Geographic Locations in English-Language Literature, 1701-2011

Contents

Dataset	Volume Count
volumemeta_geo.tsv	205,704
recordmeta_geo.tsv	173,302
titlemeta_geo.tsv	135,365

Often researchers are interested in how ideas of place are represented in works of fiction. However, this can be a difficult task for many reasons. Some researchers have turned to computational methods in order to make the process easier, but this comes with its own challenges. Through his Textual Geographies project, Matthew Wilkens alleviates some of those difficulties by identifying all the geographical locations mentioned in works of English-language fiction from 1701-2011 found in the HathiTrust Digital Library. It includes the latitude and longitude of those locations, as well as the number of times the location is mentioned in the volume. This provides researchers with usually hard-to-generate geospatial data, with a good portion of the computational work already done on their behalf.

Attribution

Matthew Wilkens and Guangchen Ruan (2020). “Geographic Locations in English-Language Literature, 1701-2011 (1.0) [Dataset].” HathiTrust Research Center. https://doi.org/10.13012/2K5C-RF13.

Funding

National Endowment for the Humanities, Office of Digital Humanities, award number HK-250673-16

Methods

Locations were identified in the source texts via the Stanford NLP group's CFR-based Named Entity Recognition (NER) tagger (https://nlp.stanford.edu/software/CRF-NER.html). Strings that were tagged as places more than about 400 times in the full HathiTrust English-language dataset (7.6M volumes, not limited to fiction) were then associated with detailed geographic information via Google's Places and Geocoding APIs. The results were reviewed by hand for locations that occurred with high frequency in the fiction dataset; any location accounting for more than about 0.1% of all location mentions in English-language fiction has been checked by hand. These high-frequency locations together account for 60% of all location mentions. The overall F1 score is slightly above 0.8.

This dataset corresponds to a corpus defined in Underwood et al.'s NovelTM English-language fiction metadata. Details about these corpora and how they were selected can be found in a report for the NovelTM dataset.

Currently, HTRC has made three variations of the dataset available: volumemeta, recordmeta, and titlemeta, each containing 130,000+ volumes. These are the latest evolution of the work that originally began as the Word Frequencies in English-Language Literature, 1700-1922 dataset, developed in collaboration with HTRC, and developed into NovelTM Datasets for English-Language Fiction, 1700-2009 when more volumes were analyzed. These were deemed as the best targets for integration, since they are HTRC’s best known estimate for what constitutes "all the English-language fiction in HathiTrust."

The datasets include geographic data for nearly all the volumes in each of the NovelTM corpora. There is missing geo data for a little less than 2% of the NovelTM volumes. Some of this gap is because the Textual Geographies data was extracted earlier than the NovelTM work, so a subset of volumes had not been included in the HathiTrust at that point, and some of it is because there are volumes that don't have any locations in them. This leaves better than 98% coverage, but not 100%.

The files here are much larger than the metadata records in the NovelTM English-language fiction dataset, both because there is additional geographic data and because there are multiple entries for each volume as there is one entry for each unique location in each volume. The largest of these files (volumemeta_geo.tsv.gz) is over 18M lines long. It would be possible to minimize size at the cost of increased query complexity by separating the bibliographic data from the geographic data, linking them via a third table of identifiers. However, the result is a three-way join whenever one needs to query the data. Performing the join on the front end and producing and then making the tables available is easier.

The three files are tab-separated, minimally escaped via Pandas, and gzipped. They're UTF-8 encoded. Below is what is contained in each file:

‘volumemeta’

This data includes all the volumes found and identified as fiction in Ted Underwood’s NovelTM dataset: 205,000+ volumes between 1701 and 2011. It includes many duplicates, such as multiple editions/printings of the same title, as well as multiple copies of each printing. The outer boundary of Underwood’s list was shaped by probabilistic models that identified fiction and attempted to filter out other genres (error enters at this step of the process).

‘recordmeta’

This data tries to exclude duplicate copies of the same printing, using HathiTrust “record ids” and “volume numbers” to identify duplicate copies. At this level of deduplication, there are 173,000+ records.

‘titlemeta’

This data tries to identify one copy of each fiction "title"—by preference the earliest copy available in HathiTrust. In other words, different editions of a novel, possibly with different prefatory material or even different language used in the title text itself, will usually be collapsed into a single title. This level of deduplication produces a list of 135,000+ distinct titles. To identify different records as examples of “the same title” Underwood used a probabilistic model, which again, introduces a source of error.

Data columns

In addition, each of the meta files has 37 columns. They are as follows:

'docid'

The unique volume identifier used by Underwood et al. matches the HathiTrust Volume ID (HTID), except that identifiers have been made filename-safe by substituting '+' for ':' and '=' for '/'.

'htid'

Volume HTID. Substantially redundant with 'docid' except as noted above.

'author'

A short, cleaned version of the volume author, copied from Underwood et al.

'shorttitle'

A short, cleaned version of the volume title, copied from Underwood et al.

'inferreddate'

A best guess of the volume's publication date, copied from Underwood et al.

'geographics'

MARC entries indicating the geographic subject matter of the text, copied from Underwood et al. Coverage is uneven (about 10% of volumes).

'text_string'

A normalized version of the unique string identified as a location in a text. Original strings have been lowercased, unaccented, have had punctuation replaced with spaces, and have been whitespace regularized. For example, 'London' becomes 'london'; 'Île-de-France' becomes 'ile de france'; and ' New York ' becomes 'new york'.

'occurs'

Integer count of the number of times a given normalized string occurs in the volume.

'publication_country'

The two-letter country code representing the publication location of the volume, where known. This can be useful as a proxy for the national origin of the volume.

'formatted_address'

A standard, long-form representation of the geographic location to which the text_string corresponds. Example: 'New York, NY, USA'.

'location_type'

The type of location. Examples include 'country', 'administrative_area_level_1' (like a US state, Canadian province, or 'England' in the UK), 'locality', 'church', 'bar', and 'route'. There are 118 distinct values.

'country_long'

Long-form name of the country within which the location is found.

'country_short'

Short-form (two letter) version of the country name where the location is found.

'admin_1_long'

Long-form version of the top-level administrative area to which the location belongs. Corresponds to US states, French regions, etc.

'admin_1_short'

Short-form version of the admin_1 label. Two letters for US states, but can be of arbitrary length.

'admin_2'

Second-level administrative area. For example, US counties.

'admin_3' | 'admin_4' | 'admin_5'

Third-, fourth-, and fifth-level administrative areas.

'locality'

City, town, village, etc.

'sublocality_1'

First-level administrative subunit of a locality. The boroughs of New York City correspond to this level.

'sublocality_2' | 'sublocality_3'

Second- and third-level administrative subunits of a locality.

'neighborhood'

Possibly colloquial, does not necessarily correspond to an administrative district.

'premise'

A named building or similar.

'subpremise'

Part of a premise.

'street_number'

The number portion of a street address. May contain non-numeric data.

'street_address'

Full street address of buildings and similarly addressable point locations. Lightly used. Not to be confused with 'formatted_address', which is always present.

'route'

Name of the road for road-like locations.

'post_code'

Alphanumeric postal code of the location, if applicable.

'natural_feature'

The name of lakes, rivers, mountain ranges, continents, etc. Can be especially useful for locations that span administrative boundaries, e.g., 'Atlantic Ocean'.

'point_of_interest'

Inconsistently applied label by Google for locations of some sort of interest. PI deemed it unreliable for general use, but it was left in as it might be of use in rare cases.

'colloquial_area'

Unofficial names for locations and regions, including those that do not correspond to any existing administrative unit. Examples: 'Central Africa', 'Caribbean'.

'continent'

The name of a continent, in cases of direct reference to the continent itself. Unfortunately, while most other fields are supplied hierarchically (that is, locations at lower administrative levels include data for the higher-level administrative areas to which they belong), this field is not. If one needs to aggregate locations by continent, one must supply an independent mapping between countries (or other location types) and continents.

'other'

A lightly used catch-all for unclassifiable location types.

'lat' | 'lon'

The specific latitude and longitude of the locations. For non-point locations, this is a single point near the geographic center of the area.

Download the dataset

You can download the dataset by clicking the link below for the version of the dataset you want. The options are “volumemeta”, “recordmeta”, or “titlemeta”. You can also download just the unique volume ids with no metadata if that is more helpful. To download just a list of the HTRC volume ids for each version of the data click on the corresponding link below. Your options are “volumeids”, “recordids”, and “titleids”.

In addition, you can use rsync to download the different dataset versions. If you have the Windows operating system you may need to download and install rsync. If you are using MacOS or Linux, rsync should already be installed. To download via rsync you will use the command line. If you are not comfortable with the command line, then use one of the links under “Downloads”. Below is a list of the rsync commands to download each version as well as the lists of volume ids:

NOTE: The final . is necessary and indicates the destination location where the file should be transferred (. represents "current directory" in UNIX)

Rsync dataset versions

rsync data.htrc.illinois.edu::textual-geographies/volumemeta_geo.tsv.gz .

rsync data.htrc.illinois.edu::textual-geographies/recordmeta_geo.tsv.gz .

rsync data.htrc.illinois.edu::textual-geographies/titlemeta_geo.tsv.gz .

Rsync ID Lists

rsync data.htrc.illinois.edu::textual-geographies/volumemeta_geo.tsv.ids .

rsync data.htrc.illinois.edu::textual-geographies/recordmeta_geo.tsv.ids .

rsync data.htrc.illinois.edu::textual-geographies/titlemeta_geo.tsv.ids .

Downloads

Datasets

volumemeta

recordmeta

titlemeta

ID Lists

volumeids

recordids

titleids

Explore the dataset

A Google Colab notebook (it is a .ipynb file, so basically a Jupyter Notebook run in Google Colab) has been created to assist with searching and manipulating the dataset. This notebook allows you to search the dataset by volume id or terms of interest and includes step by step instructions and explanations. The “Dataset search notebook” link below will take you to the Google Colab notebook.

Use Case

An example use case has also been created and made available as a Google Colab notebook. The notebook contains step by step instructions and explanations. You can access the use case by clicking on the “Use case” link below.

Google Colab Notebooks

NOTE: You will need to sign in to a Google account in order to be able to interact with the notebook, however, you can view it without one. Additionally, the notebooks are in "view" mode which prohibits any edits from being saved. You can make edits and run the notebook with those edits (only if you are signed in to a Google account), but when you close the notebook and come back, those edits will be gone. There are more instructions in the notebooks themselves on how to save a copy of the notebooks so you can make and save changes.

Dataset search notebook

Use case