Stereograph Cards 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 accessible to the public, accessible in a format that is both comprehensive and more easily digestible for both computational and traditional research purposes.

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

While the full collection contains approximately 52,000 stereographs, this dataset focuses on rights free stereographs from the 1800s through 1924. The original query that was used against the Library of Congress's API to generate the data is listed Compilation Methods section further down in the document.

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

- 39,597 rights free stereograph image (JPG) files from the 1800s through 1924 pulled from the Prints and Photographs Stereograph Cards Collection. The images were the medium sized image as determined by the data available in the API response. These images either had a width of 640 or a height of 640. More technical explanations can be found in the Compilation Methods section.
- metadata.json: a JSON file containing the metadata for all 39,597 images in the dataset
- 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 dataset.

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

Stereographs

Composition

- 39,597 rights free stereograph image (JPG) files from the 1800s through 1924 pulled from the Prints and Photographs Stereograph Cards Collection. The images were the medium sized image as determined by the data available in the API response. These images either had a width of 640 or a height of 640. More technical explanations can be found in the Compilation Methods section.
- metadata.json: a JSON file containing the metadata for all 39,597 images in the dataset
- 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 dataset.

Provenance

The image files in this dataset are derived from the Stereograph Cards digital collection on loc.gov (https://www.loc.gov/collections/stereograph-cards/about-this-collection/). 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 and comprises a scoped portion of the stereographs and not every item in the collection. Subject matter experts were consulted in the creation of a JSON API query (https://www.loc.gov/collections/stereograph-cards?dates=1800/1924&fa=access-restricted:false&q=no%
to produce rights free stereographs from the 1850s through 1924. This original query returned 44,694 results.

The JSON results were then placed in a new JSON structure of key/value pairs where the key is the LCCN and the value its corresponding entry in the content.results section of the original LOC API query.

The images chosen for the dataset were those in the medium range. First, if applicable, the item.service*medium field was used. If this field was not available, an image with a width of 640px was selected from the image*url field. Finally, if there were no images with a width of 640px, then one with a height of 640px was used.

The final dataset, after filtering out duplicates and items that had no images available, is comprised of 39,597 items.

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 Library's stereograph collection includes thousands of photographs submitted for copyright protection. Those that were copyrighted or published in the U.S. more than 95 years ago are in the public domain because the copyright has expired. The collection also includes images not submitted for copyright that may not have been published. The term of copyright for unpublished images is the life of the creator plus 70 years. For unpublished anonymous works and works where the death date of the creator is not known, the copyright term is 120 years from the date of creation.