Selected Digitized Books Dataset Data Processing Plan

Section A: General

A1: Goals of experiment

The goal of providing access to this data in the context of a data package is to make available the entirety of the collection that is currently accessible to the public, accessible in a format that is both comprehensive and more easily digestible for both computational and traditional research purposes. Note that the source collection Selected Digitized Books continues to develop and can be found on loc.gov.

A2: Describe the scope of the intended workflow or pipeline.

The text data available in this dataset was created from the images of the Selected Digitzed Books using optical character recognition (OCR) technologies. This dataset focuses on full text from 90,414 selected digitized books. The original query that was used against the Library of Congress's API to generate the data is listed in the Compilation Methods section further down in the document.

A3: Data delivery format and specifications for data generated in the experiment.

A folder containing *166,218* .txt and JSON files containing full text from *90,414* selected digitized books

  - As of 2022-09-27 there are 1,894 items from the metadata that do not have a corresponding full text extract. These will be updated as they are made available.

- metadata.json : a JSON file containing the metadata for all *90,414* selected digitized books
- metadata.csv: a CSV transformation of the original JSON metadata
- manifest.txt : a text file listing the image id, MD5 hash, and location of the images in the data set
- README.md : technical overview of how the data set was created

- Data coversheet: a more substantive overview of the data and the collection from which it is derived
- sample data: 1,000 randomly selected items from the *90,414* set and their corresponding full text extracts have been provided as sample data. Included with this are a `metadata.csv`, `metadata.json`, and `manifest.txt`.

A4: Description of intended use.

The data package will be made publicly available through a S3/Cloudfront distribution on data.labs.loc.gov.

Section B: Data Doumentation

B1: Description of dataset.

Title of dataset

Selected Digitized Books

Composition

A folder containing *166,218* .txt and JSON files containing full text from *90,414* selected digitized books

  - As of 2022-09-27 there are 1,894 items from the metadata that do not have a corresponding full text extract. These will be updated as they are made available.

- metadata.json : a JSON file containing the metadata for all *90,414* selected digitized books
- metadata.csv: a CSV transformation of the original JSON metadata
- manifest.txt : a text file listing the image id, MD5 hash, and location of the images in the data set
- README.md : technical overview of how the data set was created
- Data coversheet: a more substantive overview of the data and the collection from which it is derived
- sample data: 1,000 randomly selected items from the *90,414* set and their corresponding full text extracts have been provided as sample data. Included with this are a `metadata.csv`, `metadata.json`, and `manifest.txt`.

Provenance

The full text in this dataset is derived from the Selected Digitized Books collection on loc.gov. This page provides contextual information to situate the images contained in the dataset in relation to the source material presented on loc.gov.

Compilation methods

This dataset was created using the LOC JSON/YAML API to fetch the metadata and an internal workflow processing and data management application to pull the associated full text from an LCCN. The metadata comprises all of the selected digitized books (as of 2022-08-26) that had a date associated with it. The LOC API has a maximum result of 100,000 objects. As of 2022-08-26, there were over 118,000 selected digitized books. However, only 90,414 had a date associated with it. So in ordered to get around the API's maximum result limitation, only items with a date were gathered for this initial release. In future updates to the dataset, the additional items will be added. The two queries that were used to gather this initial data were:

Dates before 1900:
https://www.loc.gov/collections/selected-digitized-books/?c=150&dates=1000/1899&fa=access-restricted

Dates from 1900 on:
https://www.loc.gov/collections/selected-digitized-books/?c=150&dates=1900/2099&fa=access-restricted

As noted above, 1,894 items from the 90,414 items in the metadata do not currently have an extract associated with them. This set will be updated when those become available.

Preprocessing steps

No preprocessing was done in order to create the dataset. All work that was done, was completed and is described in the Compilation Methods section above.

Potential risk to people, communities, and organizations and strategies for risk mitigation.

The data package provides the following Content Advisory:

Please note that terminology in historical materials and in Library descriptions does not always match the language preferred by members of the communities depicted, and may include negative stereotypes or words that offend.

For questions or more information about this material, please contact Prints and Photographs Division staff through the Ask a Librarian service.

How will the experiment team address outdated or potentially offensive terms or elements of data that may be harmful if encountered by human users?

Please refer to the Content Advisory in the previous section.

Copyright, licensing, rights, and/or privacy restrictions

The books in this collection are in the public domain and are free to use and reuse.

Credit Line: Library of Congress

More about Copyright and other Restrictions.

For guidance about compiling full citations consult Citing Primary Sources.