Workshop Report

"Chinese Local Gazetteers (*difangzhi* 地方志): Historical Method and Computerized Data Collection and Analysis"

April 27-28, 2015

Max Planck Institute for the History of Science

This workshop explored computerized analysis of Chinese local gazetteers (*difangzhi* 地方志). It brought together historians, computer scientists and librarians to discuss how historians can benefit from 'big data' and how computer scientists and librarians can contribute to managing and analysing large amounts of digital texts.

Local gazetteers, which can be traced back to the seventh century, were copied, re-edited and collected throughout the centuries up until modern times. In them, scholars documented the social, political, and material composition of a place: the landscape, history, flora, fauna, taxes and products of a region, temples and schools, officials and celebrities, local festivities and customs, and weather records and disasters. During the workshop, historians presented their research on local gazetteers whilst computer scientists, GIS specialists, and librarians introduced digital methods for historical studies. Contributors discussed the use of computer aided data collection methods from local gazetteers and computational data analysis on a larger scale. Questions of interest were raised about the work computers can do once they have mastered the structures of knowledge embedded in local gazetteers. Which challenges and changes do digital versions of historical texts present and when and how can computational methods be applied to them?

Panel 1: Historical Studies on Local Gazetteers

In the millennium of its existence, the genre of local gazetteers developed a face of its own: the topics, though varying, became more consistent, and information presented about a topic became more standardized. At the same time local gazetteers were also always products of the moment with distinct characteristics and histories. Four historical studies exemplify the characteristics and idiosyncrasies of local gazetteers and review the state-of-the-art research on and within this genre. Chen, Schaefer, and Dennis demonstrated the role of local gazetteers as historical sources, both as reference points and the object of inquiry. Wu and Dennis critically reviewed the use of local gazetteers as "big data" for statistical analyses.

Chen Hsi-Yuan introduced a map of the Yangtze River recently found in the Grand Secretariat Archives at the Institute of History and Philology, Academia Sinica. His research concerns issues of production and the dating of the map. Chen draws on cross-referencing, discursive and cross-textual analysis using traditional means and full-text search to explore the history of postal relay stations (驛站), fishery tax offices (河泊所), and topographical and hydrologic features. He dates their periods of existence by analysing their appearance on the map and from other sources. Chen used the example of complex spatial relations

which digital humanity tools have as yet been unable to grasp. Metadata descriptions and standards for searching maps and their features could guide comparisons in the future.

Dagmar Schäfer illustrated the use of local gazetteers in research on sericulture and textiles. Information needed for such research is distributed across several chapters such as those on local products (*wuchan* 物產), customs (*fengsu* 風俗), and infrastructure and building sites. Schäfer's research draws on a comparison of textual sources and archaeological evidence that benefits from geospatial mapping visualisation. Further questions relate to the standardization of such information from a diachronic and diastratic perspective. When was information refreshed during the Yuan-Ming transition? Digital methods might also facilitate research into terminology changes or help reveal shifts in standards by enabling researchers to compare sequences in listings of, for example, textile types in the local products chapter.

Wu Micha critically examined the quality, value, and reliability of common data categories. He reviewed early scholarly data collections for cross-regional research such as Chen Zhengxiang's research on the locust temples, Xiu Hong's lists of local officials in Qing Taiwan, and Li Guoqi's quantitative analysis of Qing local officials. Wu pinpointed the importance of visualizing source bias. He urged that gaps in source materials or their inaccessibility should be addressed while attempting to preserve the varied styles and different kinds of information that local gazetteers provide. Wu asserted that information on local officials was both the most standardized in format and the most reliable.

Joe Dennis distinguished local gazetteers in terms of authored monographs. He illustrated the idiosyncrasies and inconsistencies of the genre and discussed the geographic distribution and finances of publication. While gazetteers described a specific locale, their authors or editors often originated from other areas. Printing happened somewhere else again. Also local gazetteers had to undergo several levels of accrediting before being rejected or censored by supervising officials. Para- and post-texts are crucial, giving an indication of the validity and consistency of the given information. Digital scans are lacking in quality: margins, seals, and notes are often cut off from pages; graphical information, texts on maps, and other kinds of marks are omitted from digital full texts. Metadata for images and maps require further attention.

Panel 2: Computational Methods on Local Gazetteers

This panel introduced five projects using computational methods for data collection and analysis. Conducted either by individual scholars or teams of scholars, all projects worked with big data and presented methodological issues concerning data collection and the analysis of the local gazetteers.

Chang Subin's research focuses on the classification and identification of plants and animals in the "local products" (*wuchan* 物產) chapters of Taiwanese local gazetteers. She illustrated tensions between the genre's classification schemes and those of modern botany and zoology. Chang compares terminology and descriptions of flora and fauna. Species were often given different names during different times or in different geographical locations which can cause confusion within the data. The team applied string-matching techniques to identify similar descriptions for records with different names to help resolve such confusion. She showed how historical research on fish names/species is linked to a

contemporary database for a longue durée scientific analysis of fish life in and around Taiwan.

Peter Bol discussed biographical and geographical data and their intersections, which he exemplified using building events such as temples, bridges, walls, or memorial arches. Such research reveals that groups of people who contributed financially to the building of local infrastructures such as bridges, are different from those who supported education and the building of arches. Bol elaborated on the modelling of life and career cycles in the China Biographical Database project (CBDB) and how the China Historical GIS project models historical landscapes. This demonstrates that by collecting such information and organizing in a relational database it is possible to transform narrative anecdote and items in lists into data that can be systematically analysed. Emphasis was furthermore put on the cross-regional character of local gazetteers: information on topographical features, for instance, are described across administrative boundaries.

Adam Mitchell delineated computational methods of automatic data extraction used for research on historical catastrophes. A computer program is used to transform raw text into marked-up XML documents. The table of contents is automatically identified to segment the text by sections, and the sections are further processed to obtain the entries contained within those sections. A database on auspices and unusual events (*xiangyi* 祥異) was then created with 15,000 entries from 137 Zhejiang gazetteers. Quantitative peaks and correlations, for instance, between different kinds of disasters such as famine and hyperinflation became visible. This digital analysis confirms research done by Pierre Etienne-Will four decades earlier with paper tools. Will's research also included qualitative features such as the colloquial description of the extent of a disaster which digital methods thus far have not taken into account.

Cao Ling presented a collaborative project that aims at the spatiotemporal visualization of weather variations in the region under Chinese imperial reign. The project builds on the edition of "Integration of Meteorological records during the past 3000 years" (中国三千年气象记录总集 (Nanjing 2004)). A team of scholars from the Nanjing University of Information Science and Technology (NUIST), Nanjing Agricultural University, Nanjing Normal University and Qinghua University has developed a database structure and software platform, which currently includes 150,000 records. They are analysing meteorological vocabulary and developing tools for visualization.

"Local materialities" stand at the heart of the Local gazetteers project at the MPIWG (https://www.mpiwg-berlin.mpg.de/en/research/projects/departmentSchaefer_SPC_MS_LocalMonographs), introduced by **Chen Shih-pei** and **Martina Siebert.** The project analyses the role of the Local Gazetteers in the development of a region's material identity, terminological change and standardization. In this project, digitized raw texts are semi-automatically transformed into tables through an extraction interface first developed at Harvard University and then enhanced and maintained at the MPIWG. The structuring of digital texts enables large-scale data collection across the gazetteers and thus allows cross-regional and cross-temporal comparisons of entries recorded in different gazetteers as well as formal features such as the number of items and depth of classification or length of entries over time. This data can be visualized as maps through time using the tool PLATIN (http://skruse.github.io/PLATIN).

Panel 3: GIS and Local History

Local gazetteers are highly relevant to geospatial information, and are one of the major sources when studying local history. Participants of this panel opened the view to building digital collections in terms of these two aspects.

Lex Berman presented TGAZ, a Temporal Gazetteer for Chinese History, a next-generation database for the China Historical Geographic Information System (CHGIS) (http://www.fas.harvard.edu/~chgis/). A major bottleneck of CHGIS is the difficulty of finding the corresponding geospatial information when adding new place names. The TGAZ data model allows a place name to be added to the database based on text descriptions and not only geometry. The TGAZ API (http://maps.cga.harvard.edu/tgaz/) provides a read-only search interface for CHGIS place names. Its interchangeable formats also allow it to be integrated into other web services like Pelagios (http://pelagios.org) and MARKUS (http://dh.chinese-empires.eu/beta/).

Zhao Siyuan presented an ongoing archival project from Shanghai Jiaotong University that processes local materials on Shicang, Huizhou, Southern Zhejiang, Jiangxi, Fujian, Poyang Lake, and Jiangjin. Materials are restored, digitized, and then catalogued, so that they can be accessed by other scholars. Historians apply digital humanities approaches like social network, textual, and geospatial analyses in researching these materials. Due to the high cost of restoring and digitizing the large amount of materials involved, in the second phase the project will have to be increasingly funded through commercial publication.

Panel 4: Text Mining and Analysis

This panel was devoted to digital tools that are developed to help researchers process and analyse texts.

Hou Ieong (Brent) Ho introduced MARKUS (http://dh.chinese-empires.eu/beta/), a semi-automatic mark-up platform for classical Chinese. It allows a user to upload a text and to run an "Automatic Markup" that will identify all named entities already known to MARKUS. MARKUS incorporates 530,000 historical personal names from the China Biographical Database (CBDB), 65,500 historical place names from CHGIS, historical Buddhist figures and glossaries from Dharma Drum Buddhist College (DDBC), as well as official titles and reign titles from CBDB. Users can also apply a feature to scan and tag texts based on a list of user-defined terms or regular expressions. All the automatic markups can be verified and refined manually. Results can be saved as HTML or TEI documents, and markups can be exported as Excel or CSV files in the form of "which tag, with which type, appears in which passages" that will allow the application of further analyses like co-occurrence and social network analysis.

DocuSky is a system for personal text databases developed by **Tu Hsieh-Chang** based on his experience with the Taiwan History Digital Library (THDL, http://thdl.ntu.edu.tw). Users upload their own collections of texts and can build from them their own personal databases. Metadata for texts can be imported from a folder structure or an Excel table. Once built the personal database allows a number of advanced features such as full-text search, metadata search, post-query classification, and tag analysis, all rendered dynamic

for search results. Users can overview the distribution of search results over several dimensions, such as temporal distribution or as a quantitative overview of the frequency of personal names and places. This allows the identification of co-occurring search criteria as well as names, places, dates, and other dimensions or tags.

Panel 5 & 6: Round table and general discussions

In the concluding **Roundtable** chaired by Fei Siyen; Pierre-Etienne Will, Joe Dennis, Hsiang Jieh and Peter Bol highlighted issues of accessibility and the role of contextual information for digital reading and analysis methods. It was agreed that the non-narrative character of local gazetteers is both a challenge and a chance for historical digital humanities research.

PARTICIPANTS

Host and organizers:

Shih-Pei CHEN, MPIWG, schen@mpiwg-berlin.mpg.de
Dagmar SCHAEFER, MPIWG, dschaefer@mpiwg-berlin.mpg.de
Martina SIEBERT, MPIWG, msiebert@mpiwg-berlin.mpg.de

Presenters, discussants, and chairs:

Lex BERMAN, Harvard University, mberman@fas.harvard.edu
Peter BOL, Harvard University, peter_bol@harvard.edu
Ling CAO, Nanjing University of Information Science and Technology, caoling1013@163.com
Subin CHANG, Taiwan Normal University, 109682@ntnu.edu.tw
Hsi-yuan CHEN, Academia Sinica, hchen@asihp.net
Joe DENNIS, University of Wisconsin Medison, dennis3@wisc.edu
Siyen FEI, University of Pennsylvania, siyen@sas.upenn.edu
Hou Ieong (Brent) HO, Leiden Institute for Area Studies, brent.ho@gmail.com
Jakub HRUBY, Oriental Institute, Czech Academy of Sciences, hruby@orient.cas.cz
Jieh HSIANG 項潔, National Taiwan University, jieh.hsiang@gmail.com
Andreas JANOUSCH, Universidad Autónoma de Madrid, andreas.janousch@uam.es
Adam MICHELL, Harvard University, adamkelbymitchell@gmail.com
Hsieh-Chang TU 杜協昌, National Taiwan University, hsieh.chang@gmail.com
Pierre-Etienne WILL, College de France, pierre-etienne.will@college-de-france.fr
Micha WU 吳密察, National Taiwan University, wumicha1984@gmail.com
ZHAO Siyuan 趙思淵, Shanghai Jiaotung University, titaner@sjtu.edu.cn

Participants:

Kevin CHANG, Academia Sinica, kchang@sinica.edu.tw
Esther CHEN, MPIWG, echen@mpiwg-berlin.mpg.de
Kaijun CHEN, MPIWG, kchen@mpiwg-berlin.mpg.de
Mingtong (Zoe) HONG, MPIWG, zhong@mpiwg-berlin.mpg.de
Matthias KAUN, Staatsbibliothek zu Berlin, Matthias.Kaun@sbb.spk-berlin.de
Duncan PATERSON, University of Heidelberg, duncan.paterson@asia-europe.uni-

heidelberg.de
Urs SCHOEPFLIN, MPIWG, schoepfl@mpiwg-berlin.mpg.de
Michael STANLEY-BAKER, MPIWG, mstanleybaker@gmail.com
Honghong TINN, MPIWG, htinn@mpiwg-berlin.mpg.de
Jorge URZUA, MPIWG, jurzua@mpiwg-berlin.mpg.de