Selected Digitized Books Data Package README

Version information

Version 1.1 Last updated 2024-04-17.

- *1.1 (2024-04-17)* LC Labs consolidated coversheet and README content with minor formatting updates
- *1.0 (2022-09-27)* First version

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.

About the source data or collection

Brief description & background of collection

This is a growing collection of selected books and other materials from the Library of Congress General Collections that can be made openly available. Most of the materials in this collection were published in the United States and are in English. The collection features tens of thousands of works of nonfiction, fiction, and poetry. These works cover a wide range of subjects and topics including: American history, travel, sports, cooking, agriculture, children's literature, philsophy, government publications including speeches and addresses, local history and geneology, film, and many more esoteric subjects from beekeeping to spiritualism. There are also some materials in foreign languages that were published in other countries. The materials in this collection can be read online or downloaded.

Original format

Printed books

Library of Congress reading room

Main Reading Room, https://www.loc.gov/rr/main/

Contact

For more information please contact the specialists in the Library's Main Reading Room at https://ask.loc.gov/history-humanities-social-sciences/.

Metadata type

MARC

Scale of description

Books in this collection have been digitized and described individually (at the item level). In some cases, multi-part works are represented including multi-volume books or serial publications, where one descriptive record relates several digitized resources.

Rights information

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.

Digitization information

This broad collection contains materials from a variety of digitization sources. Please direct questions about specific items to the specialists in the Main Reading rooom.

Digitization was completed over more than a decade, representing multiple phases of priorities for what to select to digitize from the General Collections. Priority was often driven by identifying, where possible, works that were not already available online from other sources, and that may not be held by other libraries. Additionally, selection was driven by subject classification (genealogy, local history, other subjects) or to assist with support for moving collections to remote storage. For several years there was an emphasis on children’s and young adult literature (class PZ). As such, the collection is not a random sample of what was published in the US from the Library's General Collections, but represents the aggregation of several areas of digitization and preservation priorities and provides a unique cross section of the holdings of the Library.

About this exploratory data package

Selected Digitized Books consists of books and other materials from the Library of Congress General Collections that can be made openly available. Most of the materials in this collection were published in the United States and are in English. The collection features thousands of works of fiction, including books intended for children, young adults, and other audiences. There are also some materials in foreign languages that were published in other countries. The materials in this collection can be read online or downloaded.

Please note: the Selected Digitized Books digital collection continues to grow. Therefore, the text in this dataset may not constitute the entirety of text that could be derived from what may be available on loc.gov.

What's included?

The data package contains:

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` : An overview of the source data or collection provenance, the contents of the data package, and how the data package was created. Available as .md, .html, and .pdf.
- 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`.

Computational readiness and possible uses

The text data available in this dataset was created from the images of the Selected Digitzed Books using optical character recognition (OCR) technologies. The corpus is quite amenable to computational text analysis methods including but not limited to keyword analysis, named entity recognition, sentiment analysis, and topic modeling.

How was it created?

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

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.

Dataset field descriptions

The data fields that follow are directly translated from the metadata.json file. The JSON file is highly nested in nature, and that nested structure is not strictly carried over into the CSV. The CSV data fields contain the top level keys and, where applicable, one nested level below. In these cases, the field names are signified by the top level key.secondary key; for example: item.call_number.

All values in each column are stored as they would be found in the JSON metadata. Meaning, that if the column's value is a list or array, it is stored as a string representation of that value. For example: the aka field's value is in list format: ['http://www.loc.gov/item/2015651359/', 'http://www.loc.gov/pictures/item/2015651359/', 'http://www.loc.gov/pictures/collection/stereo/item/2015651359/', 'http://hdl.loc.gov/loc.pnp/stereo.1s04563', 'http://hdl.loc.gov/loc.pnp/stereo.2s04563', 'http://www.loc.gov/resource/stereo.1s04563/', 'http://www.loc.gov/resource/stereo.2s04563/', 'http://lccn.loc.gov/2015651359']

Each of the fields described below appears for a result under the `content.results` section of the API response for this query. Please note that not all elements appear for each result. Elements appearing in only some results have been marked with an asterisk.

- access-restricted: used for limiting access to some items from offsite
- aka: alternative identifiers for documents (e.g., shortcut urls)
- campaigns: added at display time if a search result needs campaign codes added to the published url

- contributor: generally used for name/related names such as photographer, author, etc. but is distinct from names in subject field. For these images, this fields usually cites the photographer or copyright claimant
- coordinates: a text version of geographical coordinates; typically drawn from a MARC field
- date: a date chosen from all available dates to be the sortable date. Can be the creation date, publication date, or a date referenced in the item depending on website target and cataloging; entries in this field could be formatted as a year or YYYY-MM-DD. Items are sortable by this date.
- dates: List of dates related to the item. In ISO 8601 format, UTC. Items are facetable by these dates.
- description: often includes a short, summary description of the original physical item written to accompany the item in a list of search results; for these images, descriptions can be created from MARC records content and/or pulled from the Selected Digitized Books Collection
- digitized: true/false; indicates whether a digital surrogate exists
- extract_timestamp: timestamp of most recent ETL process
- group: ETL group
- hassegments: true/false for having segmented data (pages, bounding boxes of images, audio segmentation, etc.) in the index
- id: HTTP version of the URL for the item, including its identifier. Always appears.
- image-url: URLs for images in various sizes, if available. If the item is not something that has an image (e.g. it’s a book that’s not digitized or an exhibit), the URL for the image might be for an icon image file. The url is meant to convey enough information to be used for various result displays (i.e. list, grid, gallery, slideshow) described in more detail in the 'resources' element.
- index: the index number of the results among all results. This starts with 1 and continues through all of the results in the whole set (not just this page).
- language: languages associated with the item
- location_city: field for cities related to the item
- location_country: field for counties related to the item
- location_state: field for states related to the item
- mime_type: the formats available for a digitized item
- number: Swiss-Army knife for "numbers" - e.g., OCLC number, shelf number, etc.
- online_format: text name of the online format, usually derived from mime type; superset of mime-type(s)
- original_format: The kind of object being described (not the digitized version). If the record is for an entire collection, that is included here.
- other_title: alternative language titles, other alternative titles
- partof: Collections, divisions, units in the Library of Congress, or any of a number of less formal groupings and subgroupings used for organizing content.

- resources: details of digital content and structure
- shelf_id: Primary sorting field of item records
- site: Originally showed the source system, used in transform for display methods, tracks ETL target closely
- timestamp: time record inserted in the web index
- title: title of the item
- url: URL on the loc.gov website. If the items is something in the library catalog, the URL will start with lccn.loc.gov.
- subject: list of subjects. These are separated elements of the Library of Congress Subject Headings. Geography is not shown here, see the location element.
- type: medium of the original item
- number_lccn
- number*source*modified
- number_oclc
- number*preceding*items
- number*succeeding*items
- number*former*id
- number_issn
- number*carrier*type
- segments
- partof_title
- publication_frequency
- composite_location

The following `item` subfields of the `content.results` section are mainly for display of the item on the loc.gov website. These subfields may pull information from target-specific interpretations of MARC records.

- item.call_number
- item.contributors
- item.created_published
- item.date
- item.format
- item.language
- item.medium
- item.notes
- item.title
- item.digital_id
- item.subjects
- location

- item.location
- item.creator
- item.contents
- item.other_title
- item.genre
- item.contributor_names
- item.control_number
- item.created
- item.creators
- item.formats
- item.id
- item.link
- item.marc
- item.medium_brief
- item.mediums
- item.modified
- item.other*control*numbers
- item.place
- item.raw_collections
- item.service_low
- item.service_medium
- item.sort_date
- item.source_created
- item.source_modified
- item.subject_headings
- item.thumb_gallery
- item.summary
- item.number*former*id
- item.created*published*date
- item.stmt*of*responsibility
- item.related_items
- item.source_collection
- item.part_of
- item.repository
- item.rights_advisory
- item.access_advisory
- item.display_offsite: similar to `access_restricted`; used for limiting access to some items from offsite
- item.reproduction_number

- item.resource_links

Rights Statement

Creator and contributor information

Creator: Chase Dooley

Contributors: Eileen J. Manchester, Meghan Ferriter, Mark Cooper

Contact information

Please contact LC-Labs@loc.gov with any questions or suggestions!