Journal of Chinese History (2020), **4**, 483–527

CAMBRIDGE
UNIVERSITY PRESS

UTILITIES

Introduction to the Utilities

Peter Bol*

Harvard University
*Corresponding author. Email: pkbol@fas.harvard.edu

doi:10.1017/jch.2020.10

A variety of databases, tools and platforms have created the foundation for digital scholarship in Chinese studies. The creators of some open-access projects introduce their work below, but first I offer some notes on the kinds of utilities that make up the expanding digital universe.

Searchable text databases are the most widely used resource. Beginning in 1984 with the dynastic histories, Academia Sinica's Institute of History and Philology has set the highest standard for the creation of digital texts, and since then other institutions have followed suit. A proportion of Scripta Sinica is open access.[1] Since then, in addition to many commercial text databases, three other collections have been established that are entirely open access. CBETA, the Chinese Electronic Tripitaka Collection, began over twenty years ago.[2] The most recent is the Kanseki Repository of premodern Chinese texts, popularly known as Kanripo; it is overseen by Christian Wittern of Kyoto University and has 9,500 texts.[3] The largest is Donald Sturgeon's Ctext, the Chinese Text Project, currently holding over thirty thousand titles and more than five billion characters. The text repository is only one part of the Ctext platform, which includes a variety of tools, as Sturgeon explains in his introduction to the platform.[4] The most convenient way to discover whether a text is available in one of these three repositories (or the *Zhonghu jingdian gujiku* 中华经典古籍库) is to consult textref.org; icons show whether a text is open to view, search, or download and whether a scanned image is available. The scanned image is important because, as Sturgeon explains in the case of Ctext, the searchable text is created by applying optical character recognition (OCR) to the scanned image. Textref.org currently has 54,000 records and is an important contribution to the cyber infrastructure for Chinese studies. All providers of digital text, whether open- or licensed-access can choose to reveal the titles they have, without losing their proprietary rights.[5]

My thanks to Kwok Leong Tang for his suggestions.

[1] http://hanchi.ihp.sinica.edu.tw/. For an English language survey of digital resources from Taiwan see http://sinology.ascdc.sinica.edu.tw/index_en.html, http://thdl.ntu.edu.tw/index.html and www.digital.ntu.edu.tw/en/achievements.jsp.

[2] www.cbeta.org/.

[3] www.kanripo.org.

[4] https://ctext.org.

[5] https://textref.org/.

The Ming Qing Women's Writings Digital Archive and Database, discussed below by its director Grace Fong, is an example of **digital archive of selected writings with an online scholarly apparatus.**[6] MQWW is valuable for its collection of rare texts. It also illustrates how a database focused on a particular set of texts and people can make use of other online utilities. In this case MQWW uses an application programming interface (API) to call up information from the China Biographical Database, thus relieving it of the need to keep track of kin and social relations of the writers themselves.

The China Biographical Database (CBDB), discussed by two managers Wang Hongsu and Tsui Lik Hang, is an open-access **biographical database**, like the Dharma Drum Buddhist College (DDBC) Buddhist Studies Authority Database 佛學規範資料庫 and the Database of Names and Biographies 人物傳記資料庫 from the Institute of History and Philology.[7] One can discover persons in these databases through biogref.org, which currently has almost 520,000 records. All three databases provide categorized biographical data. There is an important difference, however, in that CBDB is a **relational database** composed of code tables and data tables. This allows it to be used in complex queries covering large numbers or persons. The existence of code tables for offices, people, places, and so on, also means that CBDB code tables, accessed through its APIs, can be used to mark up or "tag" texts on other platforms.

There are **specialized datasets** that provide code tables for tagging texts, such as the China Historical Geographic Information System (CHGIS), which also has an API to be used by online systems. Its most important use, as described in my "Visualization and Analysis of Historical Space" in this section, is the provision of data layers of administrative units for 2,000 years of China's history for use in GIS software. The CHGIS project also provides other valuable spatial datasets, including G. William Skinner's nineteenth- and twentieth-century datasets.

Platforms are online systems that allow users to upload their own data (or retrieve data that is already available on or through the platform) and use the capabilities of the platform to analyze the data. MARKUS, introduced by Hilde De Weerdt, is a platform that allows users to upload text and tag it, using code tables from CBDB, CHGIS, and other databases or by creating their own lists of terms. Tagging words in a text allows them to be extracted from the text and analyzed and visualized in diverse ways. Li Bin and his collaborators provide an example of this with their online system for the Basic Annals of Sima Qian's *Records of the Grand Historian*. In this case they tagged the text manually for person names and place names, thus allowing users to visualize the connections between persons and between persons and places statistically and geographically. Doing this with single texts manually is manageable but only an automated system would make this possible across large text corpora.

As Chen Shih-pei explains, the Local Gazetteers Research Tools (LoGaRT) is a platform devoted to extracting all manner of data from local gazetteers. Gazetteers are databases to-be: they contain structured information using common categories. But in fact there is just enough variation in the original texts to make this complicated. In her contribution to this section, on the composition of the Qing bureaucracy, Chen Bijia notes how challenging this can be in discussing the mining of the Roster of Appointments (*Jinshen lu*). The LoGaRt system is quite powerful, but the platform must be installed on a local server using texts available to that institution or users must arrange to

[6] http://digital.library.mcgill.ca/mingqing/.
[7] http://authority.dila.edu.tw/ and http://archive.ihp.sinica.edu.tw/ttsweb/html_name/。

spend time at the Max Planck Institute in Berlin. Currently most searchable gazetteers are in licensed text databases. PhiloLogic is another powerful platform or framework for text analysis to scale; when it is on a local server users can create their own instances with their own corpora. Jeffrey Tharsen and Clovis Gladstone explain its capabilities with the example of the *Twenty-Four Chinese Histories*; they have opened their instance to readers.

Various platforms provide various kinds of **tools for analyzing text corpora**. Ctext's text tools allow comparison between chosen texts, visualizing similarity and proximity based on statistical analysis, and more. In addition to the introduction to the Ctext platform and text tools for the analysis, the website offers detailed introductions and guides.[8] Some of the same capabilities are also part of PhiloLogic and they are being built into MARKUS. A different kind of platform is 10,000 Rooms, as introduced by Nicholas Frisch, which is aimed at enabling users to upload and annotate images of texts. This facilitates the study historical editions of texts that may exist in multiple woodblock editions.

Platforms typically allow users to upload texts for one-time analysis and, with registration, to store texts (taken from Ctext or Kanripo for instance) on the platform's servers, employ various tools, download data, and produce visualizations. The Digital Humanities Research Platform at Academia Sinica[9] and DocuSky from National Taiwan University[10] provide these services to registered users. DocuSky, first developed by Tu Hsieh-chang and discussed by Hsiang Jieh, is unusual in that it is a system created to enable users to transform the texts and spreadsheets they may have on their own computer into their own online database. It provides a specific XML format which acts as a bridge between content and tools. Users can convert texts and spreadsheets into the XML format and use them with the tools provided by DocuSky or other open access platforms, such as MARKUS. In addition it is meant to create a link between researchers and developers.

The text corpora, databases, datasets, code tables, APIs, tools, and platforms discussed here are examples of the kinds of utilities and digital resources available for the study of China's history. The list is not exhaustive. I have not covered the various software packages being used in data analysis. Of the several challenges in using digital resources I will draw attention here to one: the lack of a reliable means of segmenting words or phrasemes in texts written in literary Chinese. The lack of white space between words is a problem for East Asian texts generally, which is exacerbated by the lack of punctuation and ambiguity of the status of a string of characters as a word, although there is a parser for modern spoken Chinese.[11] There are open-access utilities that have had some success with punctuating literary Chinese and identifying parts of speech.[12]

The increasing number of searchable text databases, most of them commercial, presents researchers with new challenges. First, researchers would like to search metadata (that is, information about the text such as author, title, edition) across databases, even

[8] https://ctext.org/instructions; https://ctext.org/tools; https://ctext.org/digital-humanities.

[9] 數位人文研究平台 http://dh.ascdc.sinica.edu.tw.

[10] https://docusky.org.tw.

[11] https://nlp.stanford.edu/software/lex-parser.shtml#Tools.

[12] Long Quan Temple, punctuation: http://gj.cool/gjcool/index; Beijing Normal University, punctuation: https://seg.shenshen.wiki/; Academia Sinica, part-of-speech tagging: https://ckip.iis.sinica.edu.tw/service/ckiptagger/; Nanjing Normal University, part-of-speech tagging: http://47.100.116.59/suiyuan/index.php.

if the local library does not have a license to the content. The HOLLIS catalog at Harvard, for example, reveals metadata for those Erudition databases it licenses, although unaffiliated users cannot access the content. The CrossAsia Fulltext Search Catalog from the State Library in Berlin does this as well for the domain of Asian studies.[13] The lack of interconnectivity across the digital universe of Chinese studies, led to the 2018 Shanghai conference on "Cyberinfrastructure for Historical China Studies."[14] At this point there is no one agreed path forward, but there are several possibilities. The Max Planck Institute for the History of Science has developed the Research Infrastructure for the Study of Eurasia (RISE) which, through its API, is meant to enable institutions to create secure linkages between third-party research tools and various third-party textual collections.[15] The organizers of the 2018 conference, together with major libraries and research institutes in China and around the world, are working with the Chaoxing group to see whether a sophisticated, wide-ranging search, retrieval, and analysis system could be the basis for a common multi-lingual platform of open and licensed content.[16] Another approach, represented by the aforementioned textref.org and biogref.org, is for database providers to agree on a common standard for the basic metadata necessary to identify texts and individuals in their systems. The challenge is to build this into library and database workflow so that new data is entered automatically. A third option will take shape with the sixth and final edition of Endymion Wilkinson's *Chinese History: A New Manual*, to appear in 2021–2022. The *Manual* will then also appear as a curated online database that can continue to evolve, a kind of a hub whose spokes are links through APIs to library catalogs and other databases, at the same time that internal hyperlinks make it easy to explore the rich content of the book itself.

Digitizing Premodern Text with the Chinese Text Project

Donald Sturgeon

Durham University, email: donald.j.sturgeon@durham.ac.uk

doi:10.1017/jch.2020.19

Abstract
The widespread availability of digitized premodern textual sources – together with increasingly sophisticated means for their manipulation – has brought enormous practical benefits to scholars whose work relies upon reference to their contents. While great progress has been made with the construction of ever more comprehensive database systems and archives, far more remains not only possible but also realistically achievable in the near

[13] https://crossasia.org/de/service/crossasia-lab/crossasia-itr/.

[14] The program and other materials for the conference are available on the Ctext website: https://ctext.org/digital-humanities/shanghai2018

[15] https://rise.mpiwg-berlin.mpg.de/

[16] With thanks to the support of Mr. Shi Chao the model being considered is based on an open-source version of 超星发现 at www.chaoxing.com/.