Modern China Geospatial Database

Discussion published by Christian Henriot on Monday, January 10, 2022

The ENP-China project (ERC AdG) is pleased to announce the release of three geospatial datasets of its Modern China Geospatial Database. The main dataset — MCGD_Data_V2 — is the historical gazetteer that we have been developing to identify place names in China under their various denominations (in Chinese) and spellings (in different historical transliteration systems and in English) and their geocoordinates (Latitude/Longitude).

This is meant to serve as a reference tool for place names in modern China and facilitate the identification of places under their varying designations in historical sources. To this end, the ENP-China project has also developed a search interface to be released in the coming weeks. All through the preparation of the gazetteer, we have been keeping a record of how we developed this gazetteer. In this H-Asia post, I will only introduce the three datasets. For a full introduction on the making of the MCGD dataset, please read our post on the ENP-China blog.

Note: As a rule, all the names in Chinese in our datasets are provided in traditional script, except for MCGD_PRC_V1 for which we added a column with the names in simplified Chinese.

We built the initial place name dataset based on data from the Geonames database and from the Harvard Historical GIS database. The construction of the three datasets resulted from a process of selecting, curating, cleaning, and enriching the original data.

The three datasets are available at the ENP-China Data Repository on Zenodo (see the direct links to each dataset below).

Brief description of the datasets:

<table>
<thead>
<tr>
<th>File</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCGD_Data_V2</td>
<td>MCGD_Data_V2 contains all the data that we have collected on locations in modern China. Altogether there are 466,165 entries. The data include the name of locations and their variants in Chinese, pinyin, and in any recorded transliteration system; the name of the province in Chinese and in pinyin; Province ID; the latitude and longitude; the Name ID and Location ID. The Name IDs all start with N followed by seven digits, except for locations in Taiwan that start with “T” (data from Geonames). This is the internal ID system of MCGD. Locations IDs that start with “DH” are data points extracted from Harvard Historical GIS (Harvard University); those that start with “D” are locations extracted from the data points in Geonames; those that have only digits (8 digits) are data points we have added from various map sources.</td>
</tr>
</tbody>
</table>
MCGD Rep V2 is a sample of data extracted from MCGD Data V1 for the Republican period. It is based mostly on the list of xian (county) seats in 1931 [Source: Zang, Lihe 臧励龢, ed. Zhongguo gujin diming da cidian 中国古今地名大辞典. Shanghai 上海: Commercial Press, 1931], with the addition of some external sources. In fact, it covers more than just Republican China. It presents a list of the major locations in China between 1900 and 1949. It contains 1,977 entries with the name in Chinese and in pinyin; the name of the province in Chinese and in pinyin; the latitude and longitude, and the Name ID and Location ID.

MCGD PRC V1 is a list of cities in today’s PRC. It includes 2,461 locations, with their name in Chinese (both traditional and simplified Chinese) and in pinyin, the name of the province in Chinese and in pinyin; the latitude and longitude.