Geocoding and Mapping Historical Places

Yuying Yuan

yyy_yolanda@163.com

For researchers who study history, historical geography or other relevant academic fields, there may be some cases when they have a large number of data of historical places but cannot make the most use of them due to lack of efficient integration method. This blog will give an example of how to use computer techniques to assist historical research. More specifically, as is shown in the title, it is about geocoding historical Chinese place names using Python and mapping them as well as doing relevant analysis with GIS. Although the example here only includes place names in the Tang Dynasty (AD 618 to AD 907) in China, the whole workflow provides a general method to build historical geographical database based on historical documents and other online or offline data sources.

1. Data preparation

1.1 Original data extraction

Here, the Tang place names were collected from General History of Administrative Region in Historical China, the Tang Volume (Guo, 2012) (referred to as Guo-book in this blog). This book is regarded as the most authoritative book in this field up to now. Collecting and analyzing information from hundreds of historical documents, such as the Old Tang History 《旧唐书》 and the New Tang History (《新唐书》), as well as results of previous research and archaeological work, it presents the whole detailed process of how administrative regions of different levels changed during Tang dynasty. It provides information on the administrative level, the beginning year and ending year of a place, the superior administrative unity it belonged to, and relevant historical materials. However, this book does not give the coordinate information for these places. Only some of the places have information of their corresponding contemporary administrative units.
To prepare the data of historical place names, the first step is extracting the original data from source materials and putting them into a structured CSV/XLS file that includes several columns, each column representing a certain kind of data, for example, administrative level. Since the structure and narrative of the Guo-book are highly similar in each chapter, it is easy to extract information and put them in a well-structured spreadsheet with five fields: place name, corresponding administrative level, the beginning year of this administrative unit, its end year, and the upper-level administrative unit it belonged to.

1.2 Data cleaning

The second step is cleaning and standardizing the spreadsheet. Data should be concise and explicit. Data in the same column should have same written structure, such as using the same separators, all written in either lowercase or uppercase, using consistently either traditional/Fanti or simplified/Jianti Chinese characters, a standard way of mark cells as blank and so on. This step is of great importance because it will affect the result of geocoding.

2. Geocoding

Next, it comes to geocoding these Tang place names. There are mainly four resources that could be used. These four resources can be divided into three types, each of which represents a different way to use.

2.1 Local database: China Biographical Database Project (CBDB)

The CBDB has data on Tang places. By comparing the Guo book data with existing CBDB data, we can find locations for some places. The principle of comparison is to check if the data in each of the five fields listed above match. This step is quite straightforward using either Python or Excel. It is just matching information in two Excel files and finding any differences between them.

2.2 Online search without using an API: TGAZ & Buddhist Studies Authority Database Project
TGAZ (http://maps.cga.harvard.edu/tgaz/), developed by the Center for Geographic Analysis of Harvard University, provides web service for querying historical place names. The source of Chinese data is the China Historical Geographic Information System (CHGIS). The introduction on the website states: “TGAZ API - is a read-only interface designed to search the contents of the China Historical GIS placename database. The TGAZ system architecture -- based loosely on the CHGIS XML API (2006) -- has been normalized and made more generic in order to integrate multiple data sources, and to allow for vernacular spellings or transcription methods in many languages.” The json format and xml format data could easily be accessed by inputing certain information in the website address. This makes it convenient for researchers to search and get data from this website using a programming method. “http://maps.cga.harvard.edu/tgaz/placename?fmt=&n=&yr=&ftyp=&src=” is the general form of the website address. The “fmt=&n=&yr=&ftyp=&src=” part represents different search fields (table 1). If we want to search a place named “朔州” and see the result in json format, we could just add two elements into the website address – “http://maps.cga.harvard.edu/tgaz/placename?fmt=json&n=朔州&yr=&ftyp=&src=”. Since CHGIS is the only source for Chinese place names, there is no need to add anything to the “src=” part. Figure 1 shows part of the search result. The section marked by the red rounded rectangular is a typical example of the kind of information the search produces.

Table 1  Means of different parts of the website address

<table>
<thead>
<tr>
<th>Field</th>
<th>fmt=</th>
<th>n=</th>
<th>yr=</th>
<th>ftyp=</th>
<th>src=</th>
</tr>
</thead>
<tbody>
<tr>
<td>meaning</td>
<td>website format</td>
<td>place name</td>
<td>existing year</td>
<td>feature type</td>
<td>source</td>
</tr>
</tbody>
</table>
Figure 1  Search result of “朔州” in json format

In addition to the function of the website address, this example also shows that the json format result has a clear structure, which is easy for Python to capture and parse. However, as here, the search may give more than one result for one place name. Thus, when saving the several results of one place name into an Excel file, it is best to create new file-ids to record the number of results and a sub-id for each result that belongs to the same place name.

The logic of Python code is clear: (1) searching the place name in TGAZ; (2) only keeping those results that have overlapped existing year; (3) only keeping those results for which the feature type is point (because only points have longitude and latitude information); (4) capturing the content of the final result and writing them into an Excel file, while at the same time keeping both original input and final output information. Figure 2 shows part of the results, illustrating how multiple results are saved.

Figure 2  Part of the search result

The Dharma Drum Institute of Liberal Arts has a Buddhist Studies Authority Database
Project. It contains databases integrating information from various projects at the Library and Information Center at Dharma Drum Institute of Liberal Arts. The way to use this database is quite similar to TGAZ. The search and comparison process can easily be done through modifying some part of the URL. Thus, the detailed process of using this database will not be presented here.

However, based on our results from querying TGAZ and the Buddhist Studies Authority Database Project, there are still more than one thousand places that cannot be located on the map. If we plot places with matched coordinates as points on the map and create a corresponding heat map (figure 3) using QGIS, we can see that these identified places are mainly distributed in the area north to the Yangtze River basin. There are two possible reasons. First, the political, economic and cultural center was in the north during the Tang dynasty. Thus, the northern area had more population as well as more administrative units. Second, there were more written materials in the north, which makes it easier to locate places there. As a result, the geo-data of northern China are more complete. There is, however, an important exception to this: the area around modern Guangzhou (Canton), the major port for trade with Southeast and South Asia and the Middle East.
2.3 Online search with need of using API: Google/Gaode/Baidu Map

There is a third way to find the missing coordinates for those one thousand places. Thanks to the Guo-book, some of the places have information about their modern locations. Therefore, modern online map websites, such as Google Maps, Baidu Map, and Gaode Map, can be used to search for their coordinates. To do a large number of queries with programming, most online map websites require the user to apply an API Key. The whole process is similar to the previous two. The only difference is we must send the API Key to the website so that we can get results by modifying a certain part of a certain URL. Different websites have different ways of using an API Key and they all have very detailed instructions on their websites.

2.4 Data selection

After all these steps, we can find the longitude and latitude for more than ninety percent of the Tang places. However, some of them have more than one coordinate. This requires a manual check to select the correct one. Once the selecting process is finished, the cleaned, well-organized spreadsheet becomes a complete database of Tang places.

3 Mapping and GIS analysis

Researchers can either use the newly-built database as a source for searching Tang place names or they can treat this as big data and carry out macroscopic analysis. Figure 3 has already provided an example. With enough data and GIS tools, researchers may find new phenomena and trends that deserve further investigation. GIS tools provide various types of analysis. The websites of ArcGIS and QGIS are good places to find out more.

Summary

This is a brief procedure with example of how to build a new historical geographic
database and how to use computer techniques as well as different sources to facilitate historical and geographic researches. Figure 4 summarizes the general workflow.

Figure 4  Flow chart of historical geographic database building and using

References

Book

Website
ArcGIS: https://www.esri.com/zh-cn/arcgis/products/arcgis-online/resources
Buddhist Studies Authority Database: http://authority.dila.edu.tw/
CBDB: https://projects.iq.harvard.edu/cbdb
CHGIS Version 6: https://sites.fas.harvard.edu/~chgis/data/chgis/v6/
Gaode map: https://ditu.amap.com/
Gaode API instruction: https://lbs.amap.com/api/webservice/guide/api/georegeo
TGAZ: http://maps.cga.harvard.edu/tgaz/
WorldMap: http://worldmap.harvard.edu/maps/chgis

**Python code**

https://github.com/cbdb-project/CBDB-Tang-geocoding/blob/master/geocoding_tgaz.py