

Toponyms and Feature Classifications for the China Historical GIS

Merrick Lex Berman
CHGIS, Harvard Yenching Institute
<http://fas.harvard.edu/~chgis>
January 2001

One of the basic problems facing digital gazetteer projects is how to atomize placenames into their component parts so that they can be effectively arranged in the database. Chinese toponyms are no exception, and have a few characteristics deserving special attention, which are introduced below.

How such distributed systems might work, were discussed at the Digital Gazetteer Information Exchange workshop (DGIE)¹ and are part of ongoing work at the Alexandra Digital Library (ADL).² An in depth look at how ADL structured its digital gazetteer around the core components of name, location, and type, can be found in Hill,³ while a thorough explanation of the query mechanism by use of “search buckets” is presented in Frew, et al.⁴ One of the most intriguing possibilities for the development of distributed digital gazetteers is Nebert’s proposal to establish a “geographic name service (GNS),” in which a placename query would cascade through a series of distributed gazetteers and return a summary of relative hits and their sources.⁵ In the GNS model, requests are sent to authenticated gazetteer servers, some of which would be official national authorities, while others could be independent agencies or highly specialized resources. The historical toponym information being developed by the China Historical GIS project (CHGIS) would fit into this scheme as one of the highly specialized resources, as in Figure 1.

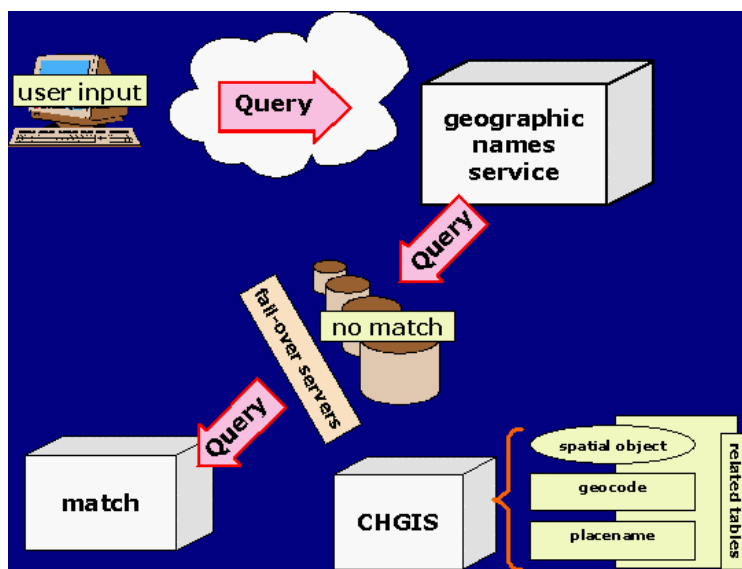


Figure 1: Geographic Name Service

For the purposes of CHGIS, it is essential to be able to match Chinese placenames (as text strings) to the appropriate records in the database. In this paper we'll concentrate on romanized placenames, and look at the component parts of a given toponym. In Chinese we often see the usage "name" (*tong ming*) and "identifier" (*zhuan ming*) in discussions of toponyms. For the purposes of this discussion, let us call them "name" and "type" respectively, and take them to mean the feature name and the feature type components of a particular toponym. We could begin by breaking a toponym up into names and types, when "types" include administrative units.

Feature Name	Feature Type
Qianping	Xian
Pinglang	Anfusi
Kunming	Fu

For feature names, we must also be sure to allow for variant spellings and alternate names (including alternate Chinese characters for the same placename) in our search process, which can be done by automatically querying a dataset of alternate names, as illustrated in Figure 2.

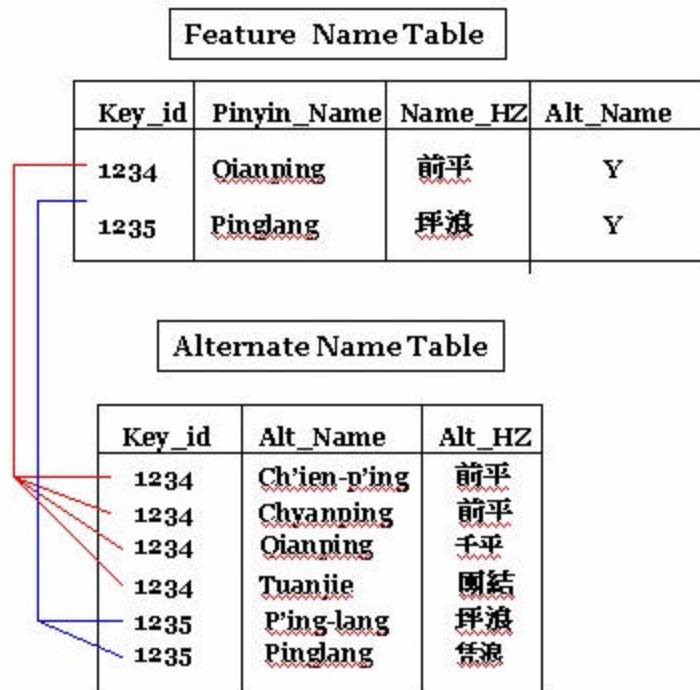


Figure 2: Alternate Name Model

The major advantage in setting up queries in this manner lies in the assignment of a unique identifier, ("Key_id" in Figure 2), that is linked to only one object in our database. Each

unique **object** has a key_id, and a one-to-many relationship with all of its associated placename records. The term “object” in this case is meant to describe a place that we can pinpoint in space and time, even though there may be more than one way to represent this place spatially. So, in addition to having more than one name, or more than one type, each object can have multiple representations: as points, lines, polygons, animations, videos, soundtracks, etc, all of which can be selected *after* we have discovered the correct match to our query from all the possible hits in the database.

This brings us to another problem: ambiguity of toponyms. If we are trying to search a digital gazetteer of U.S. placenames for “Springfield,” we will certainly have many hits. In fact, a query on “Springfield” sent to the USGS GNIS server ⁶ had 498 matches. Certainly, queries that combine both name and type will help to narrow down the list of matches somewhat. If we specify “populated places” as the type, we find only 85 Springfields. Let’s say that we knew which state to look in, for example, Georgia. Indeed, there are only 11 Springfields in Georgia, and never more than one in a given county! However, it is unreasonable to expect that our end-users will always know which county a placename is found in. Furthermore, there remains ambiguity among county names as well.

We can make use of the Ming Dynasty dataset found in Robert Hartwell’s Historical GIS of China ⁷ for an estimate. Of the 1,626 county records in Hartwell’s data for the year 1391, there were 292 records that shared placenames with at least one other record. There was one case in which the same place name was shared by seven different counties, there were three cases in which one placename was shared by four different counties, and so on...as shown in Figure 3. Altogether some 18% of the total were ambiguous.

<i># cases</i>	<i># places w/ same name</i>	<i>sub-total</i>
1	7	7
1	6	13
3	4	25
29	3	106
93	2	292
<i>total # of counties in dataset 1626</i>		
<i>total # of ambiguous placenames 292</i>		

Figure 3: Ambiguous Placenames

Obviously we must make sure that the objects are distinguished from one another before moving from placename searches to integration with other types of search engines, such as those that analyze unstructured digital texts. Once we move to free text search and classical sources, we may find that the percentage of ambiguity is unacceptable. For instance, in dealing with the

Perseus digitized classical European text archive, David Smith, estimates that the level of ambiguity among toponyms in free text searches can exceed 90%.⁸

The CHGIS project will use a combination of **key_id's** for unique places and **geocodes** to identify those places in the historical administrative hierarchy. To use Hartwell's dataset as an example, we see seven ambiguous placenames ("Xincheng") as text strings in the top table, and the same units with unique geocodes in the bottom table (Figure 4).

<i>pinyin</i>	<i>admin_type</i>	<i>sup_prov</i>	<i>province</i>	<i>circuit</i>	<i>prefect</i>	<i>dep_pref</i>
Xincheng	Xian	Ming dynasty	Jingshi		Baoding	
Xincheng	Xian	Ming dynasty	Shandong		Ji'nan	
Xincheng	Xian	Ming dynasty	Zhejiang		Hangzhou	
Xincheng	Xian	Ming dynasty	Jiangxi		Jianchang	
Xincheng	Xian	Ming dynasty	Guangxi		Qingyuan	
Xincheng	Suo	Ming dynasty	Guizhou			Annanwei
Xincheng	Wei	Ming dynasty	Jingshi			Beiping

<i>Code</i>	<i>Level1_h</i>	<i>Level2_h</i>	<i>Level3_h</i>	<i>Level4_h</i>	<i>Level5_h</i>
M001100220025		0	11	0	22
M002500210026		0	25	0	21
M004100210027		0	41	0	21
M004500260022		0	45	0	26
M008500230032		0	85	0	23
M009500003722		0	95	0	37
M001100002422		0	11	0	24

Figure 4: Geocoding to Disambiguate Placenames

Assuming that we now have a way to match our placenames with the correct objects in our digital gazetteer, let's take a closer look at **feature types**. Are we able to break up the name and type components as shown in the following example?

Feature Name	Feature Type
Xishuangbanna	Zizhizhou
Xiao Ganlanba	Gangkou
Yunnanyi	Cun

Xishuangbanna, being the transliteration of the Dai / Thai name, *sipsong panna*, actually translates as: “the twelve *panna*.” The term *panna* may have originally been derived from the Thai *pan*, meaning “a flat area,” and *na*, meaning “a rice field.” Nonetheless, the word *panna* later came to be used specifically for an administrative region, fiefdom, or taxable division of various Thai states. Therefore our placename Xishuangbanna is actually a compound of both name and type. I believe that we need to establish a clear-cut set of rules that applies to the classification of names and types, so that toponyms can be easily atomized into their appropriate elements.

In the case of Xishuangbanna, I believe that the term is one that is clearly used as a single feature name, and cannot be subdivided. But in our second example, Xiao Ganlanba, I believe that the word Xiao, should be considered a **descriptive prefix**. Since there is also a Da Ganlanba (also known as Ganlanba), we are safe to assume that the prefix is a directional indicator, and not an inherent part of the placename. Our third example, Yunnanyi, presents yet another special case: historical accretion. In the Qing period, Yunnanyi was, in fact, a postal station, *yi*. Since the Province was also called Yunnan, and the capital was called Yunnanfu, it became necessary to always use *yi* to distinguish the capital city from a tiny village with a postal station, located strategically between Yunnanfu and Dalifu. In this case, I would argue that we are dealing with three ambiguous names, and three different types, that also overlapped on the **temporal scale**.

Feature Name	Feature Type
Yunnan	Sheng
Yunnan	Fu
Yunnan	Yi

By the early 20th century, Yunnan Yi became Yunnanyi owing to popular usage, and today we could only classify it as: Yunnanyi Cun. This is a case where we must have two unique objects in the database to refer to the earlier and later forms, demarcated by beginning and ending dates. This leads me to suggest that we might adopt a feature type classification scheme along the lines of Tiger Line Files usage, from which the following has been adapted: ⁹

Feature Name Descriptive Prefix	Nan Guanfang hutong
Feature Name	Xincheng Xian
Feature Name Descriptive Suffix	Xinjiekou wai dajie
Feature Type	Baiyun feijichang

Incidentally, the placename Xinjiekou, may **seem** to be an accretion of *xin* and *jielou*, but I believe that historical documents will show that there was not a Jielou from which a Xin Jielou was to be distinguished. It’s merely a case of a name being given that way from the start. The same does not hold true for Nan Guanfang, because there actually is a Bei Guanfang as well.

Another complication, which is peculiar to Chinese, is found in mandatory name - type compounds, when the feature name is only one syllable. For example, we would normally break up Yun Xian, into the following:

Feature Name	Feature Type
Yun	Xian

But placename convention in Chinese does not allow us to ever refer to Yun by itself. Whereas we could say the name Kunming, or use it as a label on a map, and in both cases it would be perfectly understandable. But Yun will not work in spoken or written Chinese. In this case, we must use the following formula, which has unsatisfactory results in automated labeling applications:

Feature Name	Feature Type
Yun Xian	Xian

Even so, I believe that it is more effective to use the second method and to repress the addition of the second Xian, than it is to add an additional process to parse all of our Feature Names.

It is more logical to assume that the majority of multi-syllable feature names would be queried without their corresponding feature types or administrative units (*ie*, Kunming), but that single-syllable feature names can **never** be queried this way (*ie*, never query Yun by itself, but must query Yun Xian). However, if we do **not** separate the feature name and feature type for either case, and put both names and types together in the names field (*ie*, Kunming Fu, and Yun Xian), then additional automated search engines will be needed to parse out the single-syllable names! Therefore, I strongly advocate the separation of feature name and feature type into separate fields in our Chinese language digital gazetteers, **except** in the case of single-syllable feature names, which must have both an entry of name and type in the feature name field, **plus** the type in the feature type field. Additionally, I suggest that we should not allow NULL entries in the type field.

To properly display or label our units, we would allow for either feature names alone, or feature names combined with feature types. To repress the addition of a second Xian in the case of single-syllable names, we can add a conditional field (a single syllable flag), NULL by default, and N to indicate that the type should **not be added** to the name:

<i>name</i>	<i>type</i>	<i>flag</i>
Yun Xian	Xian	N
Kunming	Shi	

In the case of Kunming, the label would read “Kunming Shi,” while “Yun Xian” would remain as “Yun Xian.” This will be much easier than trying to parse every case of type that occurs in the name field.

In conclusion, I would like to advocate for the development of a multi-lingual glossary of feature types. Building on the doctoral research by Hahn,¹⁰ we can expand the list to include all of the specialized terms found in particular locales. Indeed, there may be no direct correlations between particular usages for religious and cultural sites of various kinds, but to pool our resources and examine the feature types in more detail, we will at least be able to identify each type with a unique identifier. From there, we can look forward to the ability to use our distributed gazetteer search models for locating specific types in space and time.

Notes:

1. **DGIE Workshop.** Washington D.C. Oct 12-14, 1999.
http://www.alexandria.ucsb.edu/gazetteer/dgie/DGIE_website/DGIE_homepage.htm
2. **ADL.** <http://alexandria.sdc.ucsb.edu/>
3. **Linda Hill.** “Core Elements of Digital Gazetteers: Placenames, Categories, and Footprints.”
http://alexandria.sdc.ucsb.edu/~lhill/paper_drafts/ECDL2000_paperdraft7.pdf
4. **James Frew,** Michael Freeston, Linda Hill, Greg Janee, Mary Larsgaard, Qi Zheng. “Generic Query Metadata for Geospatial Digital Libraries.”
<http://www.computer.org/proceedings/meta/1999/papers/55/jfrew.htm>
5. **Doug Nebert.** “NSDI and Gazetteer Data.”
http://www.alexandria.ucsb.edu/gazetteer/dgie/DGIE_website/session3/nebert.htm
6. **USGS Geographic Names Information Service.** <http://geonames.usgs.gov/gnisform.html>
7. **Robert Hartwell.** “China History Project.” <http://www.people.fas.harvard.edu/~chgis/data/hartwell/>
8. **David Smith.** “Extracting Geographic Information from Semi-Structured Text.”
Presented at PNC – ECAI, Hong Kong, January 2001. <http://www.perseus.tufts.edu>
9. **US Census Bureau.** “Redistricting Census 2000 TIGER/Line Files Technical Documentation,” p.59.
http://www.census.gov/geo/www/tiger/rd_2ktiger/tgrrd2k.pdf
10. **Thomas Hahn.** “Structured and classified overview of independently treated, spatially relevant elements identified in 40 Chinese local mountain gazetteers.”
http://www.library.wisc.edu/guides/EastAsia/ECAI/Spat_elements/

The slides on this topic are available at:

http://fas.harvard.edu/~chgis/data/pubs/lex_gazeteers_jan01.ppt