Challenges of Multi-Source and Bilingual Data Curation for the Research of Tribute during the Qing Dynasty

中國清朝貢品交換制度研究：多資源及雙語數據管理之案例

Loretta Kim
Assistant Professor, Department of History, Hong Kong Baptist University, Hong Kong
金由美
香港浸會大學歷史學系助理教授
E-mail: lekim@hkbu.edu.hk

Eugenia Kim
Specialist II, Dorothy and Lewis B. Cullman Center, The New York Public Library for the Performing Arts, New York, U.S.A.
Email: eugenia.kim24@gmail.com

Keywords (關鍵詞)：Historical Data (歷史數據)；Digital Curation (數據管理)；Qing Dynasty (清代)；Chinese Imperial Tribute System (中國帝制時期朝賄體系)；Multilingual Media (多語言媒體)

【Abstract】

This paper discusses the objectives, process, and outcomes of creating a digital dataset for a historical research project on the tribute system in Heilongjiang during the Qing dynasty (1644-1911). Relevant information from non-digital primary sources was compiled for the dataset to facilitate quantitative and qualitative analysis of the system’s attributes. In the course of curating the data, the investigators addressed the challenges of defining a common set of variables and matching Chinese original data with English translations. They tested methods of learning to create datasets that could accommodate heterogeneous sources and share among multiple users.

【摘要】

本文將討論清代(公元1644-1911年)黑龍江朝貢體系的歷史研究項目所創建的一套數位數據集之目標、過程與結果。研究者將非數位化資料中的相關信息編譯成數據集，以便定量和定性分析該制度的屬性。在整理數據的過程中，研究者解決了一些技術性的挑戰，“
Introduction

Advances in social science data management have generated innumerable repositories with data from one or multiple sources that reveal their rich information through convenient and user-friendly search engines. Organizations such as the Interuniversity Consortium for Political and Social Research (ICPSR) and the Institute for Quantitative Social Science provide social science scholars with the options of sharing and discovering data in online environments. However, discrete datasets have limited utility in historical research. Many historians collect information from databases and organize their findings using various data management methods, ranging from manually writing and filing note cards to employing computer software such as Microsoft Excel. Bibliographies and now wikis are also common forms of aggregating and sharing data in the practice of historical scholarship and pedagogy. In the everyday work of historians, however, creating datasets from primary sources, has not yet become a common task. Although scholars who concentrate on quantitative analysis, especially in economic and demographic history, have made substantial contributions through their construction of original datasets for their peers, datasets are far less employed as essential tools in qualitative historical research.

This case study examines the process of creating a dataset from multiple primary sources to enhance both the quantitative and qualitative data analysis through a database interface. The initial purpose for compiling the dataset was to aggregate information about the Qing dynasty (1644-1912) government’s institution of collecting tribute (Ch. *gongpin*, Ma. *alban*) from its subject populations in the area customarily known as Heilongjiang (Amur River region), and disbursing “return gifts” (Ch. *wulin*, Ma. *ulin*) to these communities[1]. The goals of compiling this dataset include 1) distinguishing this system of intra-state tribute from inter-state tribute conducted between Chinese empires and other sovereign governments[2], and 2) elucidating how the former type of transactions facilitated the evolution of two frontiers: one within the Qing state and another between the Qing and Russian empires[3]. The featured tributaries, in order of input priority, are the Orochen, Solon, Dagur, Sibe, and Heje tribes. During the process of prototype creation, the potential to use the dataset in the training of graduate students in primary source and data literacy emerged.

To achieve these aims, the authors aggregated data between May 2011 and February 2012 consisting of information from Qing official documents. The data was then compiled into a Filemaker Pro[4] database prototype for its multi-lingual capacities and ease of data entry. The frequency of tribute and return gift transactions was to be searched and measured by time period rendered in both Gregorian and Chinese imperial-lunar calendar dates, and the quantity of goods exchanged were differentiated by specific citations of items. This prototype was also intended to inform the development of a more robust online repository in the future.

Background

Impetus to design this dataset stemmed from various substantive and methodological issues. This tool was developed to aggregate a comprehensive pool of evidence regarding Qing period intra-state tribute and to examine tribute as a bilateral instead of a unidirectional transaction. Most analysis of tribute during the Qing dynasty, whether discussing flows between the Qing state and other sovereign governments or between the Qing center and its various subject populations, have concentrated on single instances[5]. Similarly, the presentation of goods and persons by a sovereign polity to an incumbent Chinese imperial government, and the latter’s distribution of “return gifts” whether luxury commodities or noblewomen given in marriage to leaders of foreign
states, have been often treated separately rather than as interconnected and symbiotic processes[6].

Much of the scholarship about tribute has concentrated on the function in establishing and maintaining relations between Chinese imperial states and other sovereign governments, such as Fairbank (1968)'s model of the Chinese world order and further substantiated by the works of Wills (1984) on the Dutch and Portuguese envoys to the Qing and others[7]. Less attention has been paid to how intra-state tribute cemented connections between the Chinese imperial centers and their subject populations, particularly in borderlands. The system of intra-state tribute has been analyzed mainly based on its conceptual operation according to administrative regulations and what factors were anomalous in isolated cases of exceptional outcomes. This research is a longitudinal study focusing on the results rather than the aims and procedures of tribute in a particular eco-cultural area. It may demonstrate that even routine transactions were distinctive and revealed certain patterns of behavior that could deviate from the legal guidelines.

This research project may modulate an imbalance of awareness about two different methods of intra-state tribute. Presentation of tribute by Qing subjects to the imperial court occurred in two principal forms, the first one being missions from a region to the capital (Rossabi, 1998, pp. 256-257, 408-429; Serruys, 1955, p. 1819). Tribute visitors to the capital were treated as honored guests, and as diplomatic representatives invested with the power and obligation to negotiate on-going relations. A second form that applies particularly to the Heje and other populations from Heilongjiang was the collection of tribute goods and subsequent disbursement of “return gifts” at designated locations within the region, where tributaries would deal with imperial military and civil officials who were overseeing the logistics instead of paying their respects to the emperor directly. In both types of tribute, the state acted as a broker of goods, effectively purchasing tribute items and paying in things that the said tributaries could not otherwise acquire.

Moreover, tributaries and the officials who were monitoring their compliance with the tribute quotas conducted private trade, and tributaries sold goods to merchants without any political fiat. However, since the conventional understanding of Chinese imperial tribute stresses the first form over the second at all but the topic specialist level of scholarship, this research shows that the latter is equally important to analyze in order to form a complete picture of the tribute system.

Literature Review

The intra-state tribute system in Heilongjiang concerned several divisions of the Qing government that raised and solved problems about its operations through written correspondence. Most of the documents about these issues are composed in standardized, bureaucratic language, but are formatted as prose rather than as tables and charts organized by clearly distinguished variables. Therefore, utilizing these sources requires extracting data, identifying common variables, and then sorting the data into variable-based groups. Although it is possible that this process has been undertaken as a step of one or more research projects, no publication to date has presented the data as a discrete set with defined parameters.

To contextualize the methodological aims of assembling this dataset, since no similar precedent exists, the authors considered its broader significance as a means to promote sharing and co-compiling after its initial construction. The dataset is most useful to persons conducting research on the characteristics of the tribute system in Heilongjiang during the Qing dynasty, so the anticipated target audience would be historians of China, but could also include sociologists, anthropologists, and economists interested in this historical institution. The principal investigator, trained as a historian, determined that the first version of the dataset would be oriented towards the anticipated needs of other historians, and then be modified subsequently for and by users in other disciplines.
In searching for relevant prototypes in history, the authors discovered that like the actual content of the dataset to be made, the issue of how to design datasets for historical research has been underexplored. Since data management is not a compulsory subject for the majority of students in history, from the undergraduate to doctoral levels, creating datasets is not an expected component of their research aptitude. Indeed, Head (2008, p. 429) has observed most humanities and social sciences undergraduates, including history students, adopt a “coping” strategy to gathering sources for research rather than applying sophisticated and discipline-specific methodology to achieve that task. Ogburn (2010, p. 244) states that scholars must not only know how to utilize but also to produce digital tools such as datasets. A major obstacle in correcting this shortcoming, however, is that in history, as in other disciplines, many faculty are able to discern the errors in how their students approach data management, but do not take proactive steps to teach appropriate methodology because they themselves do not have adequate knowledge of that subject (Carlson, Fosmire, Miller, & Nelson, 2011, p. 636). Therefore, the process of developing a dataset can be a teaching and learning experience to enhance information tool literacy and primary source literacy for both faculty and students (Archer, Hanlon, & Levine, 2009; Townsend, Brunetti, & Hofer, 2011). With this premise in mind, the authors structured their research to focus on the technical aspects of creating a tool prototype that addressed the challenges inherent to managing the data in question, pre-empting problems that users might encounter in using the compiled dataset, and constructing the tool to be used as either a scholarly or pedagogical resource.

**Research Questions and Methodology**

Several previous works of scholarship have cited discrete data for the specific empirical case of Qing-period tribute in Heilongjiang, some identifying patterns based on multiple transactions. However, no single article or monograph has revealed broad quantitative trends or the full array of items transferred in these inter-actions. Creating a comprehensive primary source dataset obviates the re-aggregation of the same data and thereby reduces the time and effort for preliminary research by scholars working on related projects. Incorporating the ability for scholars to contribute to the dataset will not only encourage the formation of a research community but also provide a pedagogical opportunity for teaching students primary source and data literacy.

Thus, the authors posited the following questions:

1. What kind of database system is needed to compile, access, disseminate, and add to the dataset?
2. What are the key elements that comprise the dataset and allow for cross-analysis?
3. How does the multi-lingual characteristic of the dataset affect usability?

The authors organized the research process into three steps. The first step was to review of primary sources, data extraction, and data organization. For the first stage of the project the authors reviewed published collections of documents from Qing official repositories stored in national and regional archives of the People’s Republic of China, and notes from unpublished documents. These sources demonstrate an inherent bias in that they were one-sided, reflecting how Qing officials conducted and perceived of the tribute transactions. The records could not be verified by comparison with texts produced by the tributaries because such information is not readily accessible since some tributaries did not produce contemporaneous records; if they did, such evidence was not preserved in lasting formats. This step took approximately one month and was based on a bibliography compiled for the general study on Heilongjiang tribute. See Appendix 1 for sample sources.

The second step was to extract relevant data from these sources. Some information was explicitly articulated
in the originating source. One such example is the Qing Imperial Album of Tributaries (Huang Qing zhigong tu), a 9-fascicle encyclopedia compiled from 1761 to 1805 under the order of the Qianlong emperor. The Album identifies each of the listed Heilongjiang tribes by the goods that they were required to submit for tribute[8]. Other sources contain details about particular tribute and return gift transactions within documents on multiple topics. To identify pertinent information from these materials, the authors conducted an exhaustive search by reading each line of these texts. Relevant evidence was isolated through filtering keywords such as marten pelt (diao pi), fox pelt (hu pi), and sturgeon (xun huang yu), which were commodities frequently submitted as tribute items. This stage lasted four months intensively and then intermittently during the next step of the research process to confirm or to add details.

The third step was to organize the data. The raw information extracted from sources during the second stage had to be sorted by variables, which were refined according to the varying commonalities in the sources, and then as necessary, reproduced in standardized formats to provide consistency of quantitative and qualitative details. The authors spent one month arranging, re-assessing, and modifying the basic in the course of introducing content into the dataset.

**Technical Design**

There were three priorities in selecting the hardware and software components for creation of the dataset. The first priority was to ensure ease of usability in manipulating datasets. Not only was it important for users of all abilities to retrieve data successfully, but also for a potential contributor to be able to add content without developing proficiency in a particular computer language or undergoing extensive training in a sophisticated software program. The second priority was to ensure the ability to transfer the local prototype to a server-based setup such as a LAMP (Linux-Apache-MySQL- PHP) stack. The third priority was to ensure the ability of researchers to transfer data from spreadsheet files due to their familiarity with software programs such as Microsoft Excel for data collection and organization. Therefore, the first form of data compilation was the creation of a spreadsheet featuring all the variables that would be eventually included in the dataset framework.

The authors took these priorities into account when assessing and testing different types of database management software, as listed in Table 1 according to the order in which they were tried with concise evaluations of their attributes. Initially, they started with an XML/XSLT test page to examine the different ways to search across disparate records. Although the presentation was effective, the entry method was not ideal for accommodating multiple languages. The next logical option was to create a PHP form for entering records into a MySQL database. However, as the pedagogical objective of this project was to develop a database that would be simple for students without programming training or experience to design, the authors did not pursue this strategy.

<table>
<thead>
<tr>
<th>Software</th>
<th>Strengths and shortcomings</th>
</tr>
</thead>
<tbody>
<tr>
<td>XML/XSLT</td>
<td>Effective presentation abilities, non-ideal entry method for language requirements</td>
</tr>
<tr>
<td>PHP/MySQL</td>
<td>Common standard option, but users must possess specific technical knowledge</td>
</tr>
<tr>
<td>Microsoft Excel</td>
<td>Familiar software program with limited graphical presentation options</td>
</tr>
<tr>
<td>Microsoft Access</td>
<td>Robust database program requiring extra training for configuration</td>
</tr>
<tr>
<td>FileMaker Pro</td>
<td>Program is familiar to primary researcher and meets the desired requirements</td>
</tr>
</tbody>
</table>
Given the lack of programming personnel, the authors re-examined a familiar data entry tool for historians: Microsoft Excel. Unfortunately, it did not present records as effectively as a web interface might. Microsoft Access, a more robust database program, was considered but required additional training on part of the authors to use. Finally, the principal investigator selected File Maker Pro based on her previous experience in working with that particular software program. Additionally, FileMaker Pro offered the following:

- Compatibility with Microsoft Excel
- The ability to transfer data from an Excel spreadsheet into the FileMaker Pro document, matching cell for cell
- The ability to view entries like single, discrete file cards, or in lists as in Excel spreadsheets
- The option to use pull-down menus and radio buttons to simplify the input procedure, obviating repetitive typing of frequently occurring content
- Design options for layering two sets of overlapping windows, one for Chinese and one for English
- Accommodation of Chinese characters in input and display (no distortion because of problems with font or code)

The authors selected variables for the cells that would be arranged on the basic canvas of the FileMaker Pro platform based on the information needed for the research for which the dataset would first be used and what details are available in the sources. Essential elements included:

- Date of transaction according to the Chinese imperial calendar as signified by reign title, year, Chinese (lunar) month and day
- Date of transaction according to the Gregorian calendar
- Location of transaction (administrative unit names)
- Giver(s) of the tribute objects
- Recipient(s) of the tribute objects
- Each item and quantity thereof
- Title of source document
- Full bibliographical reference for source document

These elements were needed to carry out qualitative analyses of what kinds of items were exchanged, and the relative quantities of those items to determine their value. The dual date system is important because historians studying China often have to convert the dates from the original imperial calendar ones into Gregorian ones so that present-day readers can understand them. Yet it is often important to retain the original dates so that one can see patterns in reign eras such as fluctuations during the reign of the Kangxi emperor (1661-1772) compared to what happened in the subsequent Yongzheng reign (1722-1735). The patterns of tribute, such as the quotas being due in the spring and autumn, were also based on the imperial calendar, not the Gregorian one. As for location of tribute, it was important to show that transactions were occurring outside of Beijing, where tribute was thought to have happened normally. If the calendar element was not included, the myth that tribute was a highly ritualistic practice with rich political meaning might be perpetuated. This interpretation might be pertinent to certain situations such as the exchanges between the Qing court and emissaries from foreign governments, but not for these tribes, of which the transactions happened in military garrisons out in the woods of Heilongjiang. The identification of givers and recipients is also critical especially for the latter because the data could reveal that whole tribes or clans were sometimes noted; whereas in other cases, specific individuals, leaders and representatives, were named.

The physical arrangement of these elements can be seen in the following screen shots. Figure 1 shows the search function pages in clear forms. Figure 2 is the Chinese and English versions of the same discrete entry,
and Figure 3 contains reign names for the Qing emperors as well as translation of conventional measurement units utilized during the Qing dynasty such as the weight unit of catty (jin) with contemporary ones such as kilograms.

Figure 1 Chinese and English search pages side by side

Figure 2 Chinese and English sample entries
Some aspects of producing the Chinese and English concordant pages simply required adequate working time and consultation of widely available references. The dates of the transactions were rendered according to the Chinese lunar calendar with the imperial reign name and year, and in the majority of cases, also with exact months and days specified. The use of a computer-based tool, the Academia Sinica’s Two-thousand Year Sino-Western calendar converter (Zhongyang yanjiu yuan liang qian nian Zhong Xi zhuang huan) [9] expedited the translation of these details into Gregorian calendar equivalents. Many location names used the established English translations, and most of the transacted items and sources were similarly facile to record in both Chinese and English.

The first version of the dataset completed by the termination of the project funding period fulfilled some but not all of the initially planned goals. It contains 759 entries and is searchable in both Chinese and English using one or more of the core variables. The appendix guides users who are not familiar with the Qing period imperial reign titles and measurement units. Any computer with FileMaker Pro software, versions 7.0 and later, can be utilized to open and to manipulate the file. The authors have extracted the data related to the tribute study that motivated the creation of the dataset.

In general, the design of this prototype was based upon the trend of the primary researcher’s students seeking quick and easy glimpses of relevant entries, primarily for specific elements, as well as aggregation of qualitative and quantitative data for use in research. Upon completion of the prototype, the authors began planning for user testing with the goal of determining whether it would be effective in meeting the design goals. Based upon the results of the user feedback, the authors intend to seek the relevant experts to construct a more robust prototype.

**Results and Discussion**

By the end of the prototype creation, the authors were able to answer the research questions by creating a FileMaker Pro database structured on the eight key elements discussed above. At the same time, the authors encountered critical problems with technical issues, translation complications, and management of multiple sources. One example was that certain multi-phase...
transactions were captured in single records rather than in separate ones to show the discrete circumstances of each phase. Another was the inability to correctly reproduce Manchu script. The most significant challenge, however, was the proper transliteration of tribal names. Most of the primary sources consulted in this project were publications in Chinese, which are translations of Manchu original documents. Neither of the two Romanization conventions, the Hanyu pinyin system used for Chinese or the Möllendorf system for Manchu, could be used for precise transliterations of terms in the tribes’ original languages that do not have indigenous phonetic scripts. Therefore, quite a few of these proper nouns, whether the names of tribes, places, or persons, were entered into the dataset in modified forms, aligned with Chinese or Manchu transliteration standards. Since the original languages belong to the Tungusic language family like Manchu and absorbed many loanwords from Mongolian before and during the Qing dynasty, the authors consulted reference works on Manchu and Mongolian names (see Appendix 2 for examples).

The diversity of data was also difficult to present in an optimally manipulable format. The graphical user interface was not sophisticated enough to inspire creative manipulation, in the sense that users might not be motivated to look at one entry and start a new search based on something discovered in the prior search. Rather they would probably be looking for a specific detail, find it, and then discontinue the search. This tendency stops users from accessing more valuable search feature such as the ability to search for a tribute or return gift translation which could be found in one source document but mentioned in other documents comparable or lesser detail. For instance, some sources would mention the names of all individual tributary envoys; whereas in others, they would be listed collectively as a group identified by their place of origin or the state that they represented. In certain cases where evidence could not be confirmed as descriptions of the exact same event, they formed individual entries in the data set. The authors expect that with more corroborating details and scrutiny for typographic errors or ambiguities, searches across entries may be reconciled in the future so that all relevant entries will emerge in a single query result or be verified as separate but inter-related events.

**Future Developments**

In carrying out this project, the authors expect to reconsider certain assumptions that are prevalent in the current practice of historical research. The first such notion is that datasets are created and utilized by a single scholar or an exclusive group of co-researchers. The second, unlike in other social sciences disciplines, students of history are not required to undergo training in dataset management. Moreover, students just utilize many datasets in a passive manner by retrieving information rather than actively contributing new content to them.

Conceptually, this dataset could be a tool that engenders new habits in data management by increasing the pool of users who could also be contributors. The first would be to expand the dataset into a more web-friendly format that would be maintained through crowdsourcing from the academic community. Users would be welcome to add new entries that would be vetted by a moderator for correct style and appropriate content. Adding a GIS (geographic information systems) component to the datasets would also engage both experienced scholars and students alike. This level of engagement would also strengthen its potential as a pedagogical tool for introducing students to the process of compiling primary sources and managing their data for others to use.

To fulfill this potential, many technical improvements remain to be made. The database must be converted from a FileMaker Pro file to a more robust structure, either through digital repository software or a customized solution. At present, the dataset is only available by request to the authors. Online conversion may also include modifying the file so that it can be opened without having FileMaker Pro software installed on the receiving computer.
The online version should furthermore be designed so that users can contribute content under the administration of the authors. User guides will be provided to help new users browse and search the dataset. Moreover, the dataset content must be refined in four key ways: 1) link creation between related entries, 2) an increase in the number of elements included with each entry, 3) the ability to include images or symbols, and 4) a Manchu-language interface.

All of these alterations will enhance the dataset’s capacity to broaden the scope of research and pedagogical collaboration, which is both a key purpose and intended outcome of the project. The authors have created a rapid prototype for what could potentially become a communal online repository. An existing example would be the eBird[10] website which collects bird observations from users all around the world. This data is then compiled into a usable format for analysis and other forms of re-use. It also incorporates other features to engage users and encourage them to return. The authors hope to expand the dataset to multiple databases to collect data from other scholars over time. This sort of collaborative environment would lend itself well to pedagogical purposes and increase dialogue between historians who share a common research interest but have access to different sources. They can contribute in various ways while making use of the aggregated data.

Conclusion

In pursuing this project, the authors encountered a number of topics common to both the history and archival/library science disciplines. The completed dataset satisfied the needs of its creators’ research, generating both quantitative and qualitative results that could be analyzed individually and collectively. Each entry provides enough evidence for a given transaction so that it can be explored from either the perspective of the giver or recipient, the items’ quantities, and their relative values. This information, as displayed in the dataset, highlights a more multi-faceted and multifocal approach to understanding the tribute system. Furthermore, the project generated new questions regarding accessibility, discovery, and dissemination of data for the inspiration of new research. The potential for merging pedagogical practice with pure scholarly activity through dataset development may emerge as a future direction that can be explored for historical research. As historians expand the parameters of acceptable source materials, they can process increasing amount of evidence effectively by building datasets that can be enhanced by other users through web-based digital channels. Historians can produce outputs more efficiently to formulate their interpretations based on larger pools of evidence. Students training in history will also be benefited by being able to pursue more data-rich research.

Notes

[1] “Ch.” and “Ma.” will be the abbreviations for “Chinese-language” and “Manchu-language” respectively if the same concept is presented in both languages. Non-English, italicized terms without specific abbreviations are Chinese words.

[2] The term “tribute” has most commonly been used in the study of China’s imperial age as a multi-dimensional concept of diplomacy and trade, as defined by the work of John King Fairbank and Ssu-yu Teng in “On the Qing’s Tributary System,” Harvard Journal of Asiatic Studies, 6(2), 135-246. Studies published in the past decade have explored intra-state tribute as a means of creating economic and social links between the Qing heartland and frontier zones, such as C. Patterson Giersch, Asian Borderlands: the Transformation of Qing China’s Yunnan Frontier (Cambridge, Massachusetts: Harvard University Press, 2006).

[3] Owen Lattimore was the pioneering thinker of the “Inner Asia” or internal frontier concept in the historiography of early modern China. See Inner Asian Frontiers of China (London and New York: Oxford University Press, 1940). The Sino-Russian frontier is a
well-developed research area. Paradigm-setting works in English include Mark Mancall, Russia and China: their Diplomatic Relations to 1728 (Cambridge, Massachusetts: Harvard University Press, 1971) and S.C.M. Paine, Imperial Rivals: China, Russia, and their Disputed Frontier (Armonk, New York: M.E. Sharpe, 1996).

[4] Filemaker Pro is a database management system developed by FileMaker Inc. (formerly Claris).


[7] See Fairbank, The Chinese World Order, and Wills, Embassies and Illusions. The sum total of the literature on inter-state tribute is too great to represent with just a few citations, but the bibliographies of these two works give some insight into the breadth of work on this topic.


[10] http://ebird.org/content/ebird

References


Appendix 1


Appendix 2

