China’s Rural Statistics: The Contemporary Chinese Village
Gazetteer Data Project

Yuanziyi Zhang
University of Pittsburgh

Follow this and additional works at: https://scholarsarchive.byu.edu/jeal

Part of the Chinese Studies Commons, Digital Humanities Commons, and the Library and Information Science Commons

BYU ScholarsArchive Citation
Available at: https://scholarsarchive.byu.edu/jeal/vol2020/iss171/4

This Report is brought to you for free and open access by the Journals at BYU ScholarsArchive. It has been accepted for inclusion in Journal of East Asian Libraries by an authorized editor of BYU ScholarsArchive. For more information, please contact scholarsarchive@byu.edu, ellen_amatangelo@byu.edu.
China’s Rural Statistics:  
The Contemporary Chinese Village Gazetteer Data Project  

Yuanzhiyi Zhang  

*Project Assistant for the CCVG Data Project*  
*East Asian Library, University Library System, University of Pittsburgh*  

**Introduction**  

The existing sources on Chinese rural data appear to be inadequate in many aspects. For example, a majority of rural data provided by the National Bureau of Statistics of China database\(^1\) are annual, seasonal, and monthly nationwide data. Provincewide rural data are provided in annual and seasonal bases, while very few annual citywide data have been recorded. In 2004, the Institute of Geographic Sciences and Natural Resources Research created a series of datasets extracted from the 2005 China Rural Statistical Yearbook\(^2\). The datasets cover a wide range of indices including gross output value, cultivated area, and output of major agricultural products, etc. The time span, however, is limited to a single year. The China Rural Studies Database\(^3\) constructed by the Social Sciences Academic Press (China) has digitized multiple sources including surveys, gazetteers, formal and informal documents, photographs, audios, videos, books, and articles. Villages are categorized by administrative division, region, topography, mode of production, and ethnic groups. The database covers 79 villages spanning in 19 provinces. Current sources of China’s rural data are problematic in the following ways. First, the sample size of most of the databases in the field is relatively small, which in practice often leads to large estimation errors. Second, for the few databases that do contain an adequate amount of data, the sources are usually inclusive and thus it is difficult for users to conduct either horizontal or vertical analysis without admitting sufficient assumptions.  

In response to scholars’ needs over the past few years, initiated by Haihui Zhang,\(^4\) the *Contemporary Chinese Village Gazetteer Data* (CCVG Data) project was proposed by the East Asian Library (EAL) of the University of Pittsburgh Library System (ULS) in 2018. The CCVG Data project is funded by the Office of Chancellor, Asian Studies Center, and University Center for International Studies at the University of Pittsburgh. Building on EAL’s extensive Chinese village gazetteer (cunzhi, 村志) collection (about 2,600 and continually growing), the project aims to create a web repository that is openly accessible online for data extracted from the gazetteers. A Chinese village gazetteer is a grassroots-level primary source that contains both qualitative and quantitative data on various topics including local history, genealogy, legends and myths, natural disaster records, economics, education, etc. According to Chen et al.,\(^5\) administrative villages (xingzhengcun, 行政村) are the administrative domain of rural governance established by the villagers’ committees based on the *Organic Law of the Villagers’ Committees of the People’s Republic of China*. To ensure data consistency, the project limits the scope of the gazetteers within administrative villages. While the team is endeavored to include a minimum of one village from each province of China if possible, a
random selection process is applied when there are multiple gazetteers of administrative villages available within a province.

The current datasets cover 1,000 villages spanning in 31 provinces, 212 cities, and 510 counties. A total of 18 categories of data are included in the datasets with 147 subcategories:

1. Gazetteer Information
2. Village Information
3. Natural Environment
4. Natural Disasters
5. Last Names
6. Year of First Availability/Purchase
7. Ethnic Groups—Range
8. Ethnic Groups—Yearly
9. Population and Migration—Range
10. Population and Migration—Yearly
11. Military, Politics and Management—Range
12. Military, Politics and Management—Yearly
13. Economy—Range
14. Economy—Yearly
15. Family Planning—Range
16. Family Planning—Yearly
17. Education—Range
18. Education—Yearly

Most data extracted are information after 1949 with very few exceptions dated from the late-1940s. The complete datasets are available for download in CSV format through the D-Scholarship data repository of the University of Pittsburgh. A coverage map built in Google Maps integrates spatial and administrative information of villages with ULS catalog links to the gazetteers. More information on the CCVG Data project, contact, and user survey is accessible through the project website.

The CCVG Data project has received much attention from scholars and researchers in a variety of disciplines. To enhance the accessibility of data as well as to improve user experience, in January 2020, the CCVG team initiated a cooperation with the School of Computing and Information at the University of Pittsburgh which aims to store the 18 datasets as separate CSV files into a relational database. Following an entity-relationship modeling process, existing CCVG datasets were analyzed to identify important entities, their attributes, and relationships. Using an existing open-source database management system (MySQL), the database allows effective and efficient ingesting, querying, manipulating, and displaying CCVG data. Among the valuable feedback and suggestions received from users, the desire for a cross-searching tool confirms the necessity of such a database. By the end of 2020, a graphical user interface will be implemented to facilitate scholars’ interactive exploration of the CCVG data.
CCVG work protocol

The data extraction process for CCVG data is carefully designed and strictly followed. After consulting the advisory board members of the project, a peer-assessed data dictionary is developed that functions both as a data entry instruction and a user guide. To ensure the accuracy of data, every piece of data follows a 3-step protocol: Examine a gazetteer for different categories of data and use page number as the label of the source.

1. Manually enter the extracted data into the data entry platform developed in Drupal.
2. Manually proofread all data entries.

While there is no unified guideline in the composition of the gazetteers, nomenclatures are usually used interchangeably and inconsistency on both terms and data can be found within a single gazetteer. Whenever such a discrepancy occurs, data from the most related chapters are considered representative. In general, data are extracted unprocessed with the exceptions of clear additions and subtractions and the few eliminations of obviously erroneous data. If there are further disputes regarding the accuracy of the data, users are encouraged to contact the CCVG team or to access the gazetteers physically or digitally.

Once the extraction from a certain amount of gazetteer completes, multiple departments in the University Library System work together to update the datasets, the maps, and the website, including the Office of Scholarly Communication and Publishing, Digital Scholarship Services, and Communications and Web Services, etc. If there are any changes in database and interface design, colleagues from the School of Computing and Information will address them accordingly.

Engagements and Applications

The CCVG Data project is a highly collaborative work that involves scholars in various institutions and departments. Within the University Library System (ULS), the Acquisitions Management Unit and the Catalog Management Unit are responsible for purchasing village gazetteers and managing the electronic catalog. The Hillman Library Stacks Maintenance, the Library Collections Storage Unit, and the Library Management Team oversee the circulation of books. At the nascent stage of the project, the Digital Scholarship Services and the Metadata and Discovery Unit work together to refine the mechanism and to offer suggestions to data extraction and input. The Information Technology Department and the Technical Services developed the data entry platform on Drupal and manage the D-Scholarship data repository. Our GIS Librarian from the Digital Scholarship Services incorporates administrative information of villages and links to gazetteers through Google Maps to construct a spatial-textual connection. The Media Services, the Office of Scholarly Communication and Publishing, and the Web Services and Communication work together to develop the website of the CCVG Data project. Meanwhile, eight scholars from the University of Pittsburgh, Harvard University, St. Vincent College, the Chinese University of Hong Kong, Hong Kong University of Science and Technology, and the University of California, Los Angeles are on the advisory board of the project. Scholars who work on the CCVG Data project come from diverse academic backgrounds including East Asian Languages and Literatures, Library and Information Science, Computer Science, etc. Consisting of many
graduate students who work as student workers and volunteers, the team also includes library staff and visiting scholars.

The CCVG Data project intends to support teaching, research, and reference services in general. Since the team has not yet publicly announced the project yet, the number of academic applications is limited. However, current applications and feedback we receive from the user survey already demonstrate the diversity in fields and subjects and hence the great potential of the CCVG Data project. As a graduate student in East Asian Languages and Literatures, I have worked on a project on the influence of arable land area on gross output value in Chinese villages in the course Data Information in Systems in Spring 2020. Adjusted datasets from the CCVG Data were imported in Python and conclusions were drawn based on interpretations of graphs. Starting from August 2021, I have been working on a project on the visualization of Chinese local performances from village gazetteers as part of the course Digital Studies and Methods Seminar. Because of China’s vast expanses and complex topography, various subcultures have originated and developed within relatively confined spaces. Evidenced by records of local opera performances in the village gazetteers, such diversity is featured by multiple genres of opera, distinguished troupes or actors, well-known repertoire, etc. The project aims to demonstrate and to visualize the complexity and diversity of Chinese local opera performances on the village level through ArcGIS mapping tools. By working on the project, I wish to construct a connection between opera and space as well as identify the less representative genres of local opera to draw attention to the preservation of certain cultural heritage. The project is inspired by the CCVG Data map and the data is also extracted from the ULS Chinese village gazetteer collection.

**Significance of CCVG data project**

Although developing rapidly, the field of digital humanities is featured by linguistic differences and great global inequalities. Thus, a bias can be found towards the English language and hence towards an Anglophone DH. As a digital humanities project itself, while most projects in Chinese employ scanning, photographing, and OCR, the CCVG Data project purports to fill in the gap between the abundant non-English textual materials and multiple digital tools and scholarship. The website of the CCVG Data project is bilingual and categories of data in CSV files are also displayed both in Chinese characters and in English. For village names, administrative divisions, and last names, Hanyu Pinyin (汉语拼音), a romanization of Chinese characters, is included.

In addition to linguistic diversity, the CCVG Data project also serves as a digital source for intersectional analysis. Proposed by Kimberlé Crenshaw, intersectionality originally models the patterns of racism and sexism over the experiences of women of color. Risam further generalizes the scope of intersectionality by expending the definition to “additional axes of difference including sexuality and ability.” After a thorough examination of existing DH projects, she notes in her article that it is perceived that “many digital humanities projects fail to engage with race, gender, disability, class, sexuality, or a combination thereof.” To avoid such partiality, the data structure of CCVG data is designed to be inclusive of detailed subcategories that are documented in the village gazetteers. In each demographic category of CCVG data, differences in genders and ethnic groups are considered and recorded to acknowledge the importance of intersectionality.
In terms of serving as subject reference, the CCVG Data project is also of great importance. Originated from the Neolithic Period, agriculture in China has been developing for over 8,000 years. From the traditional “men-tilling-and-women-weaving” mode to the modern mechanization and systematization, the evolution of agriculture is marked by industrial revolutions, marketization and globalization, and the wisdom of people. In recent years, however, it is witnessed that China’s agrarian economy is facing a decline as there is a significant increase in both the number of cities and urban populations. Furthermore, the uneven development of infrastructure construction, economy, education, etc. also brings up several issues including the rural-urban divide, regional differences, and other inequalities. Therefore, it is crucial to investigate China’s rural statistics to perceive a comprehensive vision over the development of China and to further obtain a holistic political or economic view of the world.

Critiques and justifications

One of the most common feedback the team receives is regarding the manual entry process. It is concerned that manual extractions would lead to mistakes and inaccuracy compared to computer-based techniques like optical character recognition. In the case of the CCVG Data project, constraints of OCR techniques on the Chinese language become one of the major obstacles. Firstly, as most village gazetteers do not have digital versions, the team would have to invest most of the time scanning and proofread the OCR results. This process may involve complicated copyright issues since the village gazetteers have been published by numerous publishers starting from the 1980s, some of which have even disappeared. Secondly, although data extracted from books are mostly numbers, the various ways of presenting numbers in the gazetteers could be troublesome. Figures could be written in Chinese characters, in Arabic numerals, or in a combination of both. Thirdly, it also requires much human effort to correctly distinguish the units of data. As there is no unified standard in the composition of gazetteers, units often vary greatly. Units for total grain output, for example, include dan (担), jin (斤), 1k jin (千斤), 10k jin (万斤), and kilogram (qianke, 千克).

The CCVG Data project currently confines its scope mostly within quantitative information. While team members notice that many categories of qualitative information are of interest to some scholars, for example, religion, myths, local customs, etc., it is difficult to incorporate such information into a database since attributes are often not fixed. However, separate projects focusing specifically on one of the categories may be applicable for future work.

Furthermore, while most people the project currently engages with are scholars and researchers in academia, the team hopes to expand the group both as users and as developers. Potential fields of application include industrial site selection, target marketing, policies of poverty alleviation, infrastructure construction, etc. In addition, as Miriam Posner discusses in her blog, it is also vital to do digital humanities work in partnership with the communities at stake. Input from local villagers and composers of the gazetteers is tremendously valuable. However, as we are experiencing such a difficult and complex historical moment due to the pandemic, it is not encouraged to conduct field studies and there is only a very limited number of villagers who can be reached through digital media.
Future work

The CCVG Data team is currently developing a user-friendly interface that enables scholars to browse and retrieve data of any category they are interested in. The first version is hoped to be launched by the end of 2020. The goal of the number of village gazetteers included in the datasets is 2,500-3,000. For derivations of the CCVG Data project, the team hopes to transfer the same methodology to other types of gazetteers including hydraulics gazetteers, rail transport gazetteers, and public health gazetteers. We welcome collaborators from all potential areas of application. Any feedback or comment is appreciated. For more information, please email ULS-EALReference@pitt.edu.

10 Ibid.
11 The team originally planned to announce the project at the Annual Conference of Association for Asian Studies in March 2020. The conference was canceled due to COVID-19.
16 Ibid.