Journal of Open Humanities Data

China Biographical Database (CBDB): A Relational Database for Prosopographical Research of Pre-Modern China

DATA PAPER

SONG CHEN
HONGSU WANG
*Author affiliations can be found in the back matter of this article

ubiquity press

ABSTRACT

The China Biographical Database (CBDB) is the largest prosopographical database for the study of Chinese history. We use regular expressions and neural network models to systematically harvest data from primary and secondary sources and employ an entity-relationship model to organize our data. As a relational database with both online and offline versions, CBDB provides freely accessible, structured data for macroscopic, quantitative studies of premodern China. The data in CBDB is continuously disambiguated and readily formatted for statistical, social network, and spatial analyses, and also has value for tagging named entities in historical texts and contextualizing other data collections.

CORRESPONDING AUTHOR:
Song Chen
Department of East Asian Studies, Bucknell University, Lewisburg, US
song.chen@bucknell.edu

KEYWORDS:
Chinese history; relational database; prosopography; geographical information system; social network analysis

TO CITE THIS ARTICLE:
Chen, S., & Wang, H. (2022). China Biographical Database (CBDB): A Relational Database for Prosopographical Research of Pre-Modern China. *Journal of Open Humanities Data*, 8(1); 4, pp. 1–6. DOI: *https://doi.org/10.5334/johd.68*

(1) OVERVIEW

REPOSITORY LOCATION

The database is available in both Microsoft Access and SQLite versions on Dataverse at *https://doi.org/10.7910/DVN/PAGGQS* and on Github at *https://github.com/cbdb-project/cbdb_sqlite*. They are regularly updated with new contents and functions.

CONTEXT

The China Biographical Database (CBDB) amasses biographical information from disparate historical sources to facilitate quantitative, prosopographical research of premodern China. The project originated with the dataset that Robert M. Hartwell (1932–1996) created between the mid-1970s and 1995, as part of his research on the social and political history of middle-period China (ca. 7th–13th century), and willed to the Harvard-Yenching Institute. In 2004–05, Michael A. Fuller restructured and converted the data from dBase first into FoxPro and then into Microsoft Access format. It has since been transferred to the Fairbank Center for Chinese Studies at Harvard University, which, together with the Center for Research on Ancient Chinese History at Peking University and the Institute of History and Philology at Academia Sinica, continued to add new contents under the direction of an international committee chaired by Peter K. Bol. Over the past sixteen years, CBDB has grown from a database of about 25,000 individuals to include approximately 491,000 individuals (as of May 2021) whose lives spanned from the seventh through nineteenth centuries and is available for scholarly use in several online and offline (Microsoft Access, Microsoft SQL Server, MySQL, and SQLite) versions.[1] The contents of CBDB benefit from, and are inevitably shaped by, China's historiographical tradition which provides rich data on family relations, literary exchanges, intellectual interactions, and careers in government, among others, but is often reticent about issues like gender relations and economic transactions. Because of this, CBDB has 275,945 records on bureaucratic appointments, 482,953 records on kinship relations, 160,219 records of non-kin social connections, but hardly any on economic activities as of May 2021.

(2) METHOD

STEPS

There are two core tasks in our data collection: data mining and disambiguation. CBDB is a relational database that uses the entity-relationship model to organize biographical information. Persons are a type of entity. So are places, texts, offices, and so forth. Each entity has its own set of attributes (e.g., each person has a birth year and a death year, and each place has a longitude and a latitude), and every life event is conceptualized as an instance of a relationship between multiple entities (e.g., a bureaucratic appointment is an instance of relationship, from the beginning to the end year of that appointment, between a person, the office he held, and the jurisdiction of that office). Data collection is, in substance, a matter of identifying named entities and their relationships in historical sources that are described in narrative forms. For this purpose, we have experimented with several data mining approaches and found value in algorithms based on regular expressions and neural network models, such as Bidirectional Encoder Representations from Transformers (BERT) and Bidirectional Long Short-Term Memory (Bi-LSTM). We use BERT, for example, to create a vector representation of each Chinese character (an approach known as "word embedding"), which allows us to capture semantic and syntactic relations between characters through mathematical operations. We also use Bi-LSTM to tag the characters and predict whether a character is part of a string that signifies a specific person, place, or bureaucratic office. Outputs from these automated data mining algorithms are reviewed by an editorial team before they are prepared for inclusion into our database.

1 The Microsoft Access and SQLite versions of CBDB are updated on a regular basis. To download the most recent version of CBDB in the Microsoft Access format, see *https://projects.iq.harvard.edu/cbdb/download-cbdb-standalone-database*. The up-to-date SQLite version is downloadable from *https://github.com/cbdb-project/cbdb_sqlite*. Our Microsoft SQL Server is currently undergoing alpha testing. The MySQL version of CBDB provides data dump for development teams and other experienced users upon request. The CBDB online querying and data visualization interface for general use is developed by our commercial collaborator and accessible via *http://www.inindex.cn/*. With collaboration from Academia Sinica and the CBDB open-source community, we have also been developing various backend APIs (CC BY-NC-SA 4.0) that support the future design of alternative online interfaces (*https://github.com/cbdb-project/cbdb-online-main-server/blob/develop/API.md*).

DOI: 10.5334/johd.68

In merging newly harvested data into CBDB, the chief challenge comes from the complex relationship in natural language between a name and the entity it signifies. CBDB assigns a unique identifier ("id" or "code") to each named entity regardless of how it is referenced in the sources, and our development team makes every effort to disambiguate all newly harvested data before incorporating them into the database. Take persons for example. While we are blessed by the fact that most people of all walks of life in Chinese society, unlike the Europeans, had possessed both a family name and a given name since the Han dynasty (202 BCE–220 CE) and had the flexibility of composing given names from almost any Chinese character, it is not rare for two persons to have exactly the same name. On the other hand, members of the elite in imperial China were typically known by a wide variety of names and could be referred to by their office titles and other honorific appellations. Therefore, it is often necessary to disambiguate personal names and appellations in historical sources. In practice, we make use of a variety of biographical information such as alternative names, birth and death year, native place, examination degree, and data on kinship and social connections to distinguish a person from his namesake and consolidate data points about the same person whom the sources reference in various ways.

We do not only disambiguate and code entities, but also disambiguate kinship relations. We have designed a set of symbols to describe kinship relations with greater precision than they are expressed in the natural language (e.g., we use FBS and MBS [father's or mother's brother's son], among others, to distinguish different kinds of paternal and maternal cousins). We also normalize social relations by aggregating varied expressions found in historical sources into coded categories. Natural language has numerous ways of describing social relations. While the nuances in these descriptions (e.g., to censure someone vs. to criticize someone) merit attention and may, at least in some cases, reflect subtle differences in the nature of actual social relationships or the perceptions thereof, the strength of CBDB lies in facilitating the analysis of a large amount of historical data in the aggregate. To achieve this goal, we classify social relations into coded categories. As of May 2021, we have 470 pairs of coded relations that are further organized into larger classes and subclasses, which include literary exchanges, teacher-disciple ties, supportive or oppositional political relations, and so forth. After fully disambiguating and normalizing ("coding") named entities and their relations, we partition the data into separate tables which are subsequently uploaded to the database. The primary key in each data table eliminates duplicate records, and the foreign key ensures proper linkage between tables.

Disambiguation and normalization are time-consuming tasks that require domain knowledge in specific historical periods and topics. To expedite the process, we launched a crowdsourcing platform in 2021 to encourage contributions from historians of premodern China.

SAMPLING STRATEGY

Our ultimate goal is to collect all biographical information in the extant historical record of premodern China. Resource constraints, however, require that we must set priorities. To produce a large collection of data for scholarly use within a reasonable timeframe, we have worked mainly with digitized, searchable texts, especially those that were written and formatted in a style particularly suitable for automated data extraction, and prioritized data sources that can systematically expand the coverage of our database. These include both modern scholarly works, such as biographical sketches and rosters of officeholders compiled by twentieth-century historians, and primary historical documents, such as biographies in official histories and local gazetteers, tomb epitaphs, records of imperial examination graduates, and the lists of letters and other writings in literary collections.

Several biographical dictionaries, compiled in the 1960s and 1970s, provide a large assemblage of material on the lives of approximately 70,000 persons between the tenth and seventeenth centuries (Chang & Wang, 1974; Wang, Li, & Pan, 1979; National Central Library, 1965). By systematically harvesting the data in these dictionaries, the CBDB team managed to create basic profiles for a large number of historical figures during an early phase of our project.

Since then, we have expanded coverage by concentrating data collection in three areas: bureaucratic appointments, family relations, and literary exchanges. We have collected data from two multi-volume compendia which contributed more than 35,000 records on prefectural appointments from the seventh to thirteenth centuries (Yu, 2000; Li, 2001).

These were recently supplemented by another 107,000 entries on local appointments taken from 158 local gazetteers compiled in Ming-Qing times (1368–1912). Using fifty-two examination records from the Ming dynasty (1368–1644), we have added roughly 14,116 metropolitan examination graduates and their 130,000 relatives into the database. We are now expanding data coverage in this area with a new dataset containing 19,576 Song-dynasty (960–1279) examination graduates based on a recent publication (Fu, Gong, & Zu, 2009). With the help of Tang historians (Yao Ping and Nicolas Tackett), we have added some 100,000 instances of kinship relations from tomb epitaphs between the seventh and tenth centuries (Zhou, 1992; Zhou & Zhao, 2001), and we are currently preparing a massive collection of officeholding data from Song-dynasty administrative documents (Xu, 2014).

At present, the majority of our data on social relations are based on records of literary exchanges. We collected 18,124 instances of poetic exchange between the seventh and tenth centuries, based on the work of a modern scholar (Wu, 1993), and some 8,800 instances of epistolary exchange between the tenth and thirteenth centuries based on *Complete Song-Dynasty Prose* (Zeng & Liu, 2006). We will soon add another 40,000 instances of epistolary exchange from Ming-dynasty (1368–1644) literary collections. For a full list of our data sources, see *https://projects.iq.harvard.edu/cbdb/cbdb-sources*.

In addition, we have also coded and incorporated data from existing databases that focus on specific social groups and historical periods. These include, for example, a massive collection of data on family relations and officeholding for more than 46,000 persons from the Database of Names and Biographies (Institute of History and Philology, Academia Sinica, n.d.) and some 5,000 female writers from Ming-Qing Women's Writings Project (McGill University, n.d.).

CBDB is a work in progress and has no end date planned. Its current contents reflect its history that began with Hartwell's dataset of Song-dynasty officials and gradually extended back into the Tang dynasty and forward into the Yuan, Ming, and Qing dynasties. As more historical texts from premodern China become available in searchable digital formats and the technology of data mining improves, the contents of CBDB will continue to grow.

QUALITY CONTROL

Our editorial group, composed of doctoral students in Chinese history who specialize in various topics and periods, review the output from data mining algorithms and, when necessary, manually input data into our database. Additionally, when new data are prepared for uploading to CBDB, the primary and foreign keys in data tables also function as a line of defense for data integrity.

(3) DATASET DESCRIPTION

OBJECT NAME

SQLite version: CBDB_20210525.7z;

Microsoft Access version: CBDB_bc_20210525.7z

FORMAT NAMES AND VERSIONS

CBDB is available for downloading in SQLite and Microsoft Access versions. Both its content and interface are constantly evolving. Data contents are dated by the most recent update in the format of yyyy-mm-dd, and the interface is versioned using two lowercase English letters (the latest release is the bc version).

Creation dates – 1970s to 2021–05–25

Dataset creators – Current executive committee members include Peter K. Bol (Harvard University, Chair), Xiaonan Deng (Center for Research on Ancient Chinese History, Peking University), Michael A. Fuller (University of California at Irvine), Song Chen (Bucknell University), Hsi-yuan Chen (Institute of History and Philology, Academia Sinica), Wenyi Chen (Institute of History and Philology, Academia Sinica), Xin Luo (Center for Research on Ancient Chinese History, Peking University). Current project managers are Hongsu Wang (Harvard University) and Yang Xu (Peking University). For a list of past and present committee members, editors,

and other contributors, see https://projects.iq.harvard.edu/cbdb/core-institutions-and-editors. For a list of crowdsourcing contributors, see https://projects.iq.harvard.edu/cbdb/cbdb-crowdsourcing-projects.

Language – Variable names are in English. Data are bilingual (English and Chinese).

License – CC BY-NC-SA 4.0

Repository name – Dataverse and Github

Publication date – 2021–05–25

(4) REUSE POTENTIAL

CBDB assembles biographical information from disparate sources and is particularly suited for data-driven, social scientific research that aims at discovering macroscopic patterns in Chinese history and complements the qualitative, humanistic approach of close reading. The current coverage of CBDB makes it particularly powerful for prosopographical studies of the Chinese elite from the seventh through nineteenth centuries. The data in CBDB is continuously disambiguated and readily formatted for statistical, social network, and spatial analyses. A growing number of articles are published every year that use CBDB data to explore topics ranging from career trajectory, regional composition, and family connections of civil officials to intellectual and social networks of Neo-Confucian moral philosophers, antiquities collectors, and members of political factions. For a full list of publications that use CBDB data, see https://projects.iq.harvard.edu/cbdb/publications-use-cbdb-data.

CBDB also has immense value for developing new digital projects. Online text markup platforms, like MARKUS (Ho & De Weerdt, n.d.), use CBDB code tables to tag persons, bureaucratic offices, places, and temporal references in user-uploaded historical texts. Specialized databases (e.g., Database of Names and Biographies) access CBDB, through our API, to provide more context to their data collections. The Chinese Text Project integrates data from CBDB and other sources to produce a knowledge graph in its Data Wiki (Sturgeon, n.d.), and the Shanghai Library uses our data for its Linked Open Data project (Shanghai Library, n.d.). Universities, such as Tsinghua, use CBDB to teach digital methods for Chinese studies and incorporate CBDB into their pedagogical platforms (Tsinghua University, n.d.) that train the next generation of digital humanists.

FUNDING INFORMATION

COL Digital Publishing Group Co., Ltd. (2018–)

The Tang Research Foundation (2015–17)

The Henry Luce Foundation (2012–15)

Institute of History and Philology, Academia Sinica (2006–)

Center for Research on Ancient Chinese History, Peking University (2010–)

Harvard University and Harvard University Asia Center (2008, 2009–2011)

The National Endowment for the Humanities (2009–2012; PW-50438-09)

Chiang Ching-kuo Foundation for International Scholarly Exchange (2011–2018)

The Social Sciences and Humanities Research Council of Canada (2011–2015)

The American Council of Learned Societies (2008)

Bequest from the Estate of Robert Hartwell to Harvard-Yenching Institute (2005–2010)

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Song Chen: Conceptualization, Methodology, Writing – original draft.

Hongsu Wang: Data Curation, Project Administration, Software, Writing – review & editing.

AUTHOR AFFILIATIONS

Song Chen orcid.org/0000-0003-3922-4792
Department of East Asian Studies, Bucknell University, Lewisburg, US

Hongsu Wang orcid.org/0000-0002-1840-2046
Institute for Quantitative Social Science, Harvard University, Cambridge, US

REFERENCES

Chang, B., & **Wang, D.** (1974). *Song ren zhuanji ziliao suoyin* 宋人傳記資料索引 [Index to Biographical Materials of Song Figures]. Taipei: Dingwen shuju.

Fu, X., Gong, Y., & **Zu, H.** (2009). *Song dengke ji kao* 宋登科記考 [Research on Examination Graduates of the Song Dynasty]. Nanjing: Jiangsu jiaoyu chubanshe.

Harvard University, Academia Sinica, and Peking University. *China Biographical Database*. *https://projects.iq.harvard.edu/cbdb*

Ho, H. L. B., & **De Weerdt, H.** *MARKUS: Text Analysis and Reading Platform*. *https://dh.chinese-empires.eu/markus/beta/*

Institute of History and Philology, Academia Sinica. *Database of Names and Biographies* 人名權威人物傳記資料庫. *https://newarchive.ihp.sinica.edu.tw/sncaccgi/sncacFtp*

Li, Z. (2001). *Songdai junshou tongkao* 宋代郡守通考 [Comprehensive Studies on Song-Dynasty Prefects]. Chengdu: Ba Shu shushe.

McGill Library. *Ming-Qing Women's Writings Project*. Directed by Grace S. Fong and Song Shi. *https://digital.library.mcgill.ca/mingqing/english/index.php*

National Central Library. (1965). *Ming ren zhuanji ziliao suoyin* 明人傳記資料索引 [Index to Biographical Materials of Ming Figures]. Taipei: Guoli zhongyang tushuguan.

Shanghai Library. *CBDB Linked Open Data*. *https://cbdb.library.sh.cn*

Sturgeon, D. *Chinese Text Project Data Wiki*. *https://ctext.org/tools/linked-open-data*

Tsinghua University. *Tsinghua Digital Humanities Teaching and Research Platform* 清華大學數字人文教學與研究平臺. *http://qh.nqcx.net*

Wang, D., Li, R., & **Pan, B.** (1979). *Yuan ren zhuanji ziliao suoyin* 元人傳記資料索引 [Index to Biographical Materials of Yuan Figures]. Taipei: Xinwenfeng chuban gongsi.

Wu, R. (1993). *Tang Wudai ren jiaowangshi suoyin* 唐五代人交往詩索引 [Indexes to the Exchange Poems of Tang and Five Dynasties]. Shanghai: Shanghai guji chubanshe.

Xu, S. (2014). *Song huiyao jigao* 宋會要輯稿 [Collected Administrative Documents from the Song Dynasty]. Shanghai: Shanghai guji chubanshe.

Yu, X. (2000). *Tang cishi kao quanbian* 唐刺史考全編 [Complete Collection of Studies on Tang-Dynasty Prefects]. Hefei: Anhui daxue chubanshe.

Zeng, Z., & **Liu, L.** (2006). *Quan Song wen* 全宋文 [Complete Song-Dynasty Prose]. Shanghai: Shanghai cishu chubanshe.

Zhou, S. (1992). *Tangdai muzhi huibian* 唐代墓誌彙編 [Collection of Tang-Dynasty Tomb Epitaphs]. Shanghai: Shanghai guji chubanshe.

Zhou, S., & **Zhao, C.** (2001). *Tangdai muzhi huibian xuji* 唐代墓誌彙編續集 [Sequel to the Collection of Tang-Dynasty Tomb Epitaphs]. Shanghai: Shanghai guji chubanshe.

TO CITE THIS ARTICLE:
Chen, S., & Wang, H. (2022). China Biographical Database (CBDB): A Relational Database for Prosopographical Research of Pre-Modern China. *Journal of Open Humanities Data*, 8(1): 4, pp. 1–6. DOI: *https://doi.org/10.5334/johd.68*

Published: 27 January 2022

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.