README: Slavery, Abolition, Emancipation and Freedom Project Collection

This README is based upon the Datasheets for Datasets template developed by Gebru et al. (2021). Elements have been modified to better reflect the collection's origin as an library collections-as-data project involving digitized archival materials held by Houghton Library at Harvard University.

The digital objects these datasets are based upon are available at: https://curiosity.lib.harvard.edu/slavery-abolition-emancipation-and-freedom. The materials represented in the dataset were selected for digitization by Dorothy Berry, with input from other Harvard curatorial staff members.

Table of Contents

README: Slavery, Abolition, Emancipation and Freedom Project Collection 1
- Motivations 3
  - For what purpose is the data created? 3
  - Who created the collection or datasets (team, research group, company, etc.)? 3
  - Who funded the creation of the dataset? 3
- Composition 3
  - What does the data represent? Are there multiple data types or relationships between items or files? 3
  - How many records are there? (Rows, observations, etc.) 4
  - Is the data a complete representation of possible observations or is it a sample? 4
  - What data is included? 4
  - Are relationships between items, observations, or files made explicit? If so, please describe. 4
  - Are there any errors, sources of noise, or redundancies in the dataset? 4
  - Is the dataset self contained? Does it link to or otherwise rely on external resources? 4
  - Does the dataset contain any confidential data? 5
  - Does the collection datasets contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? 5
- Collection Process 5
  - What mechanisms, instruments, or processes were used to collect the data? 5
  - If the dataset is a sample, what was the sampling strategy? 6
- Preprocessing, cleaning, and labeling 6
  - Was any preprocessing, cleaning, or labeling done to the data? Eg: tokenizing, feature extraction, processing of missing values, or removal of outliers? 6
  - Is the raw data included in this dataset or otherwise preserved and accessible? 7

Is the software used to preprocess the data available? Include a link to the access point. 7
Uses 7
What (other) tasks could the dataset be used for? 7
Is there anything about the composition of the dataset or the way it was collected and preprocessed that might impact future uses? 7
Distribution 7
Will the dataset be distributed to third parties or made available outside of the creating entity? 8
How will the dataset be distributed? 8
Will the dataset be distributed under a copyright or other intellectual property (IP) license and/or under applicable Terms of Use (ToU)? 8
Maintenance & Stewardship 8
Who will be supporting, hosting, and maintaining this dataset? 8
How can the owner, curator, or manager of the dataset be contacted? 8
Will the dataset be updated? 8
Will older versions of the dataset continue to be supported, hosted, or maintained? 8
If others want to extend, augment, build on, or contribute to the dataset, is there a mechanism to do so? 9
References 9

Motivations

This section is intended to clarify intentions behind the creation of a dataset and promote transparency of funding sources.

For what purpose is the data created?

- This data collection is intended to support the computational analysis of images, OCR text, and machine generated transcriptions of manuscript materials for research purposes.

Who created the collection or datasets (team, research group, company, etc.)?

- This Harvard Dataverse collection and its datasets was created by a joint team with members from Houghton Library, Harvard Library's Open Scholarship and Research Data Services, Harvard Library's Digital Scholarship Program, Harvard Library's Imaging Services and the Harvard Dataverse data curation and repository development teams.

Who funded the creation of the dataset?

- The digitization of the original source materials and the curation of this data collection was funded by the Harvard Library Advancing Open Knowledge Grant (https://harvardlibrarybulletin.org/news/2020%E2%80%932021-winners-harvard-library%E2%80%99s-advancing-open-knowledge-grants).

Composition

This section is intended to provide reusers with sufficient information to evaluate and make informed choices about how to use the dataset for their chosen tasks. Questions related to data about people are grouped together at the end.

What does the data represent? Are there multiple data types or relationships between items or files?

- Each dataset corresponds to a digital object which in turn corresponds to a primary source in the Houghton Library holdings pertaining to Black history and culture. Each dataset contains a mix of digitized images, XML metadata, OCR transcriptions, automatically generated, and experimental transcriptions of handwritten materials. A full inventory of all MIME types for each dataset is available in the following dataset: https://doi.org/10.7910/DVN/ZFVUHI.
- Handwritten material in our collection includes letters from famed Black abolitionists, letters from formerly enslaved peoples to the Freedmen’s Bureau, personal recollections of Gullah-Geechee sacred music, and more. Library collections, like SAEF, represent test beds for tools and systems that transcend discipline and media.

How many records are there? (Rows, observations, etc.)

- Total of 1229 datasets with 78,653 individual files
- See full dataset inventory: saef_hdv_dataset_inventory.tab (https://doi.org/10.7910/DVN/ZFVUHI)
- See full inventory of datafiles: saef_hdv_datafiles_inventory.tab (https://doi.org/10.7910/DVN/ZFVUHI)

Is the data a complete representation of possible observations or is it a sample?

- SAEF data is a sample representative of a larger set of Black culture and history collections at Houghton Library. The Library was not able to include everything in the dataset at the time of creation and our collection development in this area is ongoing.

What data is included?

- OCR transcription data includes 1) annotated images; 2) json files containing positional information, per transcription, and a confidence value; and 3) plaintext transcription, per image.

Are relationships between items, observations, or files made explicit? If so, please describe.

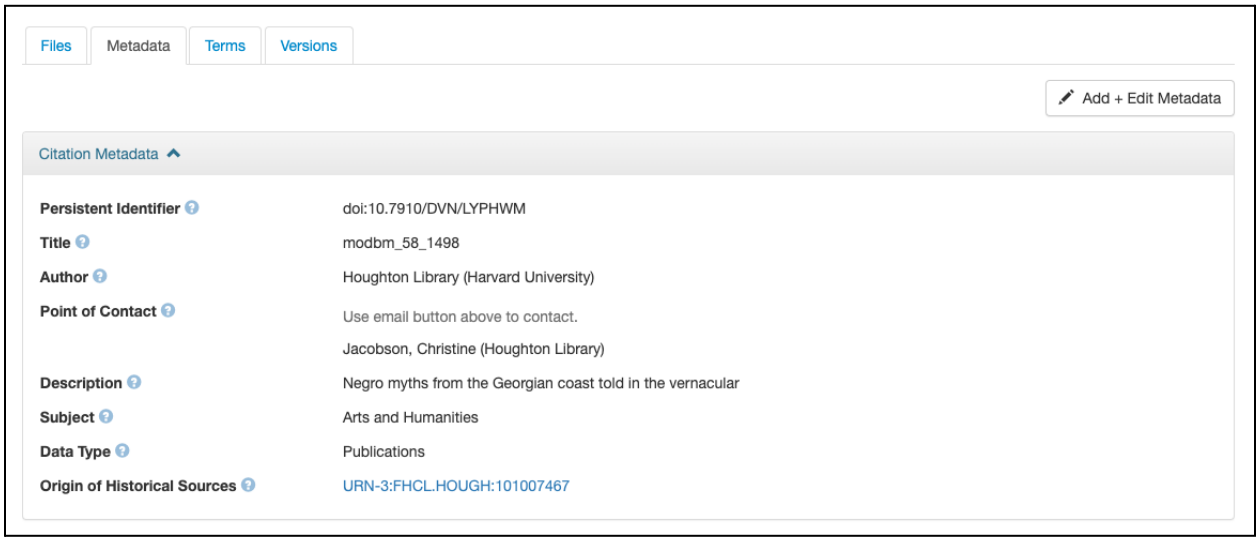
- Yes. Each dataset contains automatically generated tabular files called **Relationship Files** that map the relationship of all files in the dataset to one another.

Are there any errors, sources of noise, or redundancies in the dataset?

- OCR outputs are limited by 1) the lack of ground truth data and 2) the use of 3rd party, "big tech" vendor solutions for generating transcriptions of handwritten documents. As others have noted (See: "Further Reading" section, below), bias enters machine learning workflows at multiple points, beginning with the original digitization and proceeding to the algorithmic bias that influences.
- It's therefore useful to think about the transcriptions derived through this SAEF OCR workflow as a way for non-technical researchers who may be seeking to get a feel for the shape of, and patterns in, unwieldy datasets. So, the workflows and technologies documented here and elsewhere, while fraught, represent a low-cost, mostly automated way to begin exploring the contents of large, handwritten document collections.

Is the dataset self contained? Does it link to or otherwise rely on external resources?

- Each dataset contains all files associated with a specific digital object in the original SAEF collection. On the Metadata tab for each dataset, a link to the original digital object, in context, is made available in the dataset's **Origin of Historical Sources** field.



Does the dataset contain any confidential data?

- All datasets and datafiles are publicly available via the Harvard Dataverse Repository.

Does the collection datasets contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?

- Yes. Please see frequently asked questions about the digital collection and tangible collections upon which these datasets are based here:
    - Q: Why does this collection include some anti-Black, racist material?
        - Source: https://curiosity.lib.harvard.edu/slavery-abolition-emancipation-and-freedom/about/frequently-asked-questions

Collection Process

In addition to the goals mentioned above, these questions are designed to help other researchers create alternative datasets with similar attributes.

What mechanisms, instruments, or processes were used to collect the data?

- OCR processing moved handwritten document transcription data associated with the SAEF project between three primary platforms: Harvard's Digital Repository Service (DRS), where the source material is archived; NextCloud, our secure, medium-term storage solution, where document image data (and metadata) are staged, and (using their desktop client), absolute filepaths are simulated for us in the Python notebook; and the Harvard Dataverse, a discipline-agnostic data hosting platform for distributing

scholarship. Central to the SAEF_OCR throughput notebook is Mike Hucka's HandPrint library, which automates the batch loading of images to various 3rd party cloud services (Google, Amazon, Microsoft) for handwriting transcription

If the dataset is a sample, what was the sampling strategy?

- The materials selected for this project represent highlights of Houghton's Black history collections, but are far from comprehensive. This is a curated look at the various ways Black voices and histories appear across a variety of library holdings. We've chosen to focus on materials from the eighteenth through early twentieth centuries, and have included materials created by Black people, like novels and personal letters; materials created about Black people, like abolitionist pamphlets and speeches; and materials that implicate Black life, like business documents from the slave trade and government documents from the Freedmen's Bureau. Some of the materials are joyous celebrations of Black freedom, while others are hateful representations of White resentment and fear. There are also many complicated materials that exist in the spaces between those two vantages. Everything in this collection was selected with purpose, towards the goal of providing access to primary sources that open the door to further research into Black experiences in the Americas.

Preprocessing, cleaning, and labeling

The questions in this section are intended to provide dataset consumers with sufficient information necessary to evaluate whether the "raw" data has been transformed or prepared in ways that are compatible with their chosen tasks.

Was any preprocessing, cleaning, or labeling done to the data? Eg: tokenizing, feature extraction, processing of missing values, or removal of outliers?

- The materials represented in the dataset were selected for digitization by Dorothy Berry, with input from library staff, as a curated sample of the available materials at Houghton Library. Our goal was to, as much as possible, focus on materials created by Black people and to only include materials that could be made publicly available, hence the limited timeline ending in the early twentieth century. We consulted bibliographies of early Black print by scholars like Dorothy Porter, and also roamed Houghton's stacks looking for underappreciated material. Of the 1150 items selected for digitization, 120 were bound pamphlets transferred from Widener, cataloged only by first item or by library created spine title, i.e. "Anti-Slavery Pamphlets." As part of this project, they were cataloged by Megan McMniff. Seven archival collections were also digitized in full and their descriptions were reviewed for language and accuracy and updated accordingly.
- To create the datasets, digitized items were divided by format (pamphlet, correspondence, etc.) and dublin core metadata was added to each item, in addition to

tag's relevant to the goals of the project such as "Women's Voices," "Black voices," and "Memoirs and Slave Narratives."

Is the raw data included in this dataset or otherwise preserved and accessible?

- Yes, the datasets contain the raw data and processed data.

Is the software used to preprocess the data available? Include a link to the access point.

- The Jupyter Notebook used to process manuscript OCR data can be found here: https://github.com/Cook4986/SAEF_OCR

Uses

The following questions are intended to encourage dataset creators to reflect on the tasks for which the dataset should and should not be used. By explicitly highlighting these tasks, dataset creators can help dataset consumers make informed decisions and avoid potential risks or harms.

What (other) tasks could the dataset be used for?

- A sub-goal of the larger AOK SAEF grant effort is to provide researchers with usable OCR from our digitized collections material, to open possibilities for the full-text analysis of the collection at scale. While the bulk of the materials to be digitized are printed 19th century documents, which can be transcribed quite accurately by current OCR software without human oversight, there are many key documents that are resistant to traditional technologies-- particularly those documents comprising hand-written material.
- This collection opens up text mining possibilities and comparative image searches that expand our understanding of the African American freedom struggle.

Is there anything about the composition of the dataset or the way it was collected and preprocessed that might impact future uses?

- In some cases, OCR outputs (annotations, plaintext transcriptions) are incomplete and/or inaccurate. So, researchers should consult json files associated with each image, to determine whether the confidence value associated with a given transcription meets the standards of their individual project and to deploy manual quality control processes that may correct transcriptions as necessary.

Distribution

Dataset creators should provide answers to these questions prior to distributing the dataset either internally or externally.

Will the dataset be distributed to third parties or made available outside of the creating entity?

- These datasets are available electronically from the Harvard Dataverse Repository (https://dataverse.harvard.edu).

How will the dataset be distributed?

- Each dataset corresponds to a digital object and has a DOI and a recommended data citation.

Will the dataset be distributed under a copyright or other intellectual property (IP) license and/or under applicable Terms of Use (ToU)?

- Please see the **License/Data Use Agreement** information on the collection home page: https://dataverse.harvard.edu/dataverse/SAEF

Maintenance & Stewardship

The questions in this section are intended to encourage dataset creators to plan for dataset maintenance and communicate this plan to consumers.

Who will be supporting, hosting, and maintaining this dataset?

- This data collection is maintained by Houghton Library and Harvard Library.

How can the owner, curator, or manager of the dataset be contacted?

- Please see the **Contact** button on the dataverse collection home page: https://dataverse.harvard.edu/dataverse/SAEF

Will the dataset be updated?

- New dataset versions will be created when necessary to either add new content or to address errors identified by collection users and Harvard Library and Houghton Library staff.

Will older versions of the dataset continue to be supported, hosted, or maintained?

- Prior versions of data collection contents will be available through each dataset's Versions tab.

If others want to extend, augment, build on, or contribute to the dataset, is there a mechanism to do so?

- Researchers who use and create new datasets based on this collection are encouraged to deposit their content in the Harvard Dataverse.

References

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. “Datasheets for Datasets.” *Communications of the ACM* 64, no. 12 (December 2021): 86–92. https://doi.org/10.1145/3458723.