Working Towards a Data-Driven History of China: Three Examples from Digital Humanities Explorations

Lik Hang Tsui
Department of Chinese and History, City University of Hong Kong
lhtsui@cityu.edu.hk

Abstract

This presentation showcases approaches for harnessing digital technologies in the study of Chinese history. The digitization of historical information and the digital humanities paradigm enable new and fruitful explorations in researching premodern Chinese history.

Introduction

The Classical and modern Chinese language presents technical challenges in organizing and analyzing digitized material. The amount and heterogeneity of Chinese characters and expressions increase the difficulty of digitizing, organizing, and mining textual information, especially in executing optical character recognition (OCR) (to turn them into digital texts) and in word segmentation (since characters do not segment words and there are usually no punctuation or spaces to mark word boundaries in traditional Chinese texts). These are significant challenges any humanist dealing with Chinese materials digitally has to tackle. My presentation showcases computational approaches for addressing these problems in the study of Chinese history. I will use three main examples to elucidate the approaches: (1) the semi-automation of the accurate digitization of thousands of Tang biographies; (2) using data analysis in the personal name disambiguation for Chinese figures throughout a long historical span in premodern China; and (3) a large-scale visualization and analysis of communication networks in middle period China.

Reorganizing Chinese Biographies Digitally

The first example explores and analyzes the methods that the China Biographical Database (CBDB) project has developed and adopted to digitize reference works for Chinese history (Tsui and Wang 2020), which is part of the important process of turning them into structured biographical data for research use. The specific workflow under concern focuses on the Tang Dynasty (618-907) (Tsui and Wang 2019) and has implications for the improvement in digitization technologies for historical biographies in the Chinese language. These explorations and outcomes are about the transformation of large amounts of texts in non-Latin script into structured biographical data in a semi-automated fashion. This approach aims to strike an optimal balance between the employment of large amounts of machine-read texts and the efficient use of human curation to ensure accuracy.

Disambiguating Personal Names

The second example is about the names of Chinese historical figures. When integrating biographical data extracted from 2,000+ local gazetteers (*difangzhi*) into the biographical database, the usual protocol is to identify and link records of the same person, and thereby to "disambiguate" their names. Traditional Chinese naming customs pose big challenges to this, however, especially for a large gazetteer dataset containing 0.12 million records and 90k unique names of imperial government officials. Useful variables are missing in numerous gazetteer entries. I test and analyze solutions to disambiguating identical personal names in Chinese. First, the individuals who repeatedly took official posts in the same locality are

identified computationally. Then, the overlap of content in multiple gazetteers is cross-tabulated and singled out. Finally, the remaining data is corroborated with an external dataset, and then processed. Through these procedures, 51k personal names are disambiguated with optimal precision (Tsui 2021). Such a task is only possible if done digitally; it could not be performed manually. These techniques will be useful for disambiguation and Named Entity Recognition of other large-scale unstructured data in non-Latin script.

Visualizing and Analyzing Communication Networks of Literati Scholars

The final example of the presentation focuses on the digital representation and analysis of the communication networks among literati scholars in Song China (960-1279). These scholar-officials were constantly rotating in the empire's bureaucratic posts, and in order to communicate they wrote letters to each other regularly. The writing of such personal letters has been an archaic practice, yet the methods of letter writing were not stagnant and impacted the cultural and social exchanges of Chinese elites. Letters have not only become an increasingly important and sophisticated literary genre, but they are also the means of constructing a common cultural knowledge, a medium for the exchange of ideas, and above all, an important form of communication of political and personal information among elite men in traditional China (Richter 2015). Digital data about letters allows researchers to map and analyze the social networks exemplified in these epistolary connections, equipping historians to examine them in contextual and interpretive studies about Song epistolary culture and networks.

Built on these digital humanities approaches, this data-driven line of inquiry will revolutionize knowledge discovery in and the interpretation of the long span of Chinese history. It serves as an exciting interface for interdisciplinary collaboration across the humanities and technology in our increasingly digital society.

References

Richter, Antje, ed. 2015. *A History of Chinese Letters and Epistolary Culture*. Leiden: Brill.

Tsui, Lik Hang and Hongsu Wang. 2019. "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, 228–46. Hershey, PA: IGI Global.

Tsui, Lik Hang and Hongsu Wang. 2020. "Harvesting Big Biographical Data for Chinese History: The China Biographical Database (CBDB)." *Journal of Chinese History* 4, no. 2: 505–11.

Tsui, Lik Hang. 2021. "Rectifying Names: Digital Approaches to Name Disambiguation for Chinese Historical Figures." Conference presentation for Association for Asian Studies Virtual Annual Conference, Mar. 21-26.

Biography

Lik Hang Tsui is an Assistant Professor in the Department of Chinese and History of the City University of Hong Kong, as well as the Convenor of the Digital Society research cluster in the university's College of Liberal Arts and Social Sciences. He holds a doctoral degree in Oriental Studies from the University of Oxford. Before joining CityU, he worked as a Departmental Lecturer at Oxford and a Postdoctoral Fellow at Harvard University with the China Biographical Database. He has held visiting appointments at Academia Sinica, Peking University, and the Max Planck Institute for the History of Science. He specializes in middle period Chinese history and culture, as well as the digital humanities. He is the recipient of the New Researcher Award (2020) in his College in CityU. Recently, he was appointed to the editorial board of *IJHAC: A Journal of Digital Humanities*. He also edits book reviews for *Cultural History*.