# README: Chinese Maritime Trade Statistics Collection

About this document: This README file is based on the *Datasheets for Digital Cultural Heritage Datasets* template published by Alkemade et al. (2023).

[README: Chinese Maritime Trade Statistics Collection 1](#_Toc453181221)

[Data Description 1](#_Toc139020264)

[Composition 1](#_Toc1250498484)

[Collection Process 3](#_Toc1589605611)

[Preprocessing, cleaning, and labeling 4](#_Toc141444756)

[Uses 5](#_Toc460621304)

[Distribution 6](#_Toc1182042921)

[Maintenance & Stewardship 6](#_Toc131948015)

[References 7](#_Toc769372395)

## Data Description

**Homepage:** <https://library.harvard.edu/collections/chinese-maritime-customs>

**Repository:** <https://dataverse.harvard.edu/dataverse/chinese_trade>

**Point of Contact**: Katie Mika, Data Services Librarian, Harvard Library, katherine\_mika@harvard.edu

**Dataset Summary:** This collection of datasets contain text (.txt) and tabular data (.csv) files from four titles of the Chinese Maritime Collection from Harvard-Yenching Library:

* [Trade statistics of the treaty ports, for the period 1863-1872](https://id.lib.harvard.edu/alma/990058255570203941/catalog)
* [Returns of trade at the port...for the year...](https://id.lib.harvard.edu/alma/990058255470203941/catalog)
* [Shanghai annual returns of foreign trade, 1936](https://id.lib.harvard.edu/alma/990058255550203941/catalog)
* [Annual trade report and returns](https://id.lib.harvard.edu/alma/990058255500203941/catalog) (1923-1928)

The digital objects these datasets are based upon are available here: <https://hollis.harvard.edu/primo-explore/search?query=lsr38,exact,Chinese%20maritime%20digitization%20project,AND&tab=books&search_scope=default_scope&sortby=rank&vid=HVD2&mode=advanced&offset=0>. The materials represented in the dataset were selected for digitization by Associate University Librarian for Discovery and Access, Suzanne Wones, who convened a brainstorming session with interested stakeholders to surface several possible collections or serial titles with significant tabular data., with input from other Harvard curatorial staff members.

**Languages:** Files contain data in at least English and Chinese and may include additional languages.

**Purpose:** The dataset was created to extract computation-ready datasets from digitized volumes of Chinese maritime trade logbooks. This effort aims to facilitate digital humanities research by transforming static images into dynamic datasets, enabling computational analysis and broader accessibility.

**Keywords**: Digital humanities, digital scholarship, digitization, Chinese maritime trade, historical data, economics, trade data, China

**Supported tasks:** This dataset supports tasks related to historical economic analysis, urban development studies, and digital humanities research. It is suitable for use in both research and educational contexts.

## Composition

This dataset includes digitized Chinese maritime trade log books, covering import and export statistics, taxation of commodities, and domestic and foreign factors affecting trade.

Data from page images were extracted and transformed into .txt and .csv files in partnership with Innodata: <https://innodata.com/>. Innodata’s outputs were then curated into discrete datasets, organized by port, year, and/or data table type, depending on the contents. Curation scripts are available here: <https://github.com/kmika11/chinese-maritime-customs-data>.

#### What does the data represent? Are there multiple data types or relationships between items or files?

* For “[Trade statistics of the treaty ports, for the period 1863-1872](https://dataverse.harvard.edu/dataverse/trade_statistics_treaty_ports)" and “R[eturns of trade at the port...for the year...](https://dataverse.harvard.edu/dataverse/returns_trade_ports)“ data files are collected into datasets according to ports identified on section division pages. Within each dataset, files are labeled and tagged with port and either table type or year. Files were also treated to a Named Entity Recognition process using the python Spacy library to identify potentially relevant entities in each file. These entities are provided as file tags.
* For “[Annual Trade Reports & Returns](https://dataverse.harvard.edu/dataverse/trade_reports)“ data files are collected into datasets according to ports identified on section division pages. Datasets are collected into sub-collections by year. Within each dataset, files are labeled and tagged with port and table type. Files were also treated to a Named Entity Recognition process using the python Spacy library to identify potentially relevant entities in each file. These entities are provided as file tags.
* For “[Shanghai annual returns of foreign trade, 1936: analysis of imports and exports](https://dataverse.harvard.edu/dataverse/shanghai_returns)” data files are collected into datasets according to table type. There are three tables in this collection: “Foreign Trade of Shanghai: Analysis of Imports,” “Foreign Trade of Shanghai: Analysis of Exports,” and “Foreign Trade of Shanghai: By Countries.” All data are related to the port of Shanghai. Files are labeled with table type. Files were also treated to a Named Entity Recognition process using the python Spacy library to identify potentially relevant entities in each file. These entities are provided as file tags.
* The [Summary Inventory dataset](https://doi.org/10.7910/DVN/PCQ9QL) contains this README file, inventory files, and METS files for each digitized volume. The inventory files contain metadata for all data files as spreadsheets. These may be useful in connection with the METS files to connect data files to volume page images or to evaluate extracted entities in aggregate.

#### How many records are there?

* Trade statistics of the treaty ports, for the period 1863-1872 : 15 datasets, 759 files
* Returns of trade at the port...for the year... : 13 datasets, 1896 files
* Shanghai annual returns of foreign trade, 1936: 3 datasets, 257 files
* Annual trade report and returns (1923-1928)
  + 1923: 43 datasets, 2099 files
  + 1924: 42 datasets, 2155 files
  + 1925: 42 datasets, 2293 files
  + 1926: 41 datasets, 2303 files
  + 1927: 41 datasets, 2265 files
  + 1928: 42 datasets, 2271 files
* Summary Inventory: 11 files (9 METS files, 1 README .docx file, and 1 inventory .csv file)

#### Are relationships between items, observations, or files made explicit?

Each file is organized into a dataset and sub-collection that reflects the volume, port, and/or table type that the data represents. The inventory spreadsheet: “[chinese\_maritime\_customs\_metadata\_inventory.csv](https://dataverse.harvard.edu/file.xhtml?fileId=10394062)” also makes connections between files and volume clear. This inventory file can be used in combination with each volume’s METS file to make connections between a data file and the volume’s page image.

#### Are there any errors, sources of noise, or redundancies in the dataset?

Data were transformed according to automated processes, including Optical Character Recognition (OCR), that may introduce some errors or bias to the dataset. Data have been automatically and manually validated but are not expected to be 100% accurate.

#### Does the dataset contain any confidential data?

No. All datasets and datafiles, including linked digitized volume images and metadata files, are released publicly. The datasets in this collection on Harvard Dataverse are released under a CC0 public domain waiver.

### Collection Process

#### What mechanisms, instruments, or processes were used to collect the data?

Page images for each volume are hosted by Harvard Library’s Digital Repository Service (DRS), and the transcription vendor, Innodata, used links to [IIIF manifests](https://library.harvard.edu/services-tools/iiif-manifests-digital-objects) for each title to download the image files. After processing, the original images were returned along with the processed image files, text files containing transcribed narrative text, csv files containing tabular data, ALTO xml files containing locational information, and METS xml files containing relational metadata.

After the completed files for each volume were received from Innodata, they were stored in OneDrive during local processing. Harvard Library staff with technical expertise performed quality control on the returned files. Discussions at this stage also included conversations surrounding data curation and library-supplied metadata.

The Project Manager used an Airtable database and Sharepoint site to manage internal documentation and questions and a shared query log and written status reports to communicate externally with the vendor. The project team also met biweekly with the Innodata team to resolve questions and provide feedback about the transcription process and the received data.

During data curation discussions, it became clear that the project would benefit from researcher perspectives prior to making final decisions on how to package the datasets for deposit into Harvard Dataverse. Funding was available to hire two graduate research fellows and one May-Crane Fellow to use the output files for research and provide feedback on metadata and organization.

#### If the dataset is a sample, what was the sampling strategy?

The Chinese Maritime Collection is a fully digitized, well-described corpus, that enjoys significant scholarly interest without an overwhelming volume of content. Within the collection, four titles were selected for their high density of tabular data: [Trade statistics of the treaty ports, for the period 1863-1872](https://id.lib.harvard.edu/alma/990058255570203941/catalog), [Returns of trade at the port of Shanghai](http://id.lib.harvard.edu/alma/990058255470203941/catalog), [Shanghai annual returns of foreign trade, 1936](https://id.lib.harvard.edu/alma/990058255550203941/catalog), and the [Annual trade report and returns for Shanghai](https://id.lib.harvard.edu/alma/990058255500203941/catalog). Innodata is a known and trusted partner who has been involved in refining digital images at peer institutions and at the Harvard Law School’s Library Innovation Lab and was chosen as a partner for this project.

### Preprocessing, cleaning, and labeling

#### Was any preprocessing, cleaning, or labeling done to the data?

Data were cleaned and labeled in order to improve the searchability, comparability, standardization, and usability of the data files from Innodata.

* Cleaning: csv files were validated according to shape. Csv files with values that fell outside the header columns were flagged for manual review in which a human compared the data in the csv file to the page image to move values to correct cells.
* Preprocessing & labeling: files were renamed to reflect curation and organization
* Labeling: Named Entity Recognition was applied to all files to add searchable tags and further illuminate file contents
* Preprocessing: Metadata spreadsheets were created to connect each item to its catalog record and digitized image page. These are included in the Dataverse metadata and in the Inventory spreadsheet.

#### Is the raw data included in this dataset or otherwise preserved and accessible?

Raw metadata files are included in the Inventory dataset. Only processed csv and txt files are available in this collection.

#### Is the software used to preprocess the data available?

Yes: github link.

### Uses

#### What research uses could the dataset be used for?

The objective of depositing these data into Dataverse is to maximize their benefit for future researchers. Of course, there is a limit to how well we can estimate or predict the direction of future use. Nevertheless, the project employed two graduate student researchers to investigate potential uses for this collection.

1. There was crucial historical development in China during the 18th and 19th century. Existing research has analyzed the early capitalism development in Shanghai. A topic of special interest is the rise of cosmetics at this time and how they were localized in Shanghai. Most work surrounding this area has utilized qualitative data or quantitative data from a single factory. Thus, utilizing this large collection of statistics from this collection, one could ask questions about where the cosmetic items were imported from, how much tax was levied on them, and how the quantities imported changed over 5-10 years.
2. Another topic of interest is the meteorological infrastructure in China. The goal of the development of this infrastructure was to facilitate colonial trading. However, this development has remained understudied. One research topic focused on comparing import data to the construction of lighthouses in China. Because no natural or manufactured building materials existed locally, much of the material required to construct lighthouses must have been imported.

#### Is there anything about the composition of the dataset or the way it was collected and preprocessed that might impact future uses?

1. While the data has been manually and automatically validated, there is a high likelihood that errors are still present. Due to several inherent complexities of many tables, including multi-level headers and non-standard port names, data must be investigated and validated before any specific statistical or exploratory analysis is conducted.
2. The titles in this pilot contain text that is partially or primarily in Chinese characters. Due to language limitations of both teams and systems, Chinese-language text was not able to be processed in the same way as the English-language text. During their summer working with the dataset, graduate student research fellows confirmed that having OCR for the Chinese characters would increase the value of the dataset.

### Distribution

#### Will the dataset be distributed to third parties or made available outside of the creating entity?

These datasets are available via the Harvard Dataverse Repository (<https://dataverse.harvard.edu>).

#### How will the dataset be distributed?

Each dataset has a unique DOI and Harvard Dataverse generates a recommended citation for researchers to use to formally cite the data.

#### Will the dataset be distributed under a copyright or other intellectual property (IP) license and/or under applicable Terms of Use (ToU)?

All datasets in this collection are released under the Creative Commons CC0 Public Domain Waiver. Researchers that reuse these data are encouraged to share their content openly according to standard open data practices.

### Maintenance & Stewardship

#### Who will be supporting, hosting, and maintaining this dataset?

This data collection is maintained by Harvard Library.

#### How can the owner, curator, or manager of the dataset be contacted?

Please use the Contact Owner button on the collection home page to make inquiries about the data.

#### Will the dataset be updated?

New dataset versions will be created when necessary to either add new content or to address errors identified by collection users and Harvard Library staff.

#### Will older versions of the dataset continue to be supported, hosted, or maintained?

Prior versions of data collection contents will be available through each dataset's Versions tab.

#### If others want to extend, augment, build on, or contribute to the dataset, is there a mechanism to do so?

Researchers who use and create new datasets based on this collection are encouraged to deposit their content in the Harvard Dataverse.

### References

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. “Datasheets for Datasets.” Communications of the ACM 64, no. 12 (December 2021): 86–92. <https://doi.org/10.1145/3458723>.

Houghton Library, 2023, "SAEF: Detailed Content Inventory", https://doi.org/10.7910/DVN/ZFVUHI, Harvard Dataverse, V1, UNF:6:zrqLd6p433D8qZXX/Rda+w== [fileUNF]

Alkemade, Henk, Steven Claeyssens, Giovanni Colavizza, Nuno Freire, Jörg Lehmann, Clemens Neudecker, Giulia Osti, and Daniel van Strien. “Datasheets for Digital Cultural Heritage Datasets” 9, no. 1 (October 30, 2023): 17. <https://doi.org/10.5334/johd.124>.