Sanborn Maps 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 (2023-11-28)* First version

Content Advisory

The Sanborn maps were produced as a commercial product and sold by the Sanborn Map Company, who marketed them to fire insurance agents. Sanborn maps were created in the field by the Sanborn Map Company surveyors, rarely members of the communities that they mapped. Although intended to be technical and objective, the surveys may include language and descriptions that did not necessarily reflect the people, homes, businesses, recreational areas and civic spaces in the communities they mapped. Ethnic descriptions, in particular, may be offensive.

The Sanborn maps used English-language labels and descriptors, regardless of local language. Business or property owner names, for example, were generally translated to English.

Data Package Summary

This dataset was created as part of an LC Labs experiment in collaboration with AVP to understand the benefits, risks, quality benchmarks, workflows, compilation methods, transformations, and documentation practices required to assemble datasets for public use in the cloud. The dataset was completed with support from the Geography and Map Division.

The target audiences of this dataset are users who want to explore spatial or temporal aspects of a collection, plot data on a map or timeline, or navigate or explore data by time or place.

The dataset contains metadata records for 50,600 maps from the Sanborn Fire Insurance Maps collection. These records are included in CSV (70MB) and JSON (134MB) formats.

The dataset is organized by atlas, with each row (CSV) or JSON object representing a single atlas from the collection. Each atlas may represent one or more locations, and is geocoded to a single primary location, usually at the city level.

About the source data or collection

Brief description & background of source material

Original production by Sanborn Map Company

The Sanborn map collection consists of a uniform series of large-scale, building level maps, dating from 1867 to 1981, which are organized into over 9,800 cities in the United States as well as Canada, Mexico, and Cuba. Some atlases cover portions of neighboring towns, and thus the collection as a whole covers the commercial, industrial, and/or residential sections of some twelve thousand cities. The atlases were designed to assist fire insurance agents in determining the degree of hazard associated with a particular property and therefore show the size, shape, and construction of dwellings, commercial buildings, and factories as well as fire walls, locations of windows and doors, sprinkler systems, types of roofs, etc. . The maps also indicate widths and names of streets, property boundaries, building use, and house and block numbers. They show the locations of water mains, giving their dimensions, and of fire alarm boxes and hydrants. Sanborn maps are an unrivaled source of information for their time about the structure and use of buildings in U.S. cities.

Collection by Library of Congress

The Sanborn collection at Library of Congress includes over fifty thousand editions of fire insurance maps comprising almost seven hundred thousand individual sheets. The Library of Congress holdings represent the largest extant collection of maps produced by the Sanborn Map Company. The majority of the maps were acquired through copyright deposit, but the collection was substantially enriched in 1967 when the Bureau of the Census transferred its sizeable collection to the Library. These additional maps had been used in the field, often for decades, and include updates in the form of printed, paste-on corrections from the Sanborn Map Company. Smaller numbers of Sanborn atlases have been acquired through purchase or donation.

Digitization by Library of Congress

Most of the public domain portion of the collection is digitized and freely available online, including atlases published before 1923 and atlases for which copyright was not renewed. Some gaps may exist, due either to error or to lack of clarity in copyright renewal. Most maps which have entered the public domain since 2019 (those published between 1923 and 1928) have not been digitized, with the exception of a small number. In total, approximately 69% of the atlases in the collection are digitized and freely available online from the Library of Congress.

Creation of metadata by Library of Congress

The source data comes from the online Sanborn Maps Collection. The collection uses a local metadata schema initially developed for the 1981 publication Fire insurance maps in the Library of Congress : plans of North American cities and towns produced by the Sanborn Map Company. See section **e. Metadata type** for more information about the source collection's metadata schema.

A small portion of the Library of Congress's Sanborn collection, generally received by the library after 1981, was not yet described or digitized at the time this dataset was created and is thus not included in the dataset.

Note on representativeness of collection

Care should also be taken to understand potential sampling biases when treating the collection as a dataset. Although large, the data contained within the Sanborn collection is not an exhaustive representation of U.S. towns and cities. There is no complete record of all Sanborn maps produced, particularly the field editions that were updated over time with paste-in corrections, and so it is unknown how many maps may be missing from the Library of Congress collection. Not all cities were surveyed by the Sanborn Map Company, and not all cities were surveyed completely. Neighborhoods or certain areas of cities may be missing from the maps. Areas previously covered in a given city may be absent from subsequent editions. Coverage and selection of cities may have been determined based on commercial fire insurance markets, convenience, level of familiarity with local areas, and any number of other factors. Data may also be more complete or accurate in certain areas than in other areas.

Data on the maps may be inaccurate or misleading. For example, the sequential numbering system on Sanborn map sheets were arbitrary numbers given to the lots by the surveyor when a town did not yet have actual street address numbers. When this is the case, usually on the title page you will see the phrase: "Alternate street nos are actual, consecutive street nos are arbitrary." Labels may also contain offensive or derogatory terms that do not accurately represent the structures or communities described.

Because a small number of later additions to the collection were not yet included in the collection inventory at the time of this dataset's creation, the dataset is not wholly representative of all Sanborn maps at the Library of Congress. Placenames do not use an authoritative taxonomy (see section **e. Metadata type**).

Original format

Printed sheets, chiefly hand colored. Mix of bound and unbound sheets, generally 64 x 54 cm or smaller. Atlases that contain paste-in corrections are indicated where publication date includes a date range (e.g., May 1925 - Sep 1948).

Library of Congress reading room

Geography and Map Reading Room

Contact

For questions or more information about this material, please contact LC-Labs@loc.gov or the Geography and Map Division at maps@loc.gov.

Metadata type

The source metadata for this collection follows a local, collection-specific schema that is non-MARC. The schema is primarily based on the 1981 index Fire insurance maps in the Library of Congress : plans of North American cities and towns produced by the Sanborn Map Company. See IV. Dataset field descriptions for more details.

Location names do not use controlled vocabularies. Instead, City and County values use the placenames as they appeared on the original Sanborn atlase title pages, with the following guidelines and exceptions: 1. Sanborn usually titled each atlas with a single city, county, and U.S. state. Roughly 10% of the atlases in the collection also contain subtitles with secondary places, such as neighboring towns and subdivisions. These subtitle locations are transcribed as free text secondary locations, and appear in the source loc.gov JSON field `secondary_location` and is repeated in the `location` field alongside the primary city, county, state, and country. In the current data package, the secondary locations appear in the `Location_secondary_text` field. 2. In some cases, multiple counties or cities were listed in the title of the original publication. In these cases, the full list appears in the source JSON fields `location_city` or `location_county` field, which map to this data package's `City_text` and `County_text` fields. For example: Kansas City, Missouri ("Jackson, Clay, and Platte Counties"); Chatfield, Minnesota ("Fillmore and Olmstead Counties"); "Harrison and East Newark and Kearny", New Jersey; and "New Jersey Coast", New Jersey. 3. If the name of a given county or city changed over time, the place name is standardized across all atlases. The name used is the name on the most recent atlas, with the following exceptions. When the 1981 index was compiled, if that place name had been further updated, the compilers usually updated the place name to the name in use at the time of the index's publication. Changes in place names after that time are generally not reflected in the metadata, except in cases where an item was first acquired after the 1981 index and covers a location not previously covered by any other items in the collection. 4. Abbreviations and acronyms are usually spelled out (e.g., Saint Paul, Minnesota), and typos in the original publications are generally corrected. 5. Where there is no county (or comparable administrative unit), the county will generally appear as "Independent Cities", as in the case of Alexandria, Virginia. 6. Counties also include a range of comparable administrative entities, such as parishes in Louisiana and Census Divisions in Alaska. States may also include comparable administrative entities, such as Canadian provinces or federal districts.

Scale of description

Items are described at the atlas level. An atlas may be anywhere from one sheet to hundreds of sheets. In multi-volume editions for larger cities, each volume is considered one atlas.

Rights information

The content of the Library of Congress online Sanborn Maps Collection is in the public domain and is free to use and reuse. For more information, see https://www.loc.gov/collections/sanborn-maps/about-this-collection/rights-and-access/

About this exploratory data package

What's included?

The data package contains: - *README*: overview of the source material and how the dataset was created and the context in which it was created. Available as in .md, .pdf, and .html formats. - *metadata.json* : a JSON file containing the metadata for all 50,600 atlases - *metadata.jsonl* : a JSON lines version of the JSON data, with one record per line, useful for processing large files - *metadata.csv* : a CSV transformation of the original JSON metadata - *maps-by-state/* : A directory of maps in .jpg format organized by state or region. Each state or region is also available as a .zip file - *summary/:* directory of summary data (.csv) of the fields included in the dataset and number of records populated for each, and location distribution of dataset records (260KB) - *sample-data/:* 100 randomly selected items from the 50,600 item set has been provided as sample data. Included with this are a `metadata.csv`, `metadata.json`, and `metadata.jsonl`. (493KB)

Composition

The data package contains metadata records from 50,600 digital objects.

Metadata for these items is in a single CSV, for all 50,600 items in the collection.

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

Data on the original Sanborn maps may be incomplete or inaccurate. Although intended to be technical and objective, the surveys may include language and descriptions that did not necessarily reflect the people, homes, businesses, recreational areas and civic spaces in the communities they mapped. Ethnic descriptions, in particular, may be offensive.

Geocoded coordinates may contain errors. See **How was it created?** for details on the enrichment stage.

Provenance

The source data comes from the online Sanborn Maps Collection. The CSV metadata was compiled by the Geography and Map Division in January 2024. More details on the layered provenance of this data package can be found in sections I and III.

Computational readiness and possible uses

The data in this dataset have been selected, structured, standardized, and enriched to make the dataset more easily comprehensible and computable through a range of methods and in a variety of environments. Enrichment of locations with coordinates and structured addresses could support plotting records in mapping interfaces. Standardization of dates could enable sequencing in timelines.

How was it created?

This dataset was created through a four-stage process including data extraction, mapping and standardization to a specified schema, enrichment of certain fields with additional data, and packaging for access and use.

Preprocessing steps

This dataset was created from the online Sanborn Maps Collection, which was not altered for the purposes of this experiment.

Compilation Methods

Extraction

This dataset was created using the LOC JSON/YAML API and comprises all digitized and non-digitized map item records, retrieved through the following API query: https://www.loc.gov/collections/sanborn-maps/?fo=json. This process returned *50,600* results.

Standardization

The API fields in the response returned from the query were reviewed and selected for mapping to a schema designed for anticipated possible uses of the dataset. This schema is comprised of fields from the General Purpose schema plus additional format or collection-specific fields (see

"Section IV. Dataset field descriptions" below). In the mapping from the API response to the dataset schema, some data values were standardized for consistency across the dataset and interoperability with other datasets.

Possible standardizations may include the following methods and are listed in detail in the "Dataset field descriptions" section.
- Capitalization (method): Data value has been capitalized using title or sentence case - Fill with (string): Data value has been filled with a static string. If "empty" is present, only empty values have been filled. Otherwise, all values for that field, including mapped values, have been filled or overwritten. - Lookup (table): Data value has been looked up and replaced by a value in the specified lookup table. - Select, sum (field): A specified field in an array of objects has been selected and summed. (This is currently only in use for totalling the number of files from the API `resource` field). - Append (method): Combine multiple source fields into a single destination field.

Enrichment

After standardization, some data fields were enriched to bring additional value for potential use cases. In this dataset, locations (lists of location components-- city/town/village, county, state, country from the API `location_city`, `location_county`, `location_state`, and `location_country` field) were concatenated into *city/town/village*, *county*, *state*, *country* strings, queried in OpenStreetMap and enriched with structured location data, geocoordinates, and URLs to the structured data record in OpenStreetMap. Results were filtered to include only place types with one of the following values: "hamlet", "town", "city", "village", "county", "state", "province", "locality", "country", "suburb", "borough". If there was more than one result, the script chose the first result to encode in the output data. This approach may have produced inaccuracies in the enriched data due to idiosyncracies in recorded locations, misspellings in the original metadata, or geocoding errors from OpenStreetMap.

Packaging

After enrichment, the dataset was output in JSON, JSONL (JSON lines), and CSV formats. CSV files were flattened from JSON using the following rules: - Arrays were flattened to strings, with array items delimited by the pipe character
- Enriched locations (`Location`) are included in their original JSON form in the `Location` column. The `Location.Full_name` is included in a `Location_full_name` column. Coordinates from `Location.Coordinates` are listed as lat,long pairs in the `Coordinates` column, and split into lat and long in the `Latitude` and `Longitude` columns. `State_region` and `County` colums were parsed from the enriched `Location` and added for easier grouping and filtering of the data.

Dataset field descriptions

The data fields that follow were compiled from a "General Purpose" schema designed for the Data Transformation Services experiment and supplemented with additional fields specific to this collection and/or anticipated uses of the data. Values have been sourced from API fields or templated with static values where necessary. These mappings are indicated in the "Data source" column in the table below. Some values have been standardized for consistency across the dataset or interoperability with other datasets using similar data structures, standards, or controlled vocabularies. Types and descriptions of standardizations are listed above in the "How was it created" section and indicated in the "Standardization" column of the table below. Enrichments are also described in the "How was it created" section and are indicated below.

The data fields that follow are directly translated from the `metadata.json` file. The JSON file is nested in nature, and that nested structure is not strictly carried over into the CSV. When JSON fields have been flattened or otherwise altered to fit a CSV field, the transformation is described below.

Each of the fields described below appears for an object or row in the dataset. Please note that not all elements appear for each result. The number and percentage of results populated for each field are indicated in the table below as well as in a `summary.csv` file in this package.

Field	Datatype	Definition	Requirement	Is list	Repeatability	Data So	Standardizations
City_text	Text	Textual representation of city/town/village, copied directly from the source record	Optional	TRUE	Y	loc.gov JSON	Capital (title)
Coordinates	Text	(CSV only) The coordinate pair (lat, long) of a location matched through the OpenStreetMap location enrichment. Corresponds to the location string in the same position in the Location*full*name column.	Optional	TRUE	Y	loc.gov JSON	

Country_text	Text	Textual representation of country, copied directly from the source record	Optional	TRUE	Y	loc.gov JSON	Capital (title)
County	Text	(CSV only) County(s) of a location matched through the OpenStreetMap location enrichment.	Optional	FALSE	N	loc.gov JSON	
County_text	Text	Textual representation of county, copied directly from the source record	Optional	TRUE	Y	loc.gov JSON	Capital (title)
Date	Date (EDTF)	A structured representation of the date created.	Optional	FALSE	N	loc.gov JSON	
Digitized	Boolean	Whether or not the resource described is digitized.	Optional	FALSE	N	loc.gov JSON	
Id	Text	A unique identifier for the resource.	Required	FALSE	N	loc.gov JSON	
IIIF_manifest	URL	A IIIF manifest for the digital object, if available	Optional	FALSE	N	loc.gov JSON	
Language	Text	The language(s) of the content of the resource.	Optional	TRUE	Y	loc.gov JSON	Capital (senten

Last*updated*in_api	Timestamp	The date and time the metadata was last refreshed in the API. This may or may not reflect a change in the data.	Optional	FALSE	N	loc.gov JSON	
Latitude	Text	(CSV only) Latitude from the first coordinate in the "Coordinates" field.	Optional	FALSE	N	loc.gov JSON	
Location	Object	Structured representation of geocoded location from OpenStreetMap, including parent administrative divisions, where applicable, and geocoordinates	Optional	TRUE	Y	loc.gov JSON	
Location*full*name	Text	(CSV only) The full location string of a location matched through the OpenStreetMap location enrichment. Corresponds to the coordinate pair in the same position in the Coordinates column.	Optional	FALSE	N	loc.gov JSON	

Location*secondary*text	Text	Textual representation of secondary locations (free text), copied directly from the source record	Optional	FALSE	N	loc.gov JSON		Capital (title)
Location_temp	Text	Temporary field for storing extracted location data for enrichment stage. This is removed during the enrichment stage.	Optional	FALSE	N	loc.gov JSON		Append (from API JSON fields location *location* location *location*
Location_text	Text	Textual representation of city, county, state, country, and secondary locations, copied directly from the source record	Required	TRUE	Y	loc.gov JSON		Capital (title)
Longitude	Text	(CSV only) Longitude from the first coordinate in the "Coordinates" field.	Optional	FALSE	N	loc.gov JSON		

Mime_type	Text	The MIME type(s) of the access format files composing the digital object. Digital objects are also generally available in TIFF as a preservation format.	Optional	TRUE	Y	loc.gov JSON	
Notes	Text	Additional information about the content, context, or physical description of the resource.	Optional	TRUE	Y	loc.gov JSON	
Number*of*files	Integer	Number of files composing the digital object.	Optional	FALSE	N	loc.gov JSON	Select, sum ("files")
Online_format	Text	The format of the online version of the resource.	Optional	TRUE	Y	loc.gov JSON	
Original_format	Text	The format the resource was digitized from.	Optional	TRUE	Y	loc.gov JSON	
Part_of	Text	Groups the resource is a part of, such as source collection or repository.	Optional	TRUE	Y	loc.gov JSON	Capital (title)
Preview_url	URL	A url for a preview image or thumbnail for the digital object.	Optional	TRUE	Y	loc.gov JSON	

Repository	Text	The repository that holds the physical or digital resource.	Optional	FALSE	N	loc.gov JSON	
Source_collection	Text	The collection the resource belongs to.	Optional	TRUE	Y	loc.gov JSON	Fill with (“Sanbo Fire Insuran Map Collecti empty)
State_Region	Text	(CSV only) State(s) or region(s) represented in the map. Multiple values are delimited with pipe character.	Optional	FALSE	N	loc.gov JSON	
State_text	Text	Textual representation of state, copied directly from the source record	Optional	TRUE	Y	loc.gov JSON	Capital (title)
Subject_headings	Text	Subjects or keywords associated with the resource	Optional	FALSE	N	loc.gov JSON	
Title	Text	The primary title or description of the resource.	Required	FALSE	N	loc.gov JSON	

Type*of*resource	Text	A term that specifies the characteristics and general type of content of the resource, such as "Still image" or "Text." Based on MODS 3.7 enumerated list of values for typeOfResource: https://www.loc.gov/standards/mods/userguide/typeofresource.html	Required	FALSE	N	loc.gov JSON	Lookup (type*of*
Url	URL	The Digital Collections URL for the resource.	Optional	FALSE	N	loc.gov JSON	

Rights Statement

The content of the Library of Congress online Sanborn Maps Collection is in the public domain and is free to use and reuse. See https://www.loc.gov/collections/sanborn-maps/about-this-collection/rights-and-access/.

Creator and contributor information

Creator: AVP

Contributors: LC Labs, Geography and Map Division

Feedback

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