The data is available online and as a standalone offline database in Microsoft Access and SQLite formats.[15] New versions are currently under development. Apart from working with its collaborators in universities and research institutions on scholarly research, in 2017 the CBDB project signed an agreement with ChineseAll, a digital publishing company to develop a commercial version pre-loaded with analytical tools for institutional subscribers. An Application Programming Interface (API) for system interoperability is also available, allowing other systems such as MARKUS to conduct real-time lookup of CBDB data.[16] Collaborators of CBDB have also transformed the database entries into a file for the iOS built-in dictionary application.[17]

Training on how to make use of CBDB has also become part of courses on sinological research tools and digital scholarship.[18] Detailed explanations of the database design and structure, utilities for exporting data for network and spatial analysis, etc. can be found in the CBDB Users Guide (available in Chinese and English)[19] and various sections of the CBDB website.[20] Video tutorials are also available online.[21]

The Visualization and Analysis of Historical Space

Peter Bol*

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

doi:10.1017/jch.2020.22

Abstract
A brief introduction to historical Geographic Information Systems and the creation of the China Historical GIS. Introductions are given to the CHGIS datasets covering 221 BCE to 1911 and the many GIS datasets on nineteenth and twentieth century China created under the leadership of the late G. William Skinner.

Maps, the preeminent form of spatial visualization, can be dated back over two millennia in China, but the earliest significant extant maps date back to the Song period. It is

[15]For the online query system, see China Biographical Database Project (CBDB), https://projects.iq.harvard.edu/cbdb/accessing-cbdb-online. For the offline standalone versions, see https://projects.iq.harvard.edu/cbdb/download-cbdb-standalone-database.

[16]See https://projects.iq.harvard.edu/cbdb/cbdb-api. For another example of how the CBDB API is utilized, see Chih-Ming Chen and Chen C. Chang, "A Chinese Ancient Book Digital Humanities Research Platform to Support Digital Humanities Research," *The Electronic Library* 37 (2019), 314–36.

[17]See "Download Latest CBDB Mac Dictionary" at https://projects.iq.harvard.edu/cbdb/download-cbdb-standalone-database.

[18]Peter Bol, "How the Digital is Changing Research and Teaching on Asia," *ASIANetwork Exchange: A Journal for Asian Studies in the Liberal Arts* 25.2. (2018), 7–28, doi: http://doi.org/10.16995/ane.278.

[19]https://projects.iq.harvard.edu/cbdb/supporting-documents.

[20]http://projects.iq.harvard.edu/cbdb.

[21]For video tutorials on youtube, see www.youtube.com/watch?v=uHWJuK308Jg&list=PLGgLzyv7BMgY8Dl1AbIBsCnVxnlRQf3Hw. For those on youku, see: https://list.youku.com/albumlist/show/id_26353417.html.

https://doi.org/10.1017/jch.2020.22 Published online by Cambridge University Press

CrossMark

clear from these works that in practice there was not a standard definition of what a map should be. The famous *Yujitu* which was engraved into stone in 1136, for example, is a grid-map that on its eastern half is reasonably accurate, despite the fact that it was drawn based on information on distance and direction between administrative seats. On the reverse is the *Hua-Yi tu*, a map of the civilized and tribal peoples, which aims only at approximation of space. The differences have to do with purpose: the *Yujitu* wants to show the relationship between contemporary places and topographic features with those appearing in the earliest geographical text, the "Tributes of Yu," and the *Hua-Yi tu*, in addition to giving prefectural seats and roughly depicting mountains and rivers, is loaded with textual accounts of the surrounding tribal peoples. Both maps are propositions about spatial relationships but have adopted different methods of representation to get the point across.[1] Also from the twelfth century is the first extant historical atlas, and the earliest printed one in human history, the *Lidai zhizhang tu*, a series of maps showing spatial changes in the administrative system from antiquity to the 1080s.[2]

Cordell Yee has written the most detailed history of Chinese cartography in English,[3] and Cao Wanru has edited a collection of reproductions and studies of historical maps.[4] Significant pre-1900 US Library of Congress China maps are viewable online at the Academia Sinica Digital Center.[5] Today important historical map series are also available online, scanned and (sometimes) georeferenced. These include the twentieth-century topographic maps from the US Army Map Service at 1:250,000 scale accessible at OldMaps Online,[6] the 1:25,000 and 1:50,000 Japanese and Republican maps accessible at the Center for GIS at Academia Sinica,[7] and the excellent Soviet 1:100,000 maps available commercially. Much US overflight and satellite photography is now publicly available.[8]

There are numerous thematic atlases and maps of historical events and historical data.[9] The most comprehensive atlas of the field administration, from antiquity to 1908, is the eight-volume historical atlas edited by Tan Qixiang.[10] It is superseded in geographic accuracy and temporal specificity by the China Historical GIS (CHGIS), discussed below.

[1]On the 1136 *Yujitu* see, for example, Alexander Akin and David Mumford, "'Yu Laid Out the Lands': Georeferencing the Chinese Yujitu [Map of the Tracks of Yu] of 1136," *Cartography and Geographic Information Science* 39.3 (2012), 154–69; Cheng Yinong 成一农, *"Fei ke xue" de Zhongguo chuan tong yu tu* "非科学"的中国传统舆图 (Beijing: Zhongguo she hui kexue, 2016); Peter K. Bol, "Exploring the Propositions in Maps: The Case of the 'Yuji tu' of 1136," *Journal of Song-Yuan Studies* 46 (2016), 209–24.

[2]Shui Anli 稅安禮, *Lidai dili zhizhang tu* 歷代地理指掌圖 (Shanghai: Shanghai guji, 1989).

[3]Cordell D.K. Yee, "Cartography in China," in *Cartography in Traditional East and Southeast Asian Societies*, ed. J.B. Harley and David Woodward, *The History of Cartography 2:2: Cartography in Traditional East and Southeast Asian Societies* (Chicago: University of Chicago Press, 1994).

[4]Cao Wanru 曹婉如 and et al., eds., *Zhongguo gudai ditu ji* 中國古代地圖集, 3 vols (Beijing: Wenwu, 1990–94). For holdings of maps in Chinese in the US and Europe see Li Xiaocong 李孝聪 and Library of Congress, *Meiguo guohui tushuguan cang zhongwen guditu xulu* 美国国会图书馆藏中文古地图叙录 (Beijing: Wen wu, 2004); Li Xiaocong 李孝聪, *Ouzhou shoucang bufen zhongwen guditu xulu* 欧洲收藏部分中文古地图叙录 (Beijing: Guoji wenhua chuban gongsi, 1996).

[5]http://digitalatlas.asdc.sinica.edu.tw.

[6]https://www.oldmapsonline.org.

[7]http://map.rchss.sinica.edu.tw.

[8]https://earthexplorer.usgs.gov/ and https://landsatlook.usgs.gov/viewer.html.

[9]Indexed and discussed in Endymion Porter Wilkinson, *Chinese History: A New Manual*, 5th edn (Cambridge, MA: Endymion Wilkinson, c/o Harvard University Asia Center, 2018).

[10]Tan Qixiang 谭其骧, ed. *Zhongguo lishi ditu ji* 中国历史地图集, 8 vols (Shanghai: Ditu, 1982).

https://doi.org/10.1017/jch.2020.22 Published online by Cambridge University Press

GIS

The geographic information system (GIS) is a more recent technology, dating back only to the 1960s and widely available commercially beginning in the 1980s. We should think of a GIS as a database in two parts. The spatial part is a set of points, lines (connections between points), and polygons (lines that cross over themselves to create enclosed spaces). The core concept of a GIS is that geographic space can be represented by points, lines, and polygons on a grid where the x/y coordinates represent longitude and latitude. Once we have "spatial objects," we need the second part of the database to store information about those spatial objects. A point might represent a town, a line a river, and a polygon the boundaries of a county. Information can be attached those spatial objects (the population of a county, the name of a river, the economic output of a province). Spatial objects can thus have multiple "attributes." Using GIS software we can project the spatial objects in two dimensional space (taking into account the three dimensionality of the earth). It is possible to scan maps, peg their four corners to their longitude and latitude, and lay points, lines and polygons on top. Very often we use as a basemap a digital elevation model (or grid data) in which every pixel has a color value to represent elevation.

The power of GIS is that it allows the user to overlay different kinds of data that have spatial attributes, and thus to see correlations that would remain opaque if one relied solely on statistical analysis. With a GIS we can take datasets that have spatial references—for example, the number of Buddhist monasteries, the number of Confucian academies, the tax burden, and the number of civil service degree holders from these places—and overlay one over the other to see if there are correlations. In contrast to printed maps, a GIS is an open-ended and flexible system that gives users the ability to choose what they want to compare and how they want to represent it.[11]

GIS databases and georeferenced maps can be downloaded, but they require GIS software. ArcGIS, from ESRI the leading GIS software company, is the most sophisticated and powerful; it requires a license. Quantum GIS (QGIS) is freeware that works on both Windows and Apple operating systems; it is easy to use and quite adequate for historical studies.

Historical GIS aims to add a temporal component to spatial analysis, which under normal circumstances excels in variation across space rather than change over time.[12] It allows users to answer two questions: What was the situation at a particular moment in time? and How did that situation change over time? Most data are not only place specific, but also time specific. But to represent these spatiotemporal data one must have accurate locations of the places that existed in the past. For example, suppose one wanted to trace the growth of a population over time and its changing distribution through space. Suppose further that the population data were collected by administrative unit (county, prefecture, circuit, province) over a two-hundred year period but that the names and boundaries of the administrative units changed during this period. The prerequisite for analyzing the

[11]For a more thorough introduction to GIS see the Wikipedia article at https://en.wikipedia.org/wiki/Geographic_information_system.

[12]Anne Knowles has led the way in the application of GIS to historical studies; see Anne Kelly Knowles, ed. Special Issue: Historical GIS: The Spatial Turn in Social Science History, *Social Science History* 24.3 (2000); id. *Past Time, Past Place: GIS for History* (Redlands, CA: ESRI Press, 2002); id. "Emerging Trends in Historical GIS," *Historical Geography* 33 (2005), 7–13; Anne Kelly Knowles and Amy Hillier, eds., *Placing History: How Maps, Spatial Data, and GIS Are Changing Historical Scholarship* (Redlands, CA: ESRI Press, 2008). For practical advice see Ian Gregory, *A Place in History: A Guide to Using GIS in Historical Research* (Oxford: Oxbow, 2003); Ian Gregory and Paul S. Ell, *Historical GIS: Technologies, Methodologies and Scholarship* (Cambridge: Cambridge University Press, 2007).

https://doi.org/10.1017/jch.2020.22 Published online by Cambridge University Press

population data would be a GIS that accurately depicted the units—their location, their name changes, and their size. This would be the "base" GIS, for it would provide the basis for depicting and analyzing all the population records that were collected by county. But this would also be a common base, for it could be used to depict and analyze all the different kinds of data that were collected by administrative units for this period. The fundamental task of historical GIS is in fact a research task in historical geography, to create an historically accurate common base, something that will serve many different research interests by providing the history of the units that were used to collect information in the past. The common base GIS has value in and of itself, it becomes a digital historical atlas that can be corrected, updated, and shared at minimal expense.

China Historical GIS (CHGIS)

The CHGIS project was launched in 2000, inspired by the work of Robert M. Hartwell, who had drawn on work on a contemporary GIS for China by G. William Skinner, William Lavely, and Lawrence W. Crissman. The project, a collaboration between Harvard and the Center for Historical Geography at Fudan University, was made possible through the support of the Henry Luce Foundation, with later support from the National Endowment for the Humanities. The general editors are Peter K. Bol and Ge Jianxiong, with executive editors Merrick Lex Berman and Man Zhimin.[13]

The goal in creating CHGIS has been to make it possible for users to project their data—whether population reports, tax quotas, military garrisons, or religious institutions—onto a digital map that accurately reflects the administrative structure and settlements at any given year from 221 BCE to 1911 CE. By downloading the CHGIS datasets GIS users can join their own datasets and create new data layers appropriate to their research. They can analyze that data, test hypotheses about spatial relationships, generate historical maps for research, teaching, and publication, and share their data with others.

CHGIS serves two other functions. First, it is an historical gazetteer. Using TGAZ, the online query utility,[14] users can search by place name to see when it existed, where it was, where it belonged in the administrative hierarchy, and the historical documentation for these findings. This means that almost all administrative place names down to the county (*xian*) level can be accurately located. This free utility currently receives between 20,000 and 50,000 search requests per month. Second, it is an electronic atlas. Because CHGIS has adopted a "continuous time series" approach—that is, it traces all known changes over time, so that users are free to decide which moment(s) they wish to view—it may be used as a source of maps of the known administrative structure and settlements of China for any year between 221 BCE and 1911 CE. Users can do this for themselves, generating layers as they need them, or they may view and download CHGIS layers set at 50-year intervals from 1350 to 1900 on the WorldMap platform.[15]

There is an important caveat. CHGIS aimed to go beyond the "slices in time" of the Tan atlas and create a continuous time series, recognizing that for some periods historical documentation would be lacking. It codes multiple levels of uncertainty in the

[13]Peter K. Bol, "Creating a GIS for the History of China," in *Placing History: How Maps, Spatial Data, and GIS are Changing Historical Scholarship*, ed. Anne Kelly Knowles and Amy Hillier (Redlands, CA: ESRI Press, 2007); Peter K. Bol and Jianxong Ge, "China Historical GIS," *Historical Geography* 33 (2005); Peter K. Bol et al., "China Historical Geographic Information System 1.0–6.0," Harvard University and Fudan University, 2002–2019.

[14]http://maps.cga.harvard.edu/tgaz/.

[15]http://worldmap.harvard.edu/maps/chgis/dtg. Datasets prepared by Yuan Yuying.

https://doi.org/10.1017/jch.2020.22 Published online by Cambridge University Press

datasets. The territorial boundaries of administrative units are particularly difficult to document for some periods and some regions. Prefectural borders are only complete for north and south from 1350 on and in general CHGIS coverage is limited to those areas under the centralized civil administration.

CHGIS datasets

CHGIS datasets come in the form of GIS files ("shapefiles") which can be downloaded through the CHGIS website.[16] There are four categories of data.

1. The "time series," from 221 BCE to 1911 CE, which can be queried by year.
2. The "slices in time," for the years 1820 and 1911.
3. The "digital elevation model," which gives the topography of eastern Eurasia.
4. The lakes, rivers, and coastlines, as of 1820.

There are two further distinctions.

1. Time series and slices in time files are divided into separate files for:
   - Point files, which are divided into Villages (for 1820 and 1911), County level seats, Prefectural level seats, and Provincial Capitals.
   - Polygon files, which are divided into Prefectural boundaries (complete for 1350–1911), County boundaries (for 1911) and Provincial boundaries (for 1820 and 1911)
2. Files are offered in different Chinese character encodings and projections:
   - Encoding in either GBK (PRC standard) or UTF8
   - Projection in either Xian80 (PRC standard) or WGS84

All files can be downloaded through the CHGIS website, through the Harvard Geospatial Library[17] or directly from Harvard's Dataverse.[18]

Map 1 is an example of a map that can be generated with CHGIS datasets above.

Using the same software any data with known spatial referents can be drawn on this map. Map 2 shows the distribution of the 1880 *jinshi* degrees granted during the Yongle reign, 1403–1424.

Nineteenth- and Twentieth-Century Datasets

Datasets for nineteenth- and twentieth-century China created under the leadership of the late G. William Skinner are available with their codebooks through the CHGIS webpage. They include:

Physiographic Macroregions of China

The map of China's physiographic macroregions specifies our informed decisions as to (1) the precise location of cutpoints in river systems that extend across

[16]https://sites.fas.harvard.edu/~chgis/.

[17]https://library.harvard.edu/services-tools/harvard-geospatial-library This search engine covers holdings at Harvard and many other institutions.

[18]CHGIS Version 6 Dataverse, https://dataverse.harvard.edu/dataverse/chgis_v6 for point and polygon files; https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/M7WEFY&version=5.0, for lakes, rivers, and coastlines; https://dataverse.harvard.edu/file.xhtml?fileId=3359166&version=8.0 for the digital elevation model with elevations color-coded.

https://doi.org/10.1017/jch.2020.22 Published online by Cambridge University Press



Map 1. Prefectural boundaries, prefectural seats, and county seats as of 1400 from CHGIS v. 6.0; DEM from CHGIS, rivers lakes and coastlines as of 1820 from CHGIS v. 5.0. Labels (in Chinese or *pinyin*) can be included.

regional systems, (2) the macroregional affiliation of small coastal river systems, and (3) which internal drainage basin along the Inner Asian frontier should be included in China Proper.

ChinaW—Cities, County Seats and Yamen Units (1820—1893)

ChinaW Dataset contains point locations for all Cities, County seats, and Yamen-Level units recorded in China from 1820 to 1893, along with more than 150 attribute values for each record. A table of gazetteers consulted for each county is also provided.

Socioeconomic Macroregions of China

The regional system analysis of twentieth-century China revealed nine macroregional economies/societies, excluding Inner Asian territories. The analysis covers the urban and the associated local and regional system hierarchies of these macroregions during the 1990–2000 period. The Socioeconomic Macroregions available here are snapshots of the regional systems (which change over time depending on variable inputs). The GIS Shapefiles provided here are divided into two layers, (1) for the Main Macroregions [SMR], (2) for the 33 Subregions [SMR_SubR].

https://doi.org/10.1017/jch.2020.22 Published online by Cambridge University Press



Map 2. *Jinshi* degrees granted from 1404 to 1424. Prefectural boundaries as of 1400 from CHGIS v. 6.0. *Jinshi* data from the China Biographical Database, v. CBDB_20190424

China-CPZ (Core Periphery Zones)

China-CPZ, the core periphery zones of China, circa 1990. Core-Periphery Zones are based on a regional systems analysis of China, drawing on data about population density and agricultural intensity for county level units (see the Regional Systems Index in the ChinaA datafile). Zones were delineated considering physical geographical features such as rivers, ridge-lines, slope, and the transportation network. China's core zones represent some of the densest populations and highest levels of agricultural productivity anywhere in the post-industrial world. Traveling outward from the metropolitan cores, these levels of intensity decline, reaching their lowest levels in the peripheries along regional boundaries. This analysis complements the central place analysis of cities and towns, represented in the ChinaT datafile, which identifies nested hierarchies of regional, greater, and local city systems.

China-T (1990) Central Place Analysis

ChinaT Dataset contains a record for every central place in China that was the central city (Cc) of a *shi* (municipality) or classed as a *zhen* (town) as of 1 July 1990.

https://doi.org/10.1017/jch.2020.22 Published online by Cambridge University Press

To these were added (1) records for *qu* (centers of peri-urban urban districts that do not appear as such in the census tables), and (2) capitals of county-level units that do not rank as *zhen*.

China-A Dataset, County-Level Units 1990

The ChinaA Datasets combine the raw geographical units and the essential county-level unit variables used for Skinner's 1990 China Regional Systems analysis. ChinaA MQ Dataset is composed of the ChinaA MQ (Merged Qu) geographic units that have been joined to tabular data compiled from the 1990 Census variables.

China City Walls of Hebei and Shandong

The 1904–07 Prussian maps show the footprints of settlements including the outlines of city walls. Small areas around each capital in the ChinaW file were clipped from the georeferenced maps. Filenames contain the ChinaW ID number for each city.

Karte von Tschili und Schantung (Maps of Zhili (Hebei) and Shandong), 1904–1907

These detailed maps compiled by the Royal Prussian Land Survey provide a unique view of the North China landscape of a century ago, showing the built-up areas, or footprints, of settlements from walled cities down to small villages. Digital scanned images of these maps have been prepared for research on historical land use change.

Various other GIS datasets are also available. Examples include the Language Atlas of China (William Lavely and Lawrence Crissman), ChinaHydro NavNet (Crissman and Merrick Lex Berman) and ChinaHydro Basins (Crissman and Berman).

Online GIS Platforms for China

Today users can do some spatial analysis through online platforms, either by uploading their own data or using shared data from others. The key to a good platform (in my view) is stability, long-term institutional support, and allowing users to upload and share data. Three general platforms stand out. ESRI's ArcGIS Online has free public and institutional options and has strong support.[19] WorldMap.harvard.edu is open-source and open-access. It has map collections on contemporary and historical China, CHGIS datasets, and Skinner datasets.[20] The Academic Map Publishing Platform from Zhejiang University 學術地圖發佈平台 (based on the WorldMap platform) displays a growing number of maps pertaining to Chinese history and

[19]www.arcgis.com/home/index.html. As an example, see the online version of Map 2 at http://arcg.is/0je40u. The free public option limits the size of user datasets.

[20]Contemporary China: http://worldmap.harvard.edu/chinamap/. Historical periods: http://worldmap.harvard.edu/maps/chinaX and other mapsets in the ChinaX series. CHGIS http://worldmap.harvard.edu/maps/chgis. Skinner: http://worldmap.harvard.edu/skinner/. Silk roads: http://worldmap.harvard.edu/maps/7547.

https://doi.org/10.1017/jch.2020.22 Published online by Cambridge University Press

literature.[21] The outstanding Silk Road Historical Geography Information Open Platform 丝绸之路历史地理信息开放平台 from Capital Normal University combines extensive data on western China and Inner Asia and a variety of other utilities.[22] Academia Sinica's Chinese Civilization in Time and Space 中華文明之時空基礎架構 includes several animated maps showing change over time but is not an open platform.[23]

Creating, Linking, and Analyzing Chinese and Korean Datasets: Digital Text Annotation in MARKUS and COMPARATIVUS

Hilde De Weerdt*

Leiden University
*Corresponding author. Email: h.g.d.g.de.weerdt@hum.leidenuniv.nl

doi:10.1017/jch.2020.23

MARKUS, a multilingual digital text annotation and analysis platform, allows historians and other researchers to construct datasets from primary sources available to them in full-text digital format. Originally designed for those working with pre-twentieth-century Chinese texts, MARKUS has developed into a multifunctional annotation platform that is particularly suited for the automated annotation, referencing, and visualization of named entities in modern and literary Chinese and premodern Korean texts, but many of its additional annotation features can be used to analyze and read texts in any language, as long as the electronic documents are encoded in the most common standard for language encoding, Unicode. Below I discuss the main goals and methodological features of MARKUS and the allied text comparison utility COMPARATIVUS.[1] I will illustrate these with some examples of how MARKUS has been used in Chinese and Korean historical research.

[21] http://amap.zju.edu.cn.
[22] www.srhgis.com
[23] http://ccts.sinica.edu.tw/framework.php?lang=zh-tw.

*The development of MARKUS has been funded by the European Research Council under the European Union's Seventh Framework Programme (FP7/2007–2013) / ERC grant agreement n° 283525 (initial development by De Weerdt and Ho, and COMPARATIVUS development by De Weerdt, Gelein, and Ho) http://chinese-empires.eu/, the National Endowment for the Humanities /JISC--Digging into Data Challenge (machine learning module developed by Miao Shengfa) http://did-acte.org/, and an Asian Modernities and Traditions Large Grant, Leiden University (development of K-MARKUS by De Weerdt, Gelein, Ho, Hu, Kim, et al.) www.universiteitleiden.nl/en/dossiers/asian-modernities-and-traditions/research--funding-opportunities#caname-critical-approaches-to-new-asian-media-ecologies.

[1] On the algorithm used in COMPARATIVUS, see Paul Vierthaler and Mees Gelein, "A BLAST-based, Language-agnostic Text Reuse Algorithm with a MARKUS Implementation and Sequence Alignment Optimized for Large Chinese Corpora," *Journal of Cultural Analytics* March 18, 2019. DOI: 10.31235/osf.io/7xpqe.

https://doi.org/10.1017/jch.2020.22 Published online by Cambridge University Press