Journal of Chinese History (2020), **4**, 251–257
doi:10.1017/jch.2020.29

CAMBRIDGE
UNIVERSITY PRESS

INTRODUCTION

Introduction to the Special Issue

Peter K. Bol*

Harvard University
*Corresponding author. E-mail: peter_bol@harvard.edu

This issue of the *Journal of Chinese History* aims to take stock of the effects of "the digital" on the study of Chinese history. We are doing this through a combination of research articles whose authors have made extensive use of digital resources and technologies and a set of introductions to non-commercial, open-access utilities and tools that scholars have created to support research in a digital environment. These hardly exhaust the universe of research or tools.

When the materials with which humans found their way through time and space and communicated with others were in physical media, they could be collected, curated, and preserved for future research. When the writing and imagery became digital as well, they became ephemeral but also more accessible, and more people than ever before could become producers. The first website was created in 1991; today there are almost 2 billion of them and they are being accessed by 4.4 billion users. There are blogs and wikis, and over 20 billion text messages are being sent daily.[1] By one reckoning the entire digital universe is expected to reach 44 zettabytes (10^{21} bytes) in 2020, which would be forty times more bytes than there are stars in the observable universe.[2] We can be sure that the digital will not become less important. It is also playing an ever-larger role in the study of the pre-digital past as well.

Digital scholarship uses sources that are in digital form. Several things follow from this: sources in digital form are easily altered and manipulated, they can be treated with computational procedures (algorithms) that allow massive amounts of data to be processed very efficiently and they can be shared widely at minimal marginal cost.[3] This is an historical event that can be tracked over time and one that has been affecting the research cycle during the academic careers of the editors of this journal.

Research begins with questions. As the digital has advanced it has affected the ways in which we know what is going on in a discipline, but has it also had an effect on the questions we ask? My reading of the research articles in this issue suggests three answers: it makes it easier to address old questions by taking more information into account; it makes it possible to ask questions from multiple perspectives, for example

[1] https://blog.microfocus.com/how-much-data-is-created-on-the-internet-each-day/ and https://www.websitehostingrating.com/internet-statistics-facts.

[2] https://www.visualcapitalist.com/how-much-data-is-generated-each-day/.

[3] See the discussion in "Interchange: The Promise of Digital History," featuring Daniel J. Cohen, Michael Frisch, Patrick Gallagher, Steven Mintz, Kirsten Sword, Amy Murrell Taylor, William G. Thomas III, William J. Turkel, *Journal of American History* 95.2 (2008), 452–91

to see information in terms of both social networks and spatial distributions; and it opens up questions we had not previously attempted to answer.

The search for sources—both the search for trends in a discipline and for primary sources—once took place in the trays of the card catalog. Today the open public access catalog is the norm, but it only began in the late 1970s; the catalog of today that allows faceted search is much more recent.[4] Harvard's HOLLIS catalog, the largest academic catalog in the world, did not display CJK scripts until 2002, fourteen years after the system launched. Today that catalog can link users to sources that are online and programs such as Endnote and Zotero can automate the extraction of bibliographic data from online catalogs and websites. The advent of searchable text is the single most important development in digital scholarship. Scripta Sinica, from the Institute of History and Philology at Academia Sinica began in 1984 and continues to set the standard for accuracy.[5] Today there has been a proliferation of searchable text corpora from commercial vendors, much to the disadvantage of scholars without access to the handful of major research libraries, but also great open-access repositories such as The Chinese Text Project (ctext.org), discussed in this issue.

Ways of organizing and storing information have changed as well. Excel, the most popular commercial spreadsheet, did not appear until 1985, although the first electronic one goes back to 1969.[6] The functions built into today's spreadsheet programs have more capabilities than most historians are aware of. The relational database, which allows many tables to be joined together and facilitates complex queries, was first conceived in 1970 but the first commercial program came out in 1979.[7] Since then new kinds of databases have appeared: object databases, which represent information as objects in contrast to the tables of relational databases; and graph databases, which represent data through networks of nodes and edges.[8]

I have used the word "information," but computational methods work with "data." What is the difference? The sentence "Zhu Xi was born in the year 1130" is information. It is composed of three data (a name, an event, a date) which can be arranged as a row in a table, with two columns or variables (name or person, year of birth), or as a relationship between two nodes (a name and a year), with "birth" as the edge or link between the nodes. Information is translated into data to fit a data structure so that the data, when queried can become information again, for when taken alone the datum "1130" is merely an integer. Understanding the relationship between information and data is important when storing information in spreadsheets or databases, because the data structure defines the ways the data can be queried: the structure is what makes it possible to call up all the names of people in the dataset who were born in the year 1130.

The key to transforming information into data is the ability to identify, to "tag," elements in a text. This can work at a structural level (sentences, paragraphs, parts of speech) and at an information level (places, dates, persons). The Text Encoding Initiative (TEI) released its first guidelines for marking up text in 1990. It was soon

[4] https://en.wikipedia.org/wiki/Online_public_access_catalog (accessed September 24, 2019).

[5] http://applyonline.ihp.sinica.edu.tw/english/source/source6.htm. For information about all collections see http://hanchi.ihp.sinica.edu.tw.

[6] https://en.wikipedia.org/wiki/Spreadsheet (accessed September 24, 2019).

[7] https://en.wikipedia.org/wiki/Relational_database (accessed September 24, 2019).

[8] https://en.wikipedia.org/wiki/Object_database (accessed September 24, 2019) and https://en.wikipedia.org/wiki/Graph_database (accessed September 24, 2019).

applied to the Tripitaka by the Chinese Buddhist Electronic Text Association; a guide to the use of TEI with Chinese texts appeared in 2009. The idea of searching text for patterns of expression, such as dates, goes back to the 1950s, but in practice its use in the humanities is more recent. The Markus utility, described in this issue, allows the user to upload text, tag elements in it, and then download the tagged data.

The translation of information into data together with access to databases through the internet, makes it possible for databases to share data with each other through "application programming interfaces" (APIs), which only appeared in 2000. Almost all the utilities built for Chinese studies discussed in this issue make use of APIs, thus greatly increasing their utility by allowing knowledge distributed across different systems, collected for different purposes and managed by different people in different countries to be joined together.

The tools for **analyzing** digital information and data have proliferated. There are software packages for many kinds of cluster analysis.[9] Topic modeling, the use of "machine learning" to discover the set of topics inherent in a text corpus, first appeared in 1998.[10] Some text databases combine repository functions with analytic functions, enabling frequency analysis, comparison between editions, discovering nearest neighbors, and so on.

There is an argument that those who want to mine and analyze texts extensively need to consider learning to program, for which there are now online courses and lessons for a wide variety of analytic methods.[11] For many the free software packages for statistical analysis, spatial analysis, and network analysis are adequate. The first desktop geographic information system appeared in the 1980s.[12] Systematic social network analysis dates from the 1950s but it seems that software packages first appeared in the 1990s.[13]

Finally, digital modes of **dissemination**, the last part of the research cycle, have become ubiquitous. There are differences between print and digital editions: full color illustrations, zoomable maps, and sound files, for example, do not entail extra costs when digital. This introduction has been able to cite Wikipedia or online sources for all but one of to this point. Blogs, wikis and open-access online journals are commonplace. Individuals can publish through personal websites and the work is findable with web searches. Publishers are now experimenting with digital projects in addition to books.[14]

The first article in this issue is a theoretical essay from a scholar of Song literature, Michael Fuller: "Digital Humanities and the Discontents of Meaning." Fuller is also an experienced programmer and the chief data architect of the China Biographical Database. He argues that the digital humanities offer a way out of what he calls isolated subjectivity, when the meaning of things is whatever I as interpreter make it to be, and the hermeneutics of suspicion, when the meaning of things is revealed through deconstruction to be in service of ideology and power. The pragmatic study of language and cognition instead treats the structuring of human experience as an object of scientific

[9] https://en.wikipedia.org/wiki/Cluster_analysis.

[10] https://en.wikipedia.org/wiki/Topic_model.

[11] https://programminghistorian.org/en/.

[12] https://en.wikipedia.org/wiki/Geographic_information_system.

[13] https://en.wikipedia.org/wiki/Social_network_analysis and https://en.wikipedia.org/wiki/Social_network_analysis_software.

[14] See, for example, Stanford University Press's new series: https://www.sup.org/digital/.

inquiry. His discussion of distant and close reading as the basic strategies of the humanities begins with the mathematics of topic modeling and ends with biographical data by way of the hermeneutics of Wittgenstein, Dilthey, and Schleiermacher. The argument is in many ways straightforward: the meaning of a text is relative to the textual context in which it takes place and relative to the life experience of the writer. Distant reading—the term comes from Franco Morretti—is a way of finding in the universe of texts the patterns, structures, and clusters that the particular text that is being read closely invokes.[15] Fuller cites Schleiermacher's observation that "As every utterance has a dual relationship to the totality of the language and the whole thought of its originator, then all understanding also consists of the two moments: of understanding the utterance as derived from language, and as a fact in the thinker." His discussion of topic modeling reveals the logic of distant reading; there are many methods of analyzing very large textual corpora.

The hermeneutics of distant reading that explores texts within the context of textual corpora, Fuller argues, is complemented by an analogous approach in the study of people. In designing the China Biographical Database he sees that the way in which a relational database models life patterns does for historical figures what distant reading does for texts. That is, it makes it possible to understand "the larger life patterns within which an individual (or an era) lives." I would add that a database created for seeing the patterns of many lives should not be treated as a biographical dictionary.

The "distant reading" of biography is at the heart of Nicolas Tackett's "The Evolution of the Tang Political Elite and Its Marriage Network." Building on his earlier work on the demise of the Tang great clans, Tackett compiled kinship data from excavated epitaphs and the *New Tang History* to construct a dataset of 34,158 father–child kinship ties to explore the marriage network of political elites, the backgrounds of chief ministers, the composition of the capital elite in early, middle, and late Tang, and the composition of the provincial elite. Once the data was disambiguated—that is, determining whether two or more entities with same name referred to one or more people—he was able to construct patrilines over multiple generations as an alternative to relying on claimed choronyms (*junwang* 郡望). The one large network of political elites to which he pays particular attention contained eighty-seven patrilines, with a total of 14,444 individual members and 384 marriages. Locating individuals in these networks offered a way of knowing more about persons for whom other data was missing, such as their home base or the type of family they would marry. The distant reading informs the close reading.

Tackett combines network analysis to show kinship ties, spatial analysis to show where the political elite was based, a typology of elite epitaphs (e.g. no office, civil, military, etc.) and temporal analysis to show change over time. His results both confirm and challenge received views. Tang factions were indeed regional and the Luoyang elite began with fewer ties to the state, but Empress Wu was not in fact bringing in newly risen men but tapping well-established patrilines. After the An Lushan rebellion capital patrilines were ever more dominant at court, but the ties between them and the provinces dissipated, and provincial office-holders became provincial elites who made their careers in territorial governments.

In the past an apt quotation or anecdote might suffice to make a point, and readers could check references in the footnote. However, Tackett's findings and those of all the

[15] Moretti, Franco, *Distant Reading* (London: Verso, 2013) and *Graphs, Maps, Trees: Abstract Models for a Literary History* (London: Verso, 2005).

other research articles depend on datasets that are derived from the primary sources. To say that the work has been checked has two levels of meaning: that the analysis of the dataset is replicable and that the data itself is correctly derived from their sources. Both levels require that the dataset be made available, either to replicate the analysis or to check the data sources. Due to copyright issues the original sources usually cannot be made available. Currently there is no standard protocol for making datasets available and authors have taken different approaches. Tackett makes his data available through a personal website, Chen and Chang through Dataverse, De Weerdt through DANS (Digital Archiving and Networked Services), and Schäfer explores ways to publish her data in connection with the original sources that are kept in commercial databases.

Tackett worked with all the epitaphs he could find. Song Chen in "Writing for Local Government Schools: Authors and Themes in Song-Dynasty School Inscriptions" uses all the inscriptions for Song dynasty state schools in the digital edition of the 360 volumes of the *Complete Song Prose*, which gives him an author, a title, and usually the administrative unit. Chen aims to quantify influence.

Two aspects of his approach are particularly interesting. First, his network analysis is between inscription authors and regions rather than solely between authors. Authors are from somewhere and write for a place—is it in their native region or not? Writing for schools in more regions is a sign of greater influence. Using GIS to show where the schools with inscriptions are and the various measures of network properties built into social network software, he points out the relative isolation of the Upper Yangzi (Sichuan) but also the importance of those who, by writing for schools in different regions, bridge them.

Second, he uses distant reading approaches to analyze the content of the inscriptions. One technique is to search for key terms across the dataset to show change over time, but key word search (leaving aside the problem of word segmentation in literary Chinese) brings with it the danger of only validating or nullifying the researcher's specific hypothesis without providing an opportunity for discovering something new and unexpected. Of course we can know something about the authors, and this is also a way of differentiating likely content. Chao Buzhi, the author of numerous epitaphs and inscriptions, was a follower of Su Shi, in contrast to the even more prolific Neo-Confucian Zhu Xi. But authorship does not necessarily equate to content. Chen uses "document clustering" to show that there are three distinct thematic clusters of inscriptions and that an author could write in more than one mode. Chen's explanation of his procedure illustrates in practice the mathematical introduction given by Fuller.

In "Is there a Faction in this List?" Hilde De Weerdt and her collaborators take up three Song dynasty lists from 1104, 1142, and 1198 that were purported to represent factions and asks if they really were factions and, if so, how they were organized. The debate on Song factions they review suggests that we may not want to assume that because some said others formed a faction (and thus should be pushed out of court) that in fact they did. To get at this they adopt a novel strategy. They could have taken up those on the list and used their writings to see how involved they were with each other, but this skews analysis towards those who wrote the most. So instead they ask which names co-occur in a given piece of writing (a letter, an epitaph, etc.) in all the extant writings by persons from the time of the lists as found in the *Complete Song Prose*. This is a study of shared perceptions, as it does not matter if the writer had anything to do with a given list, only whether they thought there were relationships between people on the lists.

The amount of data involved in this procedure, from over 30,000 to over 50,000 co-occurrences, makes computational analysis essential. The results do not surprise, given changes in politics from Northern to Southern Song—those on the first list comprised a dense court network, the second list was not a group at all, and those on the third list formed conglomerations of central and provincial figures. What is interesting and useful is the method they developed to compare historical networks and sample them to test their hypotheses. By working statistically and using different ways to compare, analyze, and test the data they arrived at nuanced conclusions that recognize both different kinds of connectivity and their absence.

The kinds of social network analysis required to show this are sophisticated. Digital scholarship often involves collaboration between scholars with different kinds of expertise. In addition, the databases and platforms used (in this case CBDB and Markus) are collaborative efforts. I think this is worth mentioning because many humanists assume that we must be able to do everything ourselves, and there are indeed rare figures who can, but mastering the tools so as to use them fully and wisely is sometimes a career in itself. Collaborate!

The collaboration between the historian of science Dagmar Schäfer, the digital humanist Shih-pei Chen, and the historical geographer Qun Che combines an historical inquiry into the reporting of natural disasters related to silk production during the Yuan dynasty and the use of the LoGaRT platform for automating data collection from local gazetteers (the platform is introduced in the section on utilities). This is a good example of combining different scales and different modes of analysis. The local gazetteer, well on its way to becoming the ubiquitous form of local history during the Yuan, has generic patterns; it lends itself to becoming a database. Although there are considerable differences in how information is presented in gazetteers, it is possible in LoGaRT to search across the entire corpus—another use of "distant reading." The authors wish to explore how analytical procedures used to interpret one source reflect upon the reading of the entire corpus, and how quantitative interpretations can direct us in our reading and analysis of the particular case. Combining statistical and visual analyses on data to discover general patterns and local anomalies and historical analysis that closely examines the contexts of how data were produced, this article reveals that reporting disasters were interactive results of the historical protocols for reporting information, individual editorial decisions, and the specific conditions of a place, all of which must be taken into account. To see only at the corpus scale is not better than treating an anecdote as representative of the whole.

In their contribution Chen, Campbell, Ren, and Lee introduce the China Government Employee Database—Qing (CGED-Q). The sources for this data are the quarterly *Jinshen lu*, recording government offices and their incumbents from the mid-eighteenth century to the end of the Qing dynasty. The current database, with over three and a half million observations from 1850 to 1912, is the most important dataset for the study of the institutional and social history of Qing officialdom. This article introduces only some of conclusions that the authors are drawing from this database, showing how the qualifications that gave entry to office changed over time and how persons with different qualifications and different provincial origins were distributed across the bureaucracy. The availability of CGED-Q invites more quantitative research on Qing government.

The article also provides a very useful introduction to the challenges of creating a large, reliable database. Some challenges are predictable, such as collecting and understanding the sources and converting the information on the page into structured data.

The larger challenge, common to all large-scale collections of biographical data, is disambiguation. If every time we encountered the name of a person it was accompanied by a note telling us where he or she lived, birth and death dates, given and courtesy names, titles and offices, we would know immediately whether two persons with the same name in a historical source were one person or two. The China Biographical Database has sixty-two people named Wang Zuo 王佐 (King's Assistant), forty-nine for the Ming dynasty alone, but in many instances ambiguity remains (better to duplicate than treat two persons as one). Fortunately, the *Jinshen lu* includes the name of the place the official is from and how (and sometimes when) he qualified for office, so the disambiguation procedure was highly successful. It was far less successful in the case of bannerman, for whom the banner is given rather than the place of residence; reduced further by the frequent absence of Manchu clan names.

The final article, Charles Chang's account of the communities within urban Kunming, illustrates how new data sources can be used to study the recent past. Kunming was a small, provincial backwater in the 1930s; today it is a city of 6.2 million. How can we understand this growth and the make-up of the city today, given a dearth of official data or official data that does not correspond to ground truth? The solution lies in combining very large amounts of data from different sources. In addition to the record of remote sensing data freely available from the US government, there is data on points of interest from the Chinese commercial mapping companies (which follow government rules in distorting geographic locations), those social media posts from residents that are geo-tagged, electronic commerce data, and street view images from surveying companies.

Making large amounts of data from different sources, generated in such different ways, and offered at very different scales compatible is a challenge. Using his expertise in geographic information systems Chang builds several layers of data at different scales in such a way that each layer becomes an elaboration on the preceding one. Most of the data he uses to build this data-driven approach are themselves a byproduct of our digital era. The use of computational techniques to scrape websites, topic model hundreds of thousands of blog posts, and map geo-tagged data is becoming part of the toolkit of historians of today's world.

The articles in this issue of the Journal were written when librarians kept the libraries open, scholars were presenting papers at conferences, and students attended classes in person. The final editing took place in the spring of 2020, when we were practicing social distancing and quarantining ourselves at home. By the time this issue reaches readers the situation will have changed, but we might reflect on how we would have managed to keep on with academic life under such circumstances without digital assets and communication. The utilities introduced in this issue have provided databases, tools, and platforms that make scholarship "at a distance" possible. Although they were not written with this purpose in mind, the articles demonstrate some of the ways cutting-edge research can take place in a fully digital environment.

Cite this article: Bol PK (2020). Introduction to the Special Issue. *Journal of Chinese History* **4**, 251–257. https://doi.org/10.1017/jch.2020.29