Harvesting Big Biographical Data for Chinese History: The China Biographical Database (CBDB)

Lik Hang Tsui[1]* and Hongsu Wang[2]

[1]City University of Hong Kong and [2]Harvard University
*Corresponding author. Email: lhtsui@cityu.edu.hk

doi:10.1017/jch.2020.21

Abstract
Biographies constitute the main historical record of China. The China Biographical Database (CBDB) is an important project that tackles this vast biographical material with digital technologies. With both online and offline versions, CBDB is meant to be useful for statistical, social network, and spatial analysis, as well as serving as biographical reference. Through the wide range of data it collects through mining historical texts and reference sources, CBDB offers multiple ways to examine the lives of past groups and individuals in Chinese history. The use of CBDB data for prosopographical and other types of analysis has generated important work that interprets Chinese history in new ways, and has also fostered new forms of digital humanities collaborations. This article introduces the history of the CBDB project and its methods for populating its biographical data. It also presents the ways that historians and other scholars could utilize its data for research and teaching.

Keywords: China Biographical Database (CBDB); biographical data; prosopography; digital history; digital humanities

Introduction

Biographies constitute the main historical record of China. The various forms of biographical writing leave us a vast wealth of sources for studying China's past. According to one estimate, there are at least 500,000 extant biographies from the 3,000 years from the Shang until the Qing dynasties.[1] The actual numbers are probably much higher, giving historians an astounding number of records to study. As David Nivison remarks, "Any historian approaching the subject of traditional Chinese biography may be confident of at least one fact: he will be overwhelmed. The sheer volume of this material available in the Chinese sources is exceedingly large."[2]

In a digital age, the tools that equip us to deal with this staggering volume of material are more powerful than ever. The China Biographical Database (CBDB) is one of the projects that tackles this vast biographical material with digital technologies. CBDB is a relational database with biographical information on approximately 470,000 individuals (as of May 2020), primarily from the 7th through 19th centuries. With both online

[1]Endymion Wilkinson, *Chinese History: A New Manual* (Cambridge, MA: Harvard University Asia Center, 2012), 149.

[2]David S. Nivison, "Aspects of Traditional Chinese Biography." *Journal of Asian Studies* 21.4 (1962), 457.

and offline versions, this large-scale data is meant to be useful for statistical, social network, and spatial analysis, as well as serving as a kind of biographical reference. The long-term goal of the CBDB project is to systematically include all significant biographical material from China's historical record. Apart from being an important research source, CBDB is also one of the largest digital humanities utilities for Chinese studies, and is becoming a fundamental research tool for the field.

History and Methods

CBDB originates from the work by the late social historian Robert M. Hartwell (1932–1996). Hartwell first conceived of building a database to study the social and family networks of mid-Tang to mid-Ming dynasty elites based on a large number of historical examples that he has collected and studied. He took the first step to populate a database with such records, structuring his information around persons, locations, the bureaucratic system, kinship, and social associations.[3] He bequeathed the data (which by then contained more than 25,000 individuals mostly from the Song period, 4,500 bibliographic entries, and his work on the historical GIS of China) to the Harvard-Yenching Institute. From 2005, Peter K. Bol at Harvard organized a project to make Hartwell's data publicly available. Michael A. Fuller of UC Irvine, began redesigning the database structure and computer application, and later developed the Microsoft Access version. Deng Xiaonan of Peking University led graduate students trained in middle period history at the Center for Research on Ancient Chinese History in revising and expanding the contents of the database. Lau Nap-yin of the Institute of History and Philology at Academia Sinica arranged to make digital texts of historical sources available for CBDB. CBDB is now jointly owned and administered by the Fairbank Center for Chinese Studies at Harvard University, the Institute of History and Philology at Academia Sinica, and Center for Research on Ancient Chinese History at Peking University.

Thanks to the efforts of many project members and volunteers, CBDB has greatly expanded in both temporal coverage and scope. Over the past 15 years, it has run a series of sub-projects in the digitization of Chinese sources and research into the resultant data. It collects biographical data and organizes it into tables and fields, so that this data can be queried and analyzed on small and large scales, in Chinese or English. One main task that the CBDB project undertakes is the extraction of biographical data from historical texts. In order to provide an extensive coverage of Chinese historical figures, CBDB systematically works through texts of a given format, as opposed to doing research on each individual. It does incorporate research on individual figures done by qualified contributors, but devotes its main energies to data mining of texts that contain records about large numbers of people. The CBDB project prioritizes sources that provide systematic coverage of a group of historical figures, and especially those that are easier to turn into computer files and mine digitally. For instance, the project team incorporated a large number of reference works on Tang China (618–907) into the database from 2015 onwards.[4] The processes for including the biographical data in these

[3]Robert M Hartwell, "A Computer-based Comprehensive Analysis of Medieval Chinese Social and Economic History," in *Characters and Computers*, edited by Victor H Mair and Yongquan Liu (Amsterdam: IOS, 1991), 89–121. Also see Peter K. Bol, "The Late Robert M. Hartwell 'Chinese Historical Studies, Ltd.' Software Project," https://projects.iq.harvard.edu/files/cbdb/files/the_late_robert_m._hartwell_chinese_historical_studies_ltd._software_project.pdf.

[4]These include Wu Tingxie 吳廷燮, *Tang fangzhen nianbiao* 唐方鎮年表 (Beijing: Zhonghua shuju, 1980); Fu Xuancong 傅璇琮, *Tang Wudai renwu zhuanji ziliao zonghe suoyin* 唐五代人物傳記資料綜

works involve both semi-automated digitization as well as manual curation and proof-reading, so as to ensure the efficiency of this digitization and the accuracy of the data.[5]

Currently, CBDB persons are for the most part from the seventh through the early twentieth century (from the Tang through the Qing dynasties). Its coverage supersedes all print biographical dictionaries of China.

The structure of CBDB allows for many variables about any historical figure to be recorded. The core entity that defines biography in CBDB is (1) People. In addition to recording their basic biographical information, the database has designed tables to track their (contemporary and posthumous) relations with other people: (2) Kinship and (3) Social (Non-kin) Associations. To record their political and socio-cultural institutions as well as activities, the CBDB also contains designated tables for: (4) Status (reputation as a poet or such), (5) Modes of Entry into Government or other careers (e.g. passing the civil service examinations), (6) Postings to office (e.g. a prefect), (7) Events of significance in which a person participates, and (8) Social Institutions in which people collectively participated (e.g. temples and academies). In the database there is also information about texts that people produced and through which we learn about them: (9) Texts (including both references to primary texts and secondary texts) and (10) Data Sources (from which CBDB extracts its information). Last but not least, there are structured aspects of the world with which people interacted: the (11) Geographic Administrative Hierarchy (administrative units), (12) Longitude and Latitude (of place names), and (13) Bureaucratic Organization (changes in the bureaucracy over time).

To enrich its data, the CBDB project is collaborating with other database projects to share and incorporate their biographical data, or to develop and share resources. These include: Ming Qing Women's Writings (McGill University),[6] Academia Sinica's search engine for biographical materials 人名權威資料庫,[7] the Pers-DB Knowledge Base of Tang Persons (Kyoto University), Prosopographic and Social Network Database of the Tang and Five Dynasties (University of California, Berkeley),[8] data collections developed by the Lee-Campbell research group (Hong Kong University of Science and Technology),[9] Chronological Map of Tang-Song Literature (South-Central University for Nationalities), etc.[10]

合索引 (Beijing: Zhonghua shuju, 1982); Wu Ruyu 吴汝煜, *Tang Wudai ren jiaowang shi suoyin* 唐五代人交往詩索引 (Shanghai: Shanghai guji chubanshe, 1993); Yu Xianhao, *Tang cishi kao quanbian* 唐刺史考全編 (Hefei shi: Anhui daxue chubanshe, 2000); Yu Xianhao 郁賢皓, *Tang jiuqing kao* 唐九卿考 (Beijing: Zhongguo shehui kexue chubanshe, 2003); Xu Song 徐松, *Dengkeji kao buzheng* 登科記考補正, revised by Meng Erdong 孟二冬 (Beijing: Yanshan chubanshe, 2003). For the details of this sub-project on Tang China, see Xu Liheng 徐力恆 [Lik Hang Tsui], "Tangdai renwu da shuju: Zhongguo lidai renwu zhuanji ziliaoku he shuwei shixue" 唐代人物大數據：中國歷代人物傳記資料庫（CBDB）和數位史學, in *Shuma shidai de Zhongguo renwen xueke yanjiu* 數碼時代的中國人文學科研究, edited by Tan Guogen 譚國根 et al. (Taibei: Xiuwei zixun, 2018), 121–39.

[5]Lik Hang Tsui and Hongsu Wang, "Semi-Automating the Transformation of Chinese Historical Records into Structured Biographical Data," in *Digital Humanities and Scholarly Research Trends in the Asia-Pacific*, edited by Rebekah Wong, Haipeng Li, and Min Chou (Hershey, PA: IGI Global, 2019), 228–46.

[6]http://digital.library.mcgill.ca/mingqing/.

[7]http://archive.ihp.sinica.edu.tw/ttsweb/html_name/.

[8]https://history.berkeley.edu/nicolas-tackett.

[9]https://www.shss.ust.hk/lee-campbell-group/.

[10]https://sou-yun.cn/poetlifemap.aspx.

Period	No. of persons
Tang	53,607
Five Dynasties and Ten Kingdoms	2,612
Song	50,712
Liao	325
Jin	530
Yuan	22,438
Ming	185,475
Qing	85,751
Republican China	3,875
Others	71,172

Figure 1. Distribution of historical figures over China's historical periods in CBDB (as of April 2019)

Usage and Versions

Because CBDB records information about where people lived, where they studied, where they served in office, what offices they held, which families they were from, who they knew, etc., all these aspects of life can be correlated for large groups of historical figures. The data in CBDB and the ways that it is organized therefore lends itself to the study of social groups. Through this wide range of data it collects, CBDB offers multiple ways to examine the lives of past individuals and groups. In contrast to a single table (or spreadsheet) into which all data is loaded, the database consists of many different tables that are linked together, allowing categorization and coding of many different aspects of the life histories of people.

The transformation of historical reference works into computational data by CBDB allows for better and more flexible utilization of research outcomes on Chinese history; it allows for new ways of approaching and analyzing Chinese historical biographies and other sources. As a relational database, CBDB generates biographical data in response to simple queries (e.g. what information is there on an individual in Chinese history? Who all came from a certain locality? Who was active during a certain historical period? Who studied with whom?) as well as to complex queries (e.g. what were the social and kinship connections among all the people who entered government through the civil service examination from a certain place within a certain span of years?). This means that CBDB can be used as a biographical look-up tool since it provides detailed information about many individuals. Its more powerful use, however, is as a tool for the study of the lives of groups of people.

CBDB also supports social network analysis (SNA) as an approach to analyzing the relationships and connections of large numbers of people. A method that has been used for studying group structure in the social sciences for many decades, humanities scholars have increasingly applied SNA techniques to data derived from historical documents. CBDB also supports the use of geographic information system (GIS) technology for spatial analysis of the data, allowing users to see regional clusters and the

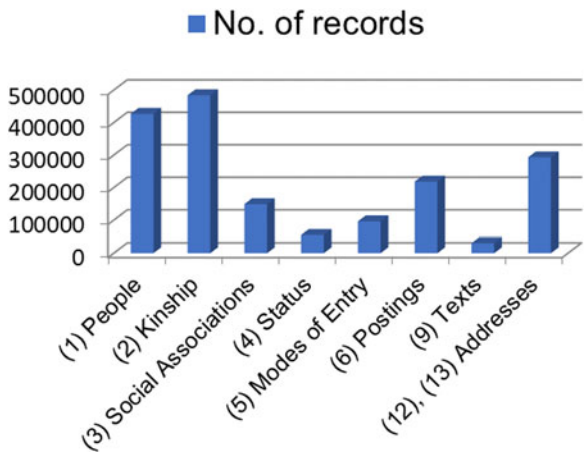

Figure 2. Number of the main types of records in CBDB (as of April 2019)

geographic spread of people and other features. CBDB draws on the China Historical GIS for most of its historical locations, but it also contributes spatial data drawn from the historical sources that it works on.[11]

The project's approach to biographical data supports the prosopographical method in the study of Chinese history. As a collective biography, prosopography is a method to examine the common characteristics of a group of people.[12] Through collecting data on phenomena that involve common aspects of people's lives, this research method sheds light on questions about social groups in history and allows scholars to make better sense of relationships between individuals and groups. Effective deployment of the prosopographical method supports discovery in texts and the relationships that they record, especially with computer-aided analysis. The use of CBDB data for prosopographical research has generated important work that helped to interpret China's past in new ways, including numerous scholarly studies in the Chinese humanities.[13]

[11]See the article on CHGIS by Peter Bol in this issue.

[12]Lawrence Stone, "Prosopography," in *Historical Studies Today*, edited by Felix Gilbert and Stephen R. Graubard (New York: Norton, 1972), 107–40.

[13]Some relatively recent examples are Peter K. Bol, "GIS, Prosopography, and History," *Annals of GIS* 18.1 (2012), 3–15; Hilde De Weerdt, Chu Ming-Kin, and Ho Hou-Ieong, "Chinese Empires in Comparative Perspective: A Digital Approach," *Verge: Studies in Global Asias* 2.2 (2016), 58–69; Song Chen, "Governing a Multicentered Empire: Prefects and Their Networks in the 1040s and 1210s," in *State Power in China, 900–1325*, edited by Patricia Buckley Ebrey and Paul J. Smith (Seattle: University of Washington Press, 2016), 101–52; Zheng Wenhao 郑文豪, "Nan-Song Fujian ren zai liang-Guang de shehui wangluo" 南宋福建人在两广的社会网络, *Fujian shifan daxue xuebao* 福建师范大学学报, no. 2 (2016), 121–29; Huang Junjie 黄军杰, "'Shuzi renwen' jishu shijiao xia quyu shi yanjiu xin qujing: yi Songdai Chuzhou jiazu qunti de shuli wei li" '数字人文'技术视角下区域史研究新取径——以宋代处州家族群体的梳理为例, *Difang wenhua yanjiu* 地方文化研究, no. 2 (2017), 106–12; Li Zonghan 李宗翰 and Zheng Li 郑莉, "Jiazu, hunyin yu daoxue: 'Xianxizhi renwu zhuan' zhong de shehui guanxi" 家族、婚姻与道学：〈仙溪志·人物传〉中的社会关系, *Tang-Song lishi pinglun* 唐宋历史评论, vol. 3 (Beijing: Shehui kexue wenxian chubanshe, 2017), 33–45; Bao Bide 包弼德 [Peter K. Bol], "Qunti, dili yu Zhonguo lishi: jiyu CBDB he CHGIS" 群体、地理与中国历史：基于CBDB 和CHGIS, *Lianghua lishi yanjiu* 量化历史研究, vol. 3–4 (Beijing: Kexue chubanshe, 2018), 213–46; Xu Yongming 徐永明, "Zhongguo gudian wenxue yanjiu de ji zhong keshihua tujing:

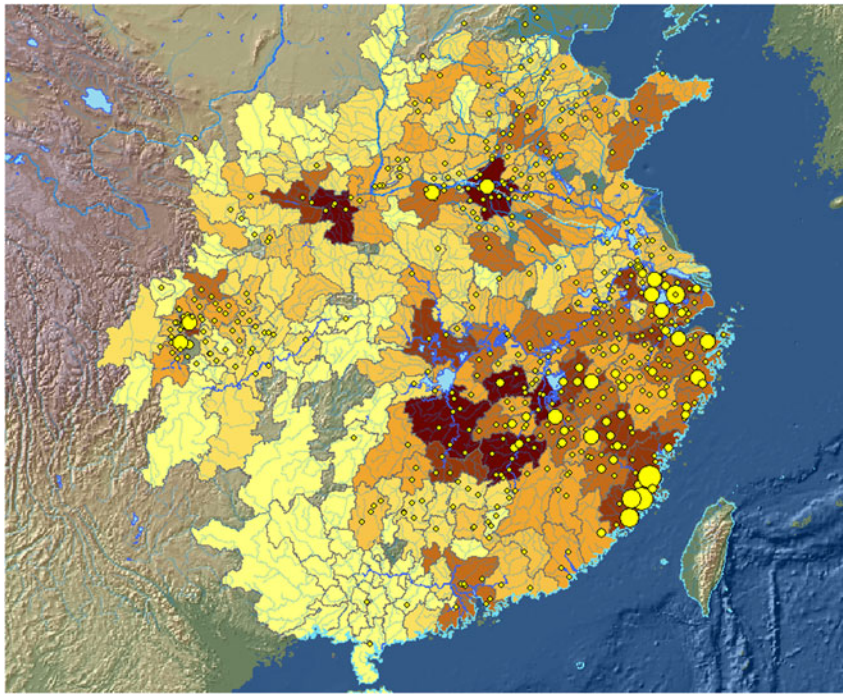

Figure 3. 4,730 examination degree holders (represented by the circles) from the Northern Song (960–1126), with population distribution as of 1080 (represented by the shades coloring the prefectures)

It has also brought about interdisciplinary collaborations with computer scientists, statisticians, visualization experts, information scientists, and so on.[14]

yi Tang Xianzu yanjiu wei li" 中国古典文学研究的几种可视化途径——以汤显祖研究为例, *Zhejiang daxue xuebao*: 浙江大学学报 no. 12 (2018), 164–74.

[14]Chao-Lin Liu, Chih-Kai Huang, Hongsu Wang, and Peter K. Bol, "Mining Local Gazetteers of Literary Chinese with CRF and Pattern Based Methods for Biographical Information in Chinese History," *Proceedings of the Third Workshop on Big Humanities Data*, 2015 IEEE International Conference on Big Data, Santa Clara, CA, October 29–November 1, 2015, 1629-38; Chao-Lin Liu and Hongsu Wang, "Matrix and Graph Operations for Relationship Inference: An Illustration with the Kinship Inference in the China Biographical Database," *Proceedings of the 2017 Annual Meeting of the Japanese Association for Digital Humanities (JADH 2017)*, Kyoto, Japan, September 11–12, 2017, 94-96; Xu Liheng 徐力恒 [Lik Hang Tsui], "Zhongguo lishi renwu da shuju" 中国历史人物大数据, *Zhongguo jisuanji xuehui tongxun* 中国计算机学会通讯, no. 4 (2018), 19–24; Yan Chengxi 严承希 and Wang Jun 王军, "Shuzi renwen shijiao: jiyu fuhao fenxi fa de Songdai zhengzhi wangluo keshihua yanjiu" 数字人文视角：基于符号分析法的宋代政治网络可视化研究, *Zhongguo tushuguan xuebao* 中国图书馆学报, no. 5 (2018), 87–103; Xiang Fan 向帆 and Zhu Shunshan 朱舜山, "Zhongguo jiapu shu de huizhi shiyan baogao: jiyu Zhongguo lidai renwu zhuanji ziliaoku de shijuehua shijian" 中国家谱树的绘制实验报告——基于中国历代人物传记资料库的视觉化实践, *Zhuangshi* 装饰, no. 10 (2018), 90–93; Chen Peihui 陈佩辉, "Renwen shujuku jianshe zhong renwen xuezhe hewei: yi 'Quan Song wen' muzhiming qinshu xinxi tiqu wei li" 人文数据库建设中人文学者何为——以〈全宋文〉墓志铭亲属信息提取为例, *Tushuguan luntan* 图书馆论坛, no. 5 (2019), 17–23. On recent explorations from a collaboration between humanists and computer scientists at Peking University, see KVisionLab, http://kvlab.org/dh.

The data is available online and as a standalone offline database in Microsoft Access and SQLite formats.[15] New versions are currently under development. Apart from working with its collaborators in universities and research institutions on scholarly research, in 2017 the CBDB project signed an agreement with ChineseAll, a digital publishing company to develop a commercial version pre-loaded with analytical tools for institutional subscribers. An Application Programming Interface (API) for system interoperability is also available, allowing other systems such as MARKUS to conduct real-time lookup of CBDB data.[16] Collaborators of CBDB have also transformed the database entries into a file for the iOS built-in dictionary application.[17]

Training on how to make use of CBDB has also become part of courses on sinological research tools and digital scholarship.[18] Detailed explanations of the database design and structure, utilities for exporting data for network and spatial analysis, etc. can be found in the CBDB Users Guide (available in Chinese and English)[19] and various sections of the CBDB website.[20] Video tutorials are also available online.[21]

The Visualization and Analysis of Historical Space

Peter Bol*

Harvard University
*Corresponding author. Email: pkbol@fas.harvard.edu

doi:10.1017/jch.2020.22

Abstract
A brief introduction to historical Geographic Information Systems and the creation of the China Historical GIS. Introductions are given to the CHGIS datasets covering 221 BCE to 1911 and the many GIS datasets on nineteenth and twentieth century China created under the leadership of the late G. William Skinner.

Maps, the preeminent form of spatial visualization, can be dated back over two millennia in China, but the earliest significant extant maps date back to the Song period. It is

[15]For the online query system, see China Biographical Database Project (CBDB), https://projects.iq.harvard.edu/cbdb/accessing-cbdb-online. For the offline standalone versions, see https://projects.iq.harvard.edu/cbdb/download-cbdb-standalone-database.

[16]See https://projects.iq.harvard.edu/cbdb/cbdb-api. For another example of how the CBDB API is utilized, see Chih-Ming Chen and Chen C. Chang, "A Chinese Ancient Book Digital Humanities Research Platform to Support Digital Humanities Research," *The Electronic Library* 37 (2019), 314–36.

[17]See "Download Latest CBDB Mac Dictionary" at https://projects.iq.harvard.edu/cbdb/download-cbdb-standalone-database.

[18]Peter Bol, "How the Digital is Changing Research and Teaching on Asia," *ASIANetwork Exchange: A Journal for Asian Studies in the Liberal Arts* 25.2. (2018), 7–28, doi: http://doi.org/10.16995/ane.278.

[19]https://projects.iq.harvard.edu/cbdb/supporting-documents.

[20]http://projects.iq.harvard.edu/cbdb.

[21]For video tutorials on youtube, see www.youtube.com/watch?v=uHWJuK308Jg&list=PLGgLzyv7BMgY8Dl1AbIBsCnVxnlRQf3Hw. For those on youku, see: https://list.youku.com/albumlist/show/id_26353417.html.