Visualizing Objects, Places, and Spaces: A Digital Project Handbook

China Biographical Database

Hongsu Wang

Published on: Nov 19, 2020

DOI: <https://doi.org/10.21428/51bee781.079bba92>

License: [Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License \(CC-BY-NC-ND 4.0\)](#).

Team

Project Managers: Hongsu Wang, senior project manager, the Institute for Quantitative Social Science at Harvard; Tong Wang, project manager of Peking University team, department of History, Peking University; Fan Xiang, project manager of Asia, Tsinghua University.

Executive Committee: Peter K. Bol, Harvard University (chair); Deng Xiaonan, Center for Research on Ancient Chinese History, Peking University; Michael A. Fuller, University of California at Irvine; Chen Song, Bucknell University; Chen Xiyuan, Institute of History and Philology, Academia Sinica; Chen Wenyi, Institute of History and Philology, Academia Sinica; Luo Xin, Center for Research on Ancient Chinese History, Peking University.

For more information contributors to CBDB, please see: <https://projects.iq.harvard.edu/cbdb/core-institutions-and-editors>

Project URL

<https://projects.iq.harvard.edu/cbdb>

Project Abstract

The China Biographical Database is a freely accessible relational database with biographical information about 521,442 individuals as of August 2022, primarily from the 7th through 19th centuries. With both online and offline versions, the data is meant to be useful for statistical, social network, and spatial analysis as well as serving as a kind of biographical reference. The image below shows the spatial distribution of a cross dynastic subset of 190,000 people in CBDB by basic affiliations (籍貫).

Spatial distribution of a cross dynastic subset of 190,000 people in CBDB
by basic affiliations (籍貫)

Time Needed

When did you begin this project? When did you complete this project?

This project were created by Harvard University; Peking University; Institute of History and Philology, Academia Sinica¹

Time Span: 2004 – present. This is an open-ended project and has no end date.

Length: 17 years

Outcomes

What is the outcome of the project?

Database: <https://projects.iq.harvard.edu/cbdb/download-cbdb-standalone-database>;

Research papers: <https://projects.iq.harvard.edu/cbdb/presentations-and-papers>;

Grants: <https://projects.iq.harvard.edu/cbdb/funders-cbdb>

Resources

What tools, resources, programs, or equipment did you use for this project?

For data query

- The China Biographical Database project uses Access and VBA to offer an open source platform as well as SQLite database for prosopography research for humanities scholars. [2](#)
- The China Biographical Database open source community (GitHub organization) is maintaining an open source online query system by Django, Vue.js and MySQL. [3](#)

For data analysis

Queries of the China Biographical Database can be exported to a variety of software packages. Scholars often use QGIS for geographical information system analysis, and Gephi for social network analysis.

For collecting data

- The China Biographical Database project open source community (GitHub organization) is maintaining an open source online inputting system for colleagues and volunteers to input data. [4](#) This inputting system were developed by Laravel, Vue.js and MySQL.
- The China Biographical Database project often uses regular expressions to capture the data if the original texts contains some obvious patterns. [5](#)
- If the data have some patterns, but the patterns are not quite obvious enough for regular expressions, we tend to use machine learning methods to mine data. For example, when we work on tagging person names, and their office postings from historical local gazetteers, we use Bi-LSTM-CRF as the architecture for the neural network, and use BERT to embed data. This has produced very reliable results.

Funding

Please describe any costs incurred for this project, and (if relevant) how you secured funding for these costs.

Bequest of Robert Hartwell at Harvard Yenching Institute (2005-2010), American Council of Learned Societies (2008), The Canadian Social Sciences and Humanities Research Council (2011-2015), The Chiang Ching-kuo Foundation (2011-2018), National Endowment for the Humanities (2009-2011), Harvard University and the Harvard University Asia Center (2008, 2009-2011), Center for Research on Ancient Chinese History at Peking University (2010), Institute of History and Philology, Academia Sinica (2006-), The Henry Luce Foundation (2012-2015), The Tang Research Foundation (2015-). [6](#)

Now we are working on commercialization with ChineseAll. [7](#)

Workflow

Please give an overview of the workflow or process you followed to execute this project, including time estimates where possible.

The basic workflow for China Biographical Database project is:

Select resources: We select the resources which are both important for historians and easy to work with.

Data mining: Regular expressions and neural networks are two very efficient tools for mining data.

Disambiguation and coding data: Historical figures used a variety of aliases in addition to their proper names, official titles and place names were often expressed in non-standard ways. This necessitated disambiguating and coding so as to standardize the data.

Importing data: Finally, we import the data from the previous step into 60+ different tables. The primary key and foreign key mechanisms in these tables remove duplicates automatically.

With new techniques emerging and so many Chinese historical documents and resources becoming available in searchable digital formats, the project has no end date planned.

Challenges & Opportunities

What, if anything, changed between beginning your project and its current/final form?

Techniques: Initially data entry was manual. Once we learned to use regular expressions to mine data we have pursued new techniques and algorithms to improve data mining, and we have changed the data structure based on access to new kinds of data. With the development of machine learning, we started to use some machine learning algorithms, for example [random forest](#). Now we have applied the neural network models such as [long short-term memory](#) and [BERT](#).

Budget: We have been supported by funds from national and private foundations, universities, and research institutions. Today we are also turning to commercialization as the best way to popularize the database.

Is there anything specific you wish you had known when beginning your project that might help other people to know?

Data mining is a better approach than inputting data by hand for a database project because a database is not a dictionary. It's necessary for a database to collect data systematically to avoid biases and to meet the needs of prosopographical research. Data mining is a good method to collect the data from thousands volumes of books in a short time, which is very efficient for creating systematic data.

Next Steps

Do you have any plans to follow up on this project or work on something similar in the future?

1. We will collect data to cover all dynasties in China;
2. We will collect more data from local histories;

3. We will seek out new data systematic sources as they appear in digital formats.

Footnotes

1. <https://projects.iq.harvard.edu/cbdb/core-institutions-and-editors> ↵
2. <https://projects.iq.harvard.edu/cbdb/download-cbdb-standalone-database> ↵
3. <https://github.com/cbdb-project/cbdb-online-query-app> ↵
4. <https://github.com/cbdb-project/cbdb-online-main-server> ↵
5.
http://projects.iq.harvard.edu/files/cbdb/files/a_regex_machine_yamangil_chen_bol.pdf
https://projects.iq.harvard.edu/files/cbdb/files/from_text_to_data_extracting_posting_dat.pdf ↵
6. <https://projects.iq.harvard.edu/cbdb/funders-cbdb> ↵
7. <http://www.col.com/> ↵