to whom it was sent, and a brief summary of the document's contents taken from official archival documentation.

Separately, a corpus of text will be created using Otter.ai transcription software, which will transcribe the author’s verbal reading of archival documents into text data. Quality control of transcription data will then be accomplished by importing text output from Otter.ai’s transcription into Microsoft Word 365 ,which will then used to render Otter.ai’s text output back into audio. The resulting audio will be compared, in real time, to the original images of the archival document used for the initial transcription. After passing this quality check, the text of the document will be added to the corpus dataset text file. Finally, transcription text will also be entered into the document's metadata in Tropy for future research involving these same documents.

Documentation and Metadata

What documentation and metadata will accompany the data?

Metadata will adhere to the standards set out by Michner et al (1997). The main organization of metadata will be done in Tropy, an open-source program, which will export the metadata as a .json file. Metadata produced by Tropy will be created and organized hierarchically at the document level, allowing for a single repository for all relevant data gathered across multiple sources. In addition, a

metadata text file adhering to the guidelines of Michner et al (1997) will be included with the finished datasets in order to communicate information, such as project methods and distribution licensing, that is important but unrelated to the archival documents being studied.

Ethics and Legal Compliance

How will you manage any ethical issues?

Owing to the length of time that has elapsed since these documents were created, privacy concerns of the individuals mentioned therein are negligible. However, as patrimony of Portugal, the utmost care was taken when handling and photographing the documents that are included in this project. Where digitized versions of the Arquivo Histórico Ultramarino's relevant holdings were available, digital versions were consulted in order to preserve the fragile condition of the archive's collection. The documents that comprise the image data of this project, however, were either not-yet-digitized or damaged to a degree of severity that rendered digitizations partially illegible. As such, the images of these documents represent a valuable resource for historians studying maroon communities in colonial Brazil, and to the extent possible, efforts will be undertaken to ensure that their textual content is widely and freely accessible.

How will you manage copyright and Intellectual Property Rights (IP/IPR) issues?

All archival data in state archives of Portugal are considered to be patrimony of the state, and consequently available for analysis. However, publication of images of archival documents requires securing permission from the holding archive. All documents that comprise this project are part of the collection of the Arquivo Histórico Ultramarino. Text of archival documents are considered to be part of the public domain under Código do Direito de Autor e dos Direitos Conexos, aprovado pelo Decreto -Lei n.º 63/85, de 14 de março. Photos of archival documents are governed by Decreto -Lei n.º 31/2019, de 3 de maio, which explicitly grants researchers permission to take digital photos of archival documents for personal or academic use. Any other use, including any publication of photos of archival documents, requires prior authorization from the Direção-Geral do Livro, dos Arquivos e das Bibliotecas. See Despacho n.º 6852/2015 de 19 de junho for additional information on authorized uses of images of archival documents. As such, the publication of document photos will require permission from the Arquivo Histórico Ultramarino before being made available to the public.

The legal barrier preventing the publication of the image dataset should not, however, be considered an impediment to independent verification of the project's findings. With the exception of the eight documents in the dataset which are from the AHU Fundo CU, Série 001 Angola collection, digitized versions of all documents in the image dataset are available for consultation online as part of the digital access project Projeto Resgate of the Brazilian National Library. The entirety of the AHU Fundo CU, Série 015 Pernambuco collection can currently be found at: http://resgate.bn.br/docreader/DocReader.aspx?bib=015_PE&pagfis=1. Additionally, the dataset containing text of the transcribed archival documents can be published and used in accordance with the legal framework of the Código do Direito de Autor e dos Direitos Conexos.

Storage and Backup

How will the data be stored and backed up during the research?

All data will be stored in three forms: external hard drive in off-site storage, local hard drive, cloud storage (Google Drive).

How will you manage access and security?

All data storage forms currently require either physical access to secure areas or single sign on (SSO) authorization utilizing robust security including two-factor authentication. The data itself is legally considered as public domain, and as such, represent a very low risk if accessed and disseminated without authorization.

Selection and Preservation

Which data are of long-term value and should be retained, shared, and/or preserved?

Until the Arquivo Histórico Ultramarino completes its long-running digitization and restoration projects and all documents are freely-available online in legible form, photos included in this project represent one of the very few -if not only- digital representations of a handful of individual colonial documents. Most notably, this includes the eight documents from the archive's Angola collection, which, as of this writing, has not yet begun to be digitized. Additionally, metadata produced by Tropy will remain relevant to historians studying maroon societies in colonial Brazil and will add to the personal archives of these historians.

What is the long-term preservation plan for the dataset?

In addition to redundant storage of data following currently-accepted standards of preservation, the project has the long-term goal of permanent online preservation utilizing an open-source data repository platform such as Zenodo. There are, however, current legal obstacles to this goal. For more information, see the Data Sharing and Ethical and Legal Compliance sections of this Data Management Plan.

Data Sharing

How will you share the data?

This project has the long-term goal of permanent online preservation utilizing an open-source data repository platform such as Zenodo. However, before the dataset containing images of archival documents can be uploaded to any publication or repository, written permission must be secured from the Arquivo Histórico Ultramarino. In the meantime, the dataset containing text of transcribed archival documents, along with all project metadata, are available for use until such time as formal authorization from the Arquivo Histórico Ultramarino is secured and all data can be held in a publicly-available repository.

Are any restrictions on data sharing required?

According to the terms of access of the Arquivo Histórico Ultramarino, images of archival documents cannot be published or otherwise widely-disseminated without the express permission of the Arquivo Histórico Ultramarino, and are only available for the personal or educational use of the person who took the photos.

Responsibilities and Resources

Who will be responsible for data management?

As sole author of the project, Austin Nelsen is fully responsible for the execution of the project's formal data management plan, including data capture, metadata production, storage and backup, archiving, and securing publication permissions.

What resources will you require to deliver your plan?

No additional resources are required.

Planned Research Outputs

Image - "Images of Archival Documents"

The cornerstone of the project, 278 photos of 23 archival documents as designated by the Arquivo Histórico Ultramarino's organizational structure. All of these documents span multiple pages, and multiple photos were taken of each page to ensure a legible image would be available for paleography and concomitant transcription. These data are formatted as .jpg files. Publication of this dataset is restricted under Portuguese law, and thus this dataset remains restricted to the personal or academic use of the individual (Austin Nelsen) who took the original photos in the archive.

Text - "Corpus of Text Transcribed from Documents"

A single text file in .txt format, this includes the combined text from all transcribed documents for the purposes of textual analysis, such as sentiment analysis. Ideally this file would include the transcribed text from all documents included in the project, however, 17 of the 23 individual documents have suffered significant damage, ranging from severe to near-total obliteration of text. Combined with the difficulties inherent to 17th-century Portuguese paleography, transcription was - and continues to be- an exceedingly-laborious endeavor. Currently, roughly 25% of the documents have been transcribed, resulting in a corpus of sufficient size so as to allow meaningful textual analysis, but transcription will continue to be an ongoing aspect of the project, as more transcriptions are completed and added to the corpus text file. In order to track versions, the online repository Github will be used to ensure the most up-to-date file is readily available.

Data paper - "Final Textual Analysis"

The final output of the project will be a journal-length article or paper extrapolating findings from the textual analysis of the documents. Said paper will detail the methodology of the project, pointing other historians toward the readily-available and easy-to-use tools for not only textual analysis but data management and other important topics common in digital scholarship, yet remain largely overlooked by historians. The selection of final publication venue will prioritize open-access journals.

Text - "Project Metadata"

All metadata from the project will be available for public use. This metadata will comprise of two files, the first is a .txt file with metadata as suggested by Michner et al (1997). The second metadata file is a .json file exported from Tropy containing all metadata extracted from OCR processes, as well as manual transcription, organized around the archival document as the fundamental unit of organization.

Planned research output details

Title	Type	Anticipated release date	Initial access level	Intended repository(ies)	Anticipated file size	License	May contain sensitive data?	May contain PII?
Images of Archival Documents	Image	2028-01-01	Restricted	Zenodo	2 GB	Creative Commons Attribution Non Commercial No Derivatives 4.0 International	Yes	No
Corpus of Text Transcribed from Documents	Text	2021-05-01	Open	GitHub	2 MB	Creative Commons Attribution Non Commercial No Derivatives 4.0 International	No	No
Final Textual Analysis	Data paper	2022-01-01	Open	None specified	3 MB	Creative Commons Attribution 4.0 International	No	No
Project Metadata	Text	2021-05-01	Open	None specified	1 MB	Creative Commons Attribution Non Commercial No Derivatives 4.0 International	No	No